Mapping 1,000+ Language Models via the Log-Likelihood Vector

Momose Oyama[1,2] **Hiroaki Yamagiwa**[1] **Yusuke Takase**[1] **Hidetoshi Shimodaira**[1,2]
[1]Kyoto University [2]RIKEN
`oyama.momose@sys.i.kyoto-u.ac.jp`, `h.yamagiwa@i.kyoto-u.ac.jp`,
`y.takase@sys.i.kyoto-u.ac.jp`, `shimo@i.kyoto-u.ac.jp`

Abstract

To compare autoregressive language models at scale, we propose using log-likelihood vectors computed on a predefined text set as model features. This approach has a solid theoretical basis: when treated as model coordinates, their squared Euclidean distance approximates the Kullback-Leibler divergence of text-generation probabilities. Our method is highly scalable, with computational cost growing linearly in both the number of models and text samples, and is easy to implement as the required features are derived from cross-entropy loss. Applying this method to over 1,000 language models, we constructed a "model map," providing a new perspective on large-scale model analysis.

1 Introduction

Language models have been evolving rapidly, and their community has expanded significantly. To understand this landscape and its future directions, it is essential to systematically analyze model similarity and positioning based on language modeling principles. On the Hugging Face Hub, models are categorized by name and attributes, while other studies assess similarity based on outputs (Yax et al., 2025) or activations (Zhou et al., 2025). Leaderboards (Beeching et al., 2023; Fourrier et al., 2024; Chiang et al., 2024) are commonly used to assess model standings.

Since language models are probability models, we propose representing each model using coordinates that capture the geometric structure of the space of probability distributions. Concretely, we define a language model's coordinates as its log-likelihood vector across a large collection of texts. Figure 1 shows a model map obtained through dimensionality reduction applied to the coordinates of 1,018 language models. This visualization reveals that models of the same type tend to cluster together, while models in close proximity often share the same primary text category, forming a continuous distribution across the map.

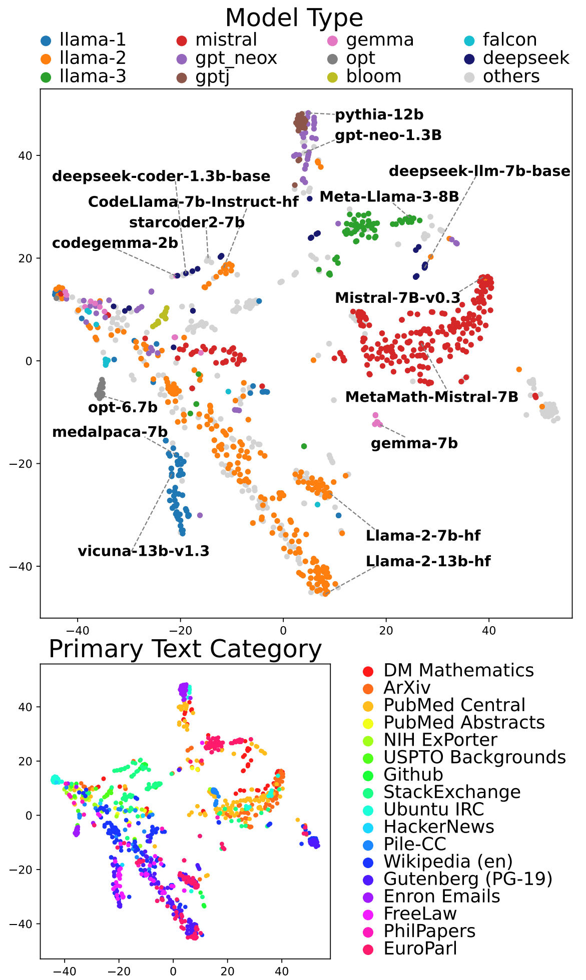

Figure 1: Map of 1,018 language models. Their log-likelihood vectors are visualized using t-SNE. (Top) Colors indicate model types. (Bottom) Colors indicate the model's "primary text category," the text category where the model achieves the highest standardized log-likelihood score among 17 text categories. See Section 4 for details.

Our code and data are available at `https://github.com/shimo-lab/modelmap`.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 32983–33038
July 27 - August 1, 2025

meta-llama/Meta-Llama-3-8B	KL [bpb]	google/codegemma-2b	KL [bpb]	deepseek-ai/deepseek-llm-7b-base	KL [bpb]
Undi95/Meta-Llama-3-8B-hf	4.56e-06	deepseek-ai/deepseek-coder-1.3b-instruct	2.46	deepseek-ai/deepseek-moe-16b-base	0.194
dfurman/Llama-3-8B-Orpo-v0.1	0.0104	bigcode/starcoderbase-1b	2.69	deepseek-ai/deepseek-llm-7b-chat	0.364
migtissera/Tess-2.0-Llama-3-8B	0.0428	deepseek-ai/deepseek-coder-1.3b-base	3.14	deepseek-ai/deepseek-moe-16b-chat	0.368
freewheelin/free-llama3-dpo-v0.2	0.101	bigcode/gpt_bigcode-santacoder	3.92	deepseek-ai/DeepSeek-V2-Lite	0.455
jondurbin/bagel-8b-v1.0	0.164	deepseek-ai/deepseek-coder-6.7b-instruct	3.97	deepseek-ai/ESFT-vanilla-lite	0.456
migtissera/Llama-3-8B-Synthia-v3.5	0.190	Qwen/CodeQwen1.5-7B-Chat	4.09	deepseek-ai/DeepSeek-V2-Lite-Chat	0.868
nvidia/Llama3-ChatQA-1.5-8B	0.191	NTQAI/Nxcode-CQ-7B-orpo	4.11	mistralai/Mistral-7B-Instruct-v0.1	1.356
ruslanmv/Medical-Llama3-8B	0.206	Salesforce/codegen-6B-multi	4.24	statking/zephyr-7b-sft-full-orpo	1.616
FairMind/Llama-3-8B-4bit-UltraChat-Ita	0.296	bigcode/starcoderbase-7b	4.44	Severian/ANIMA-Phi-Neptune-Mistral-7B	1.692
NousResearch/Hermes-2-Theta-Llama-3-8B	0.330	deepseek-ai/deepseek-coder-6.7b-base	4.87	sethuiyer/Medichat-Llama3-8B	1.718

Table 1: Top 10 nearest neighbors among the 1,018 language models for each model listed in the first row. The values indicate the KL divergence measured in bits per byte (bpb), as defined in Section 3.5. These are computed using formula (3) in Section 2, by multiplying the original KL divergence by 0.001484. Tables for the nearest neighbors of the models labeled in the top panel of Fig. 1 are provided in Appendix I.

We find that the distances in our defined coordinate system accurately capture relationships among language models. On this map, each point represents a single model, with those having similar text-generation probability distributions appearing closer together and those with more distinct distributions positioned farther apart. In Section 2, we show that the squared Euclidean distance in this coordinate system approximates the Kullback-Leibler (KL) divergence among models. Table 1 lists the nearest neighbors for each language model. For example, many of the closest neighbors of `meta-llama/Meta-Llama-3-8B` (AI@Meta, 2024) also contain `Llama-3` in their names.

Several studies have explored methods for comparing language models (see Appendix A). In particular, prior work on comparing generated text includes approaches that construct phylogenetic trees based on model-generated text (Yax et al., 2025) and approaches that measure differences in text-generation probabilities conditioned on given prompts using KL divergence (Melamed et al., 2024). However, these methods require generating text with each model, thus incurring the cost of pairwise distance computations, which becomes prohibitively expensive at large scale. By contrast, our method does not involve actual text generation; instead, we compute generation probabilities on a predefined text corpus. This enables us to derive model coordinates without pairwise comparisons, allowing efficient large-scale comparison of many models.

To gain insights from visualizing model attributes on the model map, in Section 4, we analyze various attributes[1] and their relationships. Additionally, by comparing log-likelihood and benchmark performance, we demonstrate the ability to detect data leakage. Then, in Section 5, we treat the log-likelihood vector as a feature and show how it can predict benchmark performance. Finally, in Section 6, we validate the theoretical relationship between model coordinates and KL divergence through experiments.

2 Mapping Language Models into the Space of Text Probability Distributions

In this section, we present our proposed method. Sections 2.2 and 2.3 introduce model feature vectors derived from text-generation probabilities. Section 2.4 demonstrates that the squared Euclidean distance in the coordinate system built using these features approximates the KL divergence between models. Section 2.5 offers an interpretation of the resulting model coordinates. An extension of this method, which defines model coordinates using the sequence of conditional probabilities for generating a given token sequence, is presented in Appendix E.

2.1 Autoregressive language models

Let $\mathcal{X}$ be the set of all possible texts, and let $\mathcal{V}$ be the token vocabulary. A text $x \in \mathcal{X}$ is represented as a sequence of tokens:

$$x = (y_1, \ldots, y_n), \quad y_t \in \mathcal{V}.$$

Denoting the maximum text length by $n_{\max}$, we have $\mathcal{X} = \bigcup_{n=0}^{n_{\max}} \mathcal{V}^n$. We consider a set of K language models $\{p_i\}_{i=1}^K$. With y_0 denoting the beginning-of-sequence (BOS) token, each language model p_i predicts the next token y_t given the preceding token sequence $y^{t-1} = (y_0, \ldots, y_{t-1})$. Thus, the conditional probability defined by p_i is given by

$$y_t \sim p_i(y_t \mid y^{t-1}), \quad t = 1, \ldots, n.$$

[1]Attributes include model type, primary text category, model performance, model size, and creation date.

Accordingly, the probability of a text x under model p_i, denoted $x \sim p_i$, is

$$p_i(x) = \prod_{t=1}^{n} p_i(y_t \mid y^{t-1}).$$

In addition to the K language models $p_1, \ldots, p_K$, we introduce a language model p_0 that represents an underlying distribution for theoretical purposes. We assume we have a dataset (corpus)

$$D = (x_1, x_2, \ldots, x_N) \in \mathcal{X}^N,$$

consisting of N texts, where each text is independently drawn from p_0.

2.2 Log-likelihood vector

For a model p_i, the probability of generating a text x is denoted $p_i(x)$. Following the convention in statistical model selection, we refer to $p_i(x)$ as the likelihood of model p_i given the text x. The log-likelihood $\ell_i(x) = \log p_i(x)$ is then computed as

$$\ell_i(x) = \sum_{t=1}^{n} \log p_i(y_t \mid y^{t-1}).$$

In language model implementations, $-\ell_i(x)$ corresponds to the cross-entropy loss for the text x, and $\exp(-\ell_i(x)/n)$ is known as the perplexity.

Our approach is straightforward. Given that the dataset D consists of N texts, we use the log-likelihood vector

$$\boldsymbol{\ell}_i = (\ell_i(x_1), \ldots, \ell_i(x_N))^\top \in \mathbb{R}^N$$

as the feature vector for model p_i. The first step in our model analysis is to construct the log-likelihood matrix

$$\boldsymbol{L} = (\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_K)^\top \in \mathbb{R}^{K \times N}$$

by stacking the vectors $\boldsymbol{\ell}_i$ for the K models.

2.3 Double centering

As a preprocessing step for model analysis, we apply a technique called double centering (Borg and Groenen, 2005) to $\boldsymbol{L}$. First, we perform row-wise centering. The mean of each row, referred to as the mean log-likelihood, is given by

$$\bar{\ell}_i = \sum_{s=1}^{N} \ell_i(x_s)/N.$$

Subtracting this value from each component of $\boldsymbol{\ell}_i$, we define the centered log-likelihood vector $\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iN})^\top \in \mathbb{R}^N$, where

$$\xi_{is} := \ell_i(x_s) - \bar{\ell}_i, \quad s = 1, \ldots, N.$$

Next, we apply column-wise centering to the matrix of centered feature vectors $(\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)^\top$. The mean vector is $\bar{\boldsymbol{\xi}} = \frac{1}{K}\sum_{i=1}^{K} \boldsymbol{\xi}_i$, and by subtracting this vector from each $\boldsymbol{\xi}_i$, we define the double-centered log-likelihood vector

$$\boldsymbol{q}_i = \boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}.$$

For further details, see Appendices B and C.

2.4 Kullback-Leibler divergence

The Kullback-Leibler (KL) divergence is often used to measure how far apart two models p_i and p_j are in the space of probability distributions[2]. It is defined as

$$\begin{aligned}\mathrm{KL}(p_i, p_j) &= \sum_{x \in \mathcal{X}} p_i(x) \log \frac{p_i(x)}{p_j(x)} \\ &= \mathbb{E}_{x \sim p_i}\big(\ell_i(x) - \ell_j(x)\big). \quad (1)\end{aligned}$$

We assume the dataset D is generated from an unknown underlying model p_0 and that the models p_i and p_j provide good approximations of p_0. Under this assumption, the KL divergence can be approximated as follows:

$$2\,\mathrm{KL}(p_i, p_j) \approx \mathrm{Var}_{x \sim p_0}\big(\ell_i(x) - \ell_j(x)\big). \quad (2)$$

While the definition of KL divergence in (1) involves the expectation of $\ell_i(x) - \ell_j(x)$, the approximation in (2) takes the form of a variance. This result is somewhat surprising yet quite insightful. Notably, although KL divergence is not symmetric in the two models, the approximation in (2) is symmetric. We estimate (2) from the dataset D as

$$2\,\mathrm{KL}(p_i, p_j) \approx \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2/N. \quad (3)$$

Thus, if we regard the model coordinates of p_i as $\boldsymbol{q}_i/\sqrt{N}$ by scaling with N, then the squared Euclidean distance between two points approximates $2\,\mathrm{KL}(p_i, p_j)$.

The main results, namely (2) and (3), are proved in Appendix D using the theory of exponential family of distributions (Barndorff-Nielsen, 2014; Efron, 1978, 2022; Amari, 1982), similar to the

[2] $\mathbb{E}(\cdot)$ denotes expectation and $\mathrm{Var}(\cdot)$ denotes variance.

discussion on the relationship between the norm of embeddings and KL divergence (Oyama et al., 2023). Although the concepts of model map and model coordinates have been discussed in statistics (Shimodaira, 1993; Shimodaira and Cao, 1998; Shimodaira, 2001), and there have been a few applications of model maps (Shimodaira and Hasegawa, 2005; Shimodaira and Terada, 2019), they have received little attention or use in practice.

2.5 Model coordinates

We primarily use $\boldsymbol{q}_i$ as the feature vector of model p_i and refer to it as the model coordinates[3]. As shown in (3), the squared Euclidean distance in the $\boldsymbol{q}$-coordinate system approximates the KL divergence between language models[4], indicating that $\boldsymbol{q}_i$ represents the position of p_i in the space of probability distributions. Since $\boldsymbol{\xi}_i$ differs from $\boldsymbol{q}_i$ only by an offset from the origin, $\boldsymbol{\xi}_i$ also serves as a model coordinate, and $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2$. However, we prefer $\boldsymbol{q}_i$ for its more interpretable components and thus adopt it throughout this paper.

For visualization purposes, we mainly use $\boldsymbol{\ell}_i$ as the coordinates of the model map, as $\boldsymbol{\ell}_i$ can be intuitively interpreted as encoding $\sqrt{N}\,\bar{\ell}_i$ in the "height" dimension and $\boldsymbol{q}_i$ in the "horizontal" dimensions. As shown in Appendix D.6,

$$\|\boldsymbol{\ell}_i - \boldsymbol{\ell}_j\|^2 = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 + N(\bar{\ell}_i - \bar{\ell}_j)^2, \quad (4)$$

which means the squared Euclidean distance in the $\boldsymbol{\ell}$-coordinate system can be decomposed into the sum of $2N\,\mathrm{KL}(p_i, p_j)$ and $N\,(\bar{\ell}_i - \bar{\ell}_j)^2$.

3 Experimental Setup

We describe the key components of our experiment. In particular, Section 3.1 explains the procedure for selecting the 10,000 texts used to compute the model coordinates, and Section 3.2 discusses how we selected the 1,018 language models. Further details are given in Appendix G.

3.1 Selection of text data

The texts used for computing the language models' coordinates were extracted from the Pile (Gao et al., 2020), with five categories of copyrighted material removed[5]. This yielded a dataset D consisting of

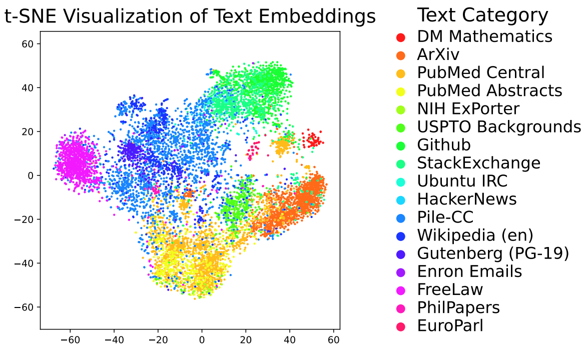

Figure 2: Text embeddings for 10,000 texts in dataset D, computed via `simcse-roberta-large` (Gao et al., 2021b) and visualized with t-SNE. Colors indicate 17 text categories.

10,000 texts, each tagged with a category label from the Pile. Figure 2 visualizes these texts.

To build the dataset, we began by dividing the first 1M texts from the Pile Uncopyrighted corpus into 1,024-byte chunks (UTF-8 encoded). In cases where decoding errors occurred, we truncated by one byte at a time. Chunks smaller than 256 bytes were discarded, resulting in about 5.7M valid chunks. From these, we randomly sampled 10,000 texts to create the final dataset used for computing model coordinates. The average length of these 10,000 texts was 972.3188 bytes.

3.2 Selection of language models

We used $K = 1{,}018$ language models in total. Of these, 1,000 were selected from models listed on Open LLM Leaderboard v1. Specifically, we considered CausalLM models ranging from 1B to 13B parameters and ranked them by their number of downloads over the 30 days preceding February 1, 2025[6]. We initially selected the top 1,100 models by download count and attempted log-likelihood calculations. Among these, 1,011 successfully produced valid log-likelihood values, and we chose the 1,000 most frequently downloaded from that set. In addition, we included 18 models from the DeepSeek language model series. We obtained information on model parameter sizes and architectures from the Leaderboard. Appendix G provides basic information on the selected models and details on how model types were defined. A complete list of models used in this study is given in Appendix L.

[3] We refer to the Euclidean space where the vectors $\boldsymbol{\ell}_i$, $\boldsymbol{\xi}_i$, or $\boldsymbol{q}_i$, possibly scaled, reside as the *log-likelihood space*.

[4] That is, $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 \approx 2N\,\mathrm{KL}(p_i, p_j)$. For simplicity, we omit the constant scaling factor in expressions of this type.

[5] `https://huggingface.co/datasets/monology/pile-uncopyrighted`

[6] Hugging Face's API provides the total number of downloads in the past 30 days.

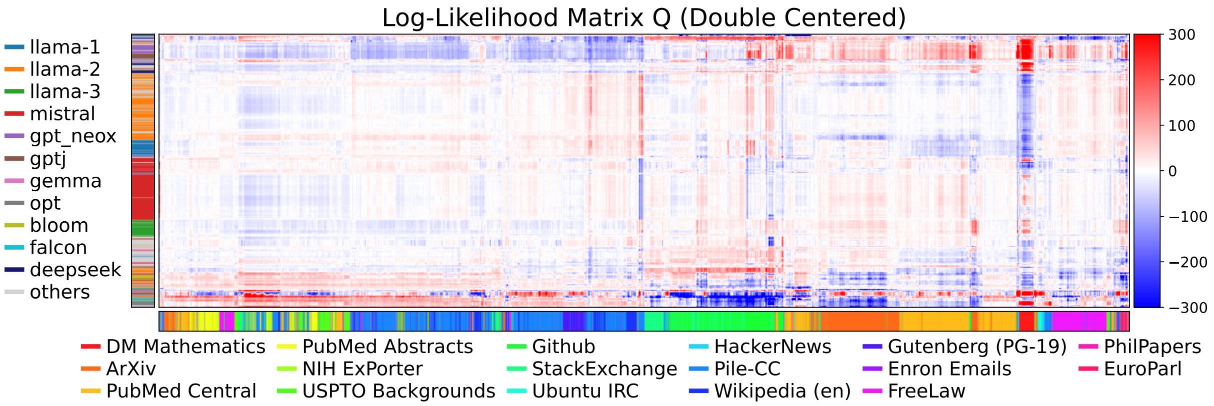

Figure 3: The double-centered log-likelihood matrix $\boldsymbol{Q}$, with rows and columns reordered by hierarchical clustering. Each row corresponds to one of the 1,018 models, color-coded by model type. Each column represents one of the 10,000 texts, color-coded by text category.

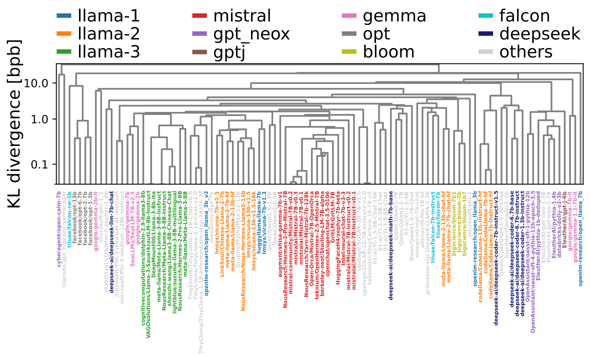

Figure 4: Hierarchical clustering of the top 100 most-downloaded models, based on their feature vectors $\boldsymbol{q}_i$. Model names are color-coded by model type. KL divergence is reported in units of bits per byte (bpb).

3.3 Computation of the log-likelihood

The log-likelihood matrix $\boldsymbol{L}$ was computed in float16 precision, with the bottom 2% of values clipped. This clipping mitigates the large impact of extremely low likelihoods on (3). After computing $\boldsymbol{L}$, we applied row-wise and column-wise centering to obtain the double-centered log-likelihood matrix $\boldsymbol{Q}$. Figure 3 visualizes $\boldsymbol{Q}$. Each value in the matrix can be interpreted in two ways: as the relative probability of a text for each model or as the relative likelihood of a model for each text. Both models and texts exhibit clustering patterns. Figure 4 shows a dendrogram of the top 100 models. We examined the effective dimension from the perspective of feature vector dimensionality reduction and found that the cumulative contribution ratio, based on the sum of squared singular values of $\boldsymbol{Q}$, reached 90% at 42 dimensions and 95% at 82 dimensions.

3.4 Obtaining the leaderboard scores

We obtained benchmark scores for the language models used in our experiments from Open LLM Leaderboard v1[7]. Between April 2023 and June 2024, this leaderboard evaluated language models on six tasks: AI2 Reasoning Challenge (ARC) (Clark et al., 2018), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021a), TruthfulQA (Lin et al., 2022a), Winogrande (Sakaguchi et al., 2019), and GSM8K (Cobbe et al., 2021). Along with individual task scores, we also use the average score across all six tasks, referred to as 6-TaskMean.

3.5 Byte-normalized KL divergence for cross-experiment comparison

The KL divergence of text generation, $\mathrm{KL}(p_i, p_j)$, generally increases with the number of tokens in the text. As a result, it cannot be directly compared with the KL divergence from other experiments using different text data. For models that use the same tokenizer, one can normalize the KL divergence by the average number of tokens in the text to obtain the KL divergence per token. However, when comparing KL divergence between models with different tokenizers, as in this study, it is more appropriate to normalize by the average text length in bytes and use the KL divergence per byte. For instance, if $\mathrm{KL}(p_i, p_j) = 1{,}000$, dividing by the average text length of 972.3188 bytes yields a KL divergence per byte of $1{,}000/972.3 = 1.028$ nats $= 1.484$ bits[8].

[7] https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard

[8] To convert the code length unit from nats to bits, multiply by $1/\log 2 = 1.4427$.

KL divergence can be understood from the perspective of coding theory as a measure of how much longer a message becomes when encoded using an incorrect probability distribution. $\mathrm{KL}(p_i, p_j)$ represents the extra code length required when encoding text generated from probability distribution p_i using a different distribution p_j. Dividing this value by the average text length in bytes gives the additional code length per byte. In the example above, this means that, due to differences between the models, an extra 1.484 bits are needed to encode each byte of text.

4 Map of Language Models

We applied t-SNE (van der Maaten and Hinton, 2008) to the log-likelihood matrix $\boldsymbol{L}$ for dimensionality reduction[9]. Using this visualization, we analyze the insights gained from the model map in this section. While this paper presents model maps using $\boldsymbol{L}$, alternative maps using the double-centered log-likelihood matrix $\boldsymbol{Q}$, as well as a model map with labels for all language models, are available in Appendix H.

4.1 Visualizing attributes on the model map

Model type. The top panel of Fig. 1 visualizes the distribution of model types, with each model color-coded according to its type. We observe that models belonging to the same type tend to cluster together, forming distinct regions on the map (e.g., `llama-2`, `mistral`, and `gemma`). In particular, models optimized for coding tasks[10] appear in a relatively compact region, suggesting that these models share notable similarities in their probability distributions.

Text category. The bottom panel of Fig. 1 shows the model map with each model color-coded according to the text category in which it achieves the highest standardized log-likelihood score (Appendix G.4). From this figure, we see that models exhibiting high likelihoods for the same text category are grouped together. Notably, the cluster containing coding-specialized models in the top panel aligns with the GitHub/StackExchange region in the bottom panel, suggesting that these models have relatively high likelihoods for text originating from GitHub and StackExchange.

Model performance. Figure 5 visualizes two evaluation metrics: mean log-likelihood and benchmark task performance. From the left and central panels, we see that both metrics exhibit similar trends on the map, where models that lie close together tend to show similar metric values. Additionally, in the right panel, the GSM8K/MMLU region corresponds to the ArXiv/PubMed Central region in the bottom panel of Fig. 1, suggesting that models with high likelihoods on academic and scientific texts also tend to perform well on mathematical reasoning and academic knowledge-intensive tasks.

Model size and creation date. Figure 6 shows the distribution of models by size and creation date. Compared to Fig. 5, newer models generally perform better, but model size does not always correlate with performance, as some smaller models perform comparably to larger ones.

4.2 Detection of data leakage

Since the text data we used was extracted from the Pile corpus, models that were pre-trained on the Pile are likely to exhibit higher log-likelihood values than their actual capabilities measured by benchmarks. We analyze this effect using the model map in Fig. 7. The left panel highlights models that used the Pile for pre-training. The right panel shows models with high mean log-likelihood relative to their 6-TaskMean score. The alignment between these two distributions suggests that models pre-trained on the Pile tend to achieve higher likelihoods on our text data, while their benchmark performance remains comparatively lower.

5 Predicting Model Performance from Model Coordinates

As shown in the central panel of Fig. 5, the positioning of models on the map suggests that a model's benchmark performance may be inferred from its coordinates. In this section, we conduct a regression analysis using the $\boldsymbol{q}$-coordinates to predict benchmark scores and evaluate predictive performance.

5.1 Benchmark scores and models

We use six benchmark scores from Open LLM Leaderboard v1, as described in Section 3.4. Our

[9] The perplexity value was set to 30, and we used the scikit-learn implementation (Varoquaux et al., 2015).

[10] For example, `starcoder2-7b` (Lozhkov et al., 2024), `deepseek-coder-1.3b-base` (Guo et al., 2024), `codegemma-2b` (CodeGemma Team et al., 2024) and `CodeLlama-7b-Instruct-hf` (Rozière et al., 2024).

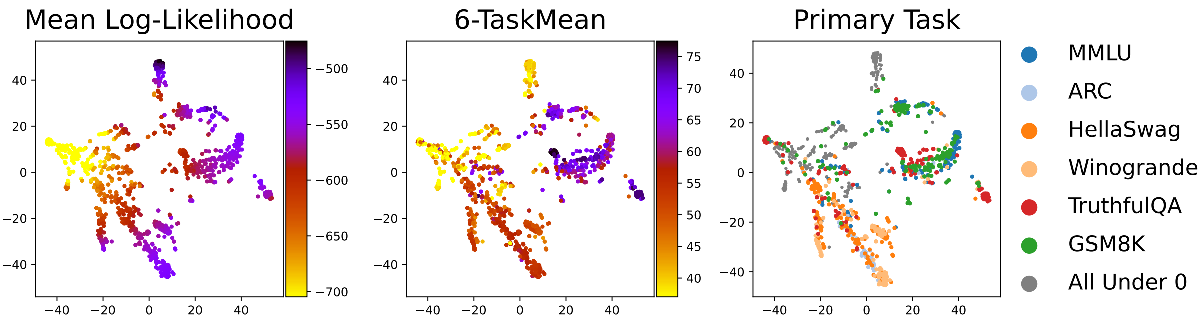

Figure 5: Model maps illustrating model performance. From left to right, the panels show each model's mean log-likelihood, 6-TaskMean score, and the "primary task," meaning the task for which each model achieves its highest standardized score among the six tasks (Appendix G.4). The color bar is clipped at the 10th percentile for mean log-likelihood and 6-TaskMean, with darker colors indicating better performance. In the primary task panel, models with standardized scores below zero on all six tasks are labeled "All Under 0."

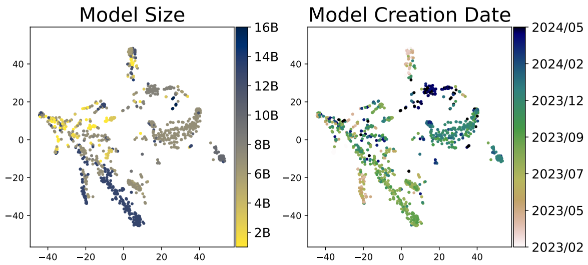

Figure 6: Model maps color-coded by (Left) number of parameters and (Right) model creation date.

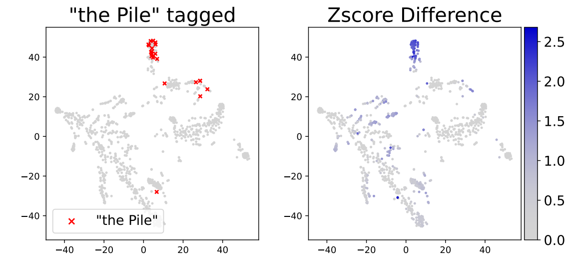

Figure 7: (Left) Models tagged with "the Pile." (Right) Difference between the standardized mean log-likelihood and the standardized 6-TaskMean score.

experiments are conducted on 996 models for which these benchmark scores are available[11].

5.2 Setting for regression analysis

For each benchmark task, the dataset is given as $\{(\boldsymbol{q}_1, v_1), \ldots, (\boldsymbol{q}_K, v_K)\}$, where $\boldsymbol{q}_i \in \mathbb{R}^N$ is the double-centered log-likelihood vector of the language model p_i, and $v_i \in [0, 100]$ is its corresponding benchmark score. We use ridge regression to predict each benchmark score. Let $\boldsymbol{Q} \in \mathbb{R}^{K \times N}$ be the matrix of explanatory variables, and let $\boldsymbol{v} = (v_1, \ldots, v_K)^\top \in \mathbb{R}^K$ be the vector for the target variable. The objective function with parameter $\boldsymbol{w} \in \mathbb{R}^N$ is given by:

$$\mathcal{L}(\boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{Q}\boldsymbol{w}\|^2 + \alpha\|\boldsymbol{w}\|^2, \quad (5)$$

where $\alpha \in \mathbb{R}_{>0}$ is a hyperparameter that controls the strength of regularization. Since the number of variables N is much larger than the sample size K ($N \gg K$), making this a high-dimensional regression setting, we carefully set α using cross-validation to avoid overfitting.

We partition the models into five folds based on model types and perform parameter training and benchmark score prediction[12]. To mitigate the effect of randomness, we repeat the data splitting with five different seeds and take the average of the predictions as the final predicted score. As evaluation metrics, we compute Pearson's r and Spearman's ρ to measure the correlation between the predicted and benchmark scores. Additionally, we conduct experiments by replacing the target variable with 6-TaskMean and mean log-likelihood, leading to a total of eight experimental settings. See Appendix K.1 for details.

5.3 Results and discussion

Regression analysis. Table 2 summarizes the results of the regression analysis using q-coordinates. Across all benchmark tasks, both Pearson's r and

[11] From the 1,018 models, we excluded four that lacked TruthfulQA scores and 18 from `deepseek-ai`, as their benchmark scores were incomplete due to the models being relatively new, leaving 996 for analysis. For simplicity, we denote the number of models as K.

[12] We used the `GroupKFold` and `RidgeCV` implementations provided by scikit-learn (Varoquaux et al., 2015).

	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	6-TaskMean	mean log-likelihood
Pearson's r	0.946	0.909	0.932	0.901	0.941	0.884	0.953	0.989
Spearman's ρ	0.948	0.956	0.934	0.884	0.948	0.857	0.960	0.974

Table 2: Results of ridge regression for predicting benchmark scores from model coordinates. Predictions for 6-TaskMean and mean log-likelihood are also included. High correlation coefficients are observed across all settings. See Table 7 in Appendix K for the mean and standard deviation of correlation coefficients across the five data splits.

	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	6-TaskMean
Pearson's r	0.453	0.598	0.346	0.072	0.508	0.239	0.395
Spearman's ρ	0.432	0.467	0.422	0.048	0.512	0.364	0.400

Table 3: Correlation between mean log-likelihoods and benchmark scores. The results show that perplexity has only moderate correlations on most tasks, especially compared to model coordinates in Table 2.

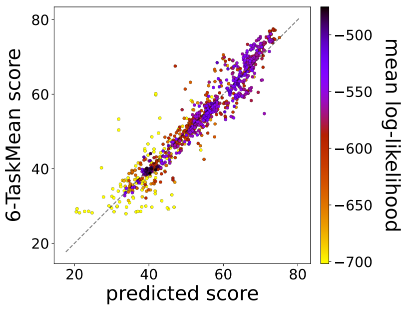

Figure 8: Scatter plot comparing predicted and actual 6-TaskMean scores for test sets (the dashed line indicates the identity line). Pearson's r is 0.953 and Spearman's ρ is 0.960, indicating strong predictive performance. Points are color-coded by the mean log-likelihood, with the color scale clipped at the 10th percentile. While higher mean log-likelihood values tend to correspond to higher benchmark scores, a few models with unusually high log-likelihoods due to data leakage (Section 4.2) deviate from this trend. Nevertheless, the predictions remain accurate overall. Scatter plots for individual benchmark tasks are shown in Fig. 18 in Appendix K.2.

Spearman's ρ are consistently high, indicating that the ridge regression model based on model coordinates achieves strong predictive performance. This trend holds not only for individual tasks but also for the 6-TaskMean, where the correlation coefficients remain high, as illustrated in the scatter plot in Fig. 8. Even when the target variable is the mean log-likelihood, defined as the simple average over the 10,000 texts, the regression model still achieves high correlation. This suggests that the model coordinates, despite being double-centered and excluding direct information about the mean, retain a sufficiently rich structure to accurately predict the average log-likelihood.

Performance prediction based on perplexity. The mean log-likelihood equals $-\log(\text{perplexity})$ and is often used as a performance indicator. We evaluated its correlation with benchmark scores, and Table 3 shows moderate correlations across all tasks, consistent with prior findings (Liu et al., 2023a; Fang et al., 2025; Thrush et al., 2025). Unlike mean log-likelihood, which uniformly averages over texts, regression using the log-likelihood matrix $\boldsymbol{L}$ can assign task-dependent weights. As shown in Table 9 in Appendix K, its prediction accuracy is nearly the same as that of $\boldsymbol{Q}$. This indicates that combining $\boldsymbol{Q}$ and the mean in $\boldsymbol{L}$ offers little improvement. Accurate prediction with $\boldsymbol{Q}$ alone suggests that model positions already encode a structure aligned with benchmark performance.

6 Empirical Validation of Theory

In Section 2, we discussed model coordinates for text probability models. Appendix E extends this framework to token-sequence probability models. These theoretical results state that the squared Euclidean distance in the log-likelihood space approximates the KL divergence between models. To validate this, we first evaluate token-level conditional probability models (Section 6.1), since token-level experiments are generally easier to conduct than text-level experiments. We then extend our analysis to text probability models (Section 6.2). Additionally, in Section 6.3, we explore the relationship between model weight parameters and model coordinates.

6.1 Validation of (32) for token-level models

Settings. Two models with a shared tokenizer `Llama-2-7b-hf` and `Llama-2-7b-chat-hf` (Touvron et al., 2023b), denoted as p_1 and p_2, were used. Following the method described in Appendix E.1, we computed the model coordinates $\zeta_1(x)$ and $\zeta_2(x)$ for each text $x = (y_1, \cdots, y_n) \in D$, where these coordinates are centered vectors with elements $\log p_i(y_t|y^{t-1})$ as defined in (28). We then calculated the squared Euclidean distance between these coordinates, $\|\zeta_1(x) - \zeta_2(x)\|^2$. To obtain the exact KL divergence between models, we used the outputs of the softmax function in the language models. We computed the sum of per-token KL divergences: $\sum_{t=1}^{n} \mathrm{KL}(p_1(y_t|y^{t-1}), p_2(y_t|y^{t-1}))$, which is also used in Lv et al. (2023).

Results and discussion. The left panel of Fig. 9 shows a scatter plot of the squared Euclidean distance and KL divergence for $x \in D$, with a Pearson's correlation coefficient of $r = 0.893$. This result indicates that (32) provides a good approximation in actual language models.

6.2 Validation of (3) for text-level models

Settings. Among the 292 language models sharing the tokenizer with `Llama-2-7b-hf` (Touvron et al., 2023b), we excluded the five models with the largest values of $\sum_{x \in D} \|\zeta_i(x)\|^2$, leaving 287 models for our experiment. Using the approach described in Section 2, we computed the model coordinates for each model and calculated the squared Euclidean distance $\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2$ between every pair of models. Because it is extremely difficult to directly compute $\mathrm{KL}(p_i, p_j)$, we instead used $\frac{1}{N} \sum_{x \in D} \left\|\zeta_i(x) - \zeta_j(x)\right\|^2$ as a proxy. For a theoretical justification that this quantity approximates $\mathrm{KL}(p_i, p_j)$, see (33) in Appendix E.

Results and discussion. As shown in the right panel of Fig. 9, the scatter plot of squared Euclidean distance versus KL divergence exhibits a Pearson's correlation coefficient of $r = 0.904$. This finding confirms that the relationship in (3) holds approximately in practical language models.

6.3 Relationship between model weights and model coordinates

For language models with the same architecture, comparison can also be conducted via weight parameters. We investigated how the structure of the log-likelihood space aligns with the structure of the weight-parameter space. We generated 36 new language models by linearly interpolating the weights of `Llama-2-7b-hf`, `Llama-2-7b-chat-hf` (Touvron et al., 2023b), and `vicuna-7b-v1.5` (Zheng et al., 2023), using a 6×6 grid of coefficients. For these 36 models, we computed text-generation log-likelihoods and visualized the resulting model coordinates in two dimensions using Principal Component Analysis (PCA).

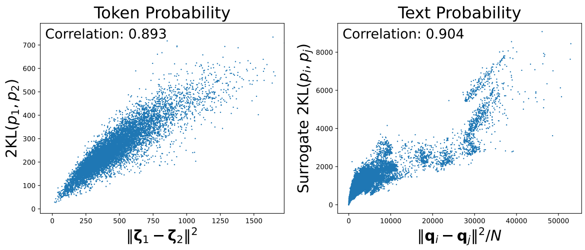

Figure 9: Relationship between the squared Euclidean distance of model coordinates and KL divergence. (Left) In the token-level experiment (Section 6.1), each point represents a text. (Right) In the text-level experiment (Section 6.2), each point represents a pair of models.

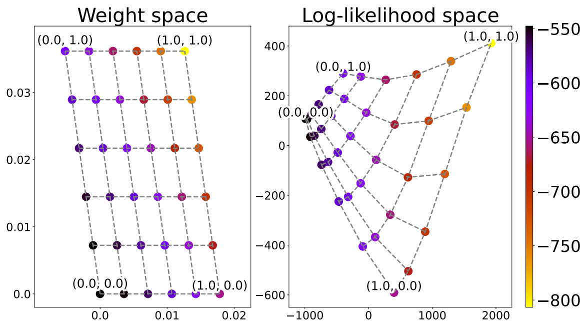

Figure 10: Visualization of 36 language models obtained by linearly interpolating pretrained model weights based on `Llama-2-7b-hf`. Each point is color-coded according to its mean log-likelihood. (Left) Models in the weight parameter space. (Right) Models in the log-likelihood space, represented by the $\boldsymbol{q}$-coordinate system.

Figure 10 shows these 36 models in both weight space and log-likelihood space, where the two-dimensional grid structure is preserved. Interestingly, the mean log-likelihoods of the interpolated models closely follow the linear interpolation of the three base models' mean log-likelihoods. This suggests that such interpolation can be a practical tool for exploring high-performing models. See Appendix J for details.

7 Conclusion

We proposed a method to compare autoregressive language models using log-likelihoods on a predefined text set. By interpreting these as model coordinates, we showed that the squared Euclidean distance approximates KL divergence. Experiments on over 1,000 models confirmed the method's effectiveness for analyzing model relationships, predicting benchmark performance, and validating the theory.

Limitations

- Changing the text data will alter the analysis results of the model map. This is both a limitation and an advantage of the proposed method, because it allows us to choose text data according to the analysis objective. For example, if we want to investigate code-focused language models in more detail, we can increase the proportion of code data from GitHub, thereby increasing the resolution of code-focused models on the model map.

- The proof (Section 2 and Appendix D) that the squared Euclidean distance in the model coordinate system approximates the KL divergence assumes that the language model's text-generation probabilities closely match the distribution of the text data. When this assumption does not hold, the approximation accuracy decreases. However, even under such circumstances, the model coordinates should still function sufficiently as model features.

- If the text data used for the model coordinates is contained in a language model's pre-training corpus, that model's mean log-likelihood may be overestimated. This data leakage, or data contamination, is generally non-negligible, as shown in Fig. 7 in Section 4.2, which illustrates the effect of using the Pile corpus. However, comparing it against benchmark scores makes it possible to detect such data leakage, and one can remove models that are affected. Furthermore, the model map based on the $\boldsymbol{q}$-coordinate system is robust to contamination in the data, as the estimation of KL divergence using the squared Euclidean distance in the $\boldsymbol{q}$-coordinate system remains valid even if the true generative model p_0 that produces the dataset D varies, as long as all compared models remain sufficiently close to p_0. For instance, even if D is generated from one of the K models, such as p_i, the estimation formula (3) remains correct (Appendix D.7).

- Beyond the data leakage mentioned above, other systematic errors introduced into the model coordinates can also affect the model map. As noted in Appendix C, $\boldsymbol{\ell}_i$ and $\bar{\ell}_i$ are susceptible to systematic biases. However, thanks to double centering, $\boldsymbol{q}_i$ is less influenced by bias terms.

- Although the calculation of model coordinates is linear time $O(KN)$ in the number of models K and the number of texts N, it still requires a non-negligible amount of computation. In our experiments, it took about 10 minutes on a single GPU (RTX 6000 Ada) to compute coordinates ($N = 10^4$, float16) for a single 7B model.

- Computing the model map visualization from the model coordinates is generally not linear in K. For example, t-SNE requires a distance matrix, incurring $O(K^2N)$ computational cost. However, in modern computing environments, as long as K is not extremely large (e.g., in the millions), the cost of visualization is negligible compared to the cost of calculating the model coordinates.

- A sufficiently large number of text samples, N, is desirable for the model coordinates. We used $N = 10^4$. Since the error in the KL divergence estimate due to randomness decreases proportionally to $N^{-1/2}$, N must be increased according to the desired resolution of the model map.

- When using model coordinates as feature vectors, $N = 10^4$ can be unwieldy. According to the experiment in Section 3.3, applying PCA to reduce the dimensionality of $\boldsymbol{q}$-coordinates to 82 dimensions still retains 95% of the information. However, the predictive performance of such dimension-reduced features has not yet been tested.

- Since the language models used in our experiments were obtained from the Open LLM Leaderboard v1 (which ran from April 2023 to June 2024), our discussion of models released after June 2024 is limited.

- The results of the task-performance prediction in Section 5 should be interpreted conservatively. We employed a proper cross-validation, splitting the models based on their model types so that the training, validation, and test sets were disjoint, thereby limiting data leakage. However, if nearly identical models appear in both the training and test partitions (for example, models labeled with different types but built on the same base model and modified only slightly by fine-tuning), prediction becomes artificially easier. As shown in Tables 7 and 8 in Appendix K.3, random splits that permit such leakage yield higher predictive performance than splits that strictly follow model-type groupings. Finally, note that models not listed on the leaderboard were excluded from our evaluation.

- The theoretical validation experiment in Section 6 is limited. Currently, it is difficult to directly compute exact KL divergence values among text-generation probability models, so conducting more precise validation experiments remains a future challenge.

- In the method of computing model coordinates from token-sequence conditional probabilities (Appendix E and Appendix F), the proof that the squared Euclidean distance in the model coordinate system approximates the KL divergence requires additional assumptions (Appendix F.3). In practice, Assumption 2 does not hold, and due to the variations in (39), $\|\zeta_i - \zeta_j\|^2$ in (32) tends to overestimate the KL divergence. Nonetheless, even in such situations, token-level model coordinates will still likely function sufficiently as features.

Acknowledgments

This study was partially supported by JSPS KAKENHI 22H05106, 23H03355, JST CREST JPMJCR21N3, JST BOOST JPMJBS2407.

References

01. AI, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Wen Xie, and 13 others. 2025. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

42dot Inc. 2023. 42dot llm: A series of large language model by 42dot.

Rishiraj Acharya. 2023. Catppt.

AI@Meta. 2024. Llama 3 model card.

Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. *Preprint*, arXiv:2305.13245.

AI@Waktaverse. 2024. Waktaverse llama 3 model card.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Duarte M. Alves, José Pombal, Nuno M. Guerreiro, Pedro H. Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André F. T. Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *Preprint*, arXiv:2402.17733.

Shun-Ichi Amari. 1982. Differential Geometry of Curved Exponential Families-Curvatures and Information Loss. *The Annals of Statistics*, 10(2):357 – 385.

Shun-Ichi Amari. 1998. Natural gradient works efficiently in learning. *Neural computation*, 10:251–276.

Yuvanesh Anand, Zach Nussbaum, Brandon Duderstadt, Benjamin Schmidt, and Andriy Mulyar. 2023. Gpt4all: Training an assistant-style chatbot with large scale data distillation from gpt-3.5-turbo.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. 2021. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *Preprint*, arXiv:2108.07732.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. 2024. Llemma: An open language model for mathematics. *Preprint*, arXiv:2310.10631.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *Preprint*, arXiv:1607.06450.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Abhinand Balachandran. 2023. Tamil-llama: A new tamil language model based on llama 2. *Preprint*, arXiv:2311.05845.

Ole Barndorff-Nielsen. 2014. *Information and exponential families: in statistical theory*. John Wiley & Sons.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *Preprint*, arXiv:2312.09993.

Mohammad Bavarian, Heewoo Jun, Nikolas Tezak, John Schulman, Christine McLeavey, Jerry Tworek, and Mark Chen. 2022. Efficient training of language models to fill in the middle. *Preprint*, arXiv:2207.14255.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. 2023. Open llm leaderboard.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *Preprint*, arXiv:2004.05150.

Leonard Bereska and Stratis Gavves. 2024. Mechanistic interpretability for AI safety - a review. *Transactions on Machine Learning Research*. Survey Certification, Expert Certification.

Stella Biderman, Kieran Bicheno, and Leo Gao. 2022. Datasheet for the pile. *Preprint*, arXiv:2201.07311.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. *Preprint*, arXiv:2304.01373.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *Preprint*, arXiv:1911.11641.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. Gpt-neox-20b: An open-source autoregressive language model. *Preprint*, arXiv:2204.06745.

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Ingwer Borg and Patrick J. F. Groenen. 2005. *Modern Multidimensional Scaling: Theory and Applications*, 2 edition. Springer Series in Statistics. Springer, New York, NY.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

CeADAR. 2023. Financeconnect-13b (revision 5f7841d).

Sahil Chaudhary. 2023. Code alpaca: An instruction-following llama model for code generation.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. 2024a. Alpagasus: Training a better alpaca with fewer data. *Preprint*, arXiv:2307.08701.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. 2023a. Symbolic discovery of optimization algorithms. *Preprint*, arXiv:2302.06675.

Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024b. Longlora: Efficient fine-tuning of long-context large language models. *Preprint*, arXiv:2309.12307.

Yukang Chen, Shaozuo Yu, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2023b. Long alpaca: Long-context instruction-following models.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *Preprint*, arXiv:2403.04132.

chiliu. 2023. Mamba-gpt-3b-v4.

François Chollet. 2019. On the measure of intelligence. *Preprint*, arXiv:1911.01547.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *Preprint*, arXiv:1905.10044.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

CodeGemma Team, Heri Zhao, Jeffrey Hui, Joshua Howland, Nam Nguyen, Siqi Zuo, Andrea Hu, Christopher A. Choquette-Choo, Jingyue Shen, Joe Kelley, Kshitij Bansal, Luke Vilnis, Mateo Wirth, Paul Michel, Peter Choy, Pratik Joshi, Ravin Kumar, Sarmad Hashmi, Shubham Agrawal, and 8 others. 2024. Codegemma: Open code models based on gemma. *Preprint*, arXiv:2406.11409.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with scaled ai feedback. *Preprint*, arXiv:2310.01377.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. 2024. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *Preprint*, arXiv:2401.06066.

Alpin Dale, Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Llongorca13b: Llama2-13b model instruct-tuned for long context on filtered openorcav1 gpt-4 dataset.

Tri Dao. 2023. FlashAttention-2: Faster attention with better parallelism and work partitioning.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Preprint*, arXiv:2205.14135.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, Guillaume Desjardins, Arnaud Doucet, David Budden, Yee Whye Teh, Razvan Pascanu, Nando De Freitas, and Caglar Gulcehre. 2024. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *Preprint*, arXiv:2402.19427.

DeciAI Research Team. 2023. Decilm-7b-instruct.

DeepSeek-AI, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, and 68 others. 2024a. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fuli Luo, Guangbo Hao, Guanting Chen, and 138 others. 2024b. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *Preprint*, arXiv:2405.04434.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *Preprint*, arXiv:2110.02861.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

Peter Devine. 2024. Tagengo: A multilingual chat dataset. *Preprint*, arXiv:2405.12612.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. *Preprint*, arXiv:2101.11718.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

Duy Quang Do, Hoang Le, and Duc Thang Nguyen. 2023. Torolama: The vietnamese instruction-following and chat model.

Bradley Efron. 1978. The geometry of exponential families. *The Annals of Statistics*, 6:362–376.

Bradley Efron. 2022. *Exponential Families in Theory and Practice*. Cambridge University Press.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and Douwe Kiela. 2023. Human-centered loss functions (halos).

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2025. What is wrong with perplexity for long-context language modeling? In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. 2024. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`.

Victor Gallego. 2024. Configurable safety tuning of language models with synthetic preference data. *Preprint*, arXiv:2404.00495.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021a. A framework for few-shot language model evaluation.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021b. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *Preprint*, arXiv:2009.11462.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Xinyang Geng and Hao Liu. 2023. Openllama: An open reproduction of llama.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2024. Tora: A tool-integrated reasoning agent for mathematical problem solving. *Preprint*, arXiv:2309.17452.

Diego Granziol, Stefan Zohren, and Stephen Roberts. 2021. Learning rates as a function of batch size: A random matrix theory approach to neural network training. *Preprint*, arXiv:2006.09092.

Griffin Team, Soham De, Samuel L Smith, Anushan Fernando, Alex Botev, George-Christian Muraru, Ruba Haroun, and Leonard Berrada et al. 2024. Recurrentgemma.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, and 24 others. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence. *Preprint*, arXiv:2401.14196.

Jan Philipp Harries. 2023. orca_mini_v2_ger_7b: An explain tuned llama-7b model based on orca mini v2 and adapted to german language.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *Preprint*, arXiv:2203.09509.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Alex Henry, Prudhvi Raj Dachapally, Shubham Pawar, and Yuxuan Chen. 2020. Query-key normalization for transformers. *Preprint*, arXiv:2010.04245.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *Preprint*, arXiv:2403.07691.

Eliahu Horwitz, Nitzan Kurer, Jonathan Kahana, Liel Amar, and Yedid Hoshen. 2025. Charting and navigating hugging face's model atlas. *Preprint*, arXiv:2503.10633.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. *Preprint*, arXiv:1902.00751.

Chan-Jan Hsu, Chang-Le Liu, Feng-Ting Liao, Po-Chun Hsu, Yi-Chang Chen, and Da-Shan Shiu. 2024. Breeze-7b technical report. *Preprint*, arXiv:2403.02712.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

IDEA-CCNL. 2021. Fengshenbang-lm.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*.

"interstellarninja", "Teknium", "theemozilla", "karan4d", and "huemin_art". 2024. Hermes-2-pro-mistral-7b.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *Preprint*, arXiv:2212.12017.

Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. 2023. Neftune: Noisy embeddings improve instruction finetuning. *Preprint*, arXiv:2310.05914.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Dongfu Jiang, Yishan Li, Ge Zhang, Wenhao Huang, Bill Yuchen Lin, and Wenhu Chen. 2024. Tigerscore: Towards building explainable metric for all text generation tasks. *Preprint*, arXiv:2310.00752.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Preprint*, arXiv:1705.03551.

Dahyun Kim, Yungi Kim, Wonho Song, Hyeonwoo Kim, Yunsu Kim, Sanghoon Kim, and Chanjun Park. 2024a. sdpo: Don't use your data all at once. *Preprint*, arXiv:2403.19270.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2024b. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling. *Preprint*, arXiv:2312.15166.

Dohyeong Kim, Myeongjun Jang, Deuk Sin Kwon, and Eric Davis. 2022. Kobest: Korean balanced evaluation of significant tasks. *Preprint*, arXiv:2204.04541.

Seungduk Kim, Seungtaek Choi, and Myeongho Jeong. 2024c. Efficient and effective vocabulary expansion towards multilingual large language models. *Preprint*, arXiv:2402.14714.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, Sungho Park, and Kyubyong Park. 2023. A technical report for polyglot-ko: Open-source large-scale korean language models. *Preprint*, arXiv:2306.02254.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Aran Komatsuzaki, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2023. Sparse upcycling: Training mixture-of-experts from dense checkpoints. *Preprint*, arXiv:2212.05055.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. *Preprint*, arXiv:2309.00237.

Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. 2023. Openassistant conversations – democratizing large language model alignment. *Preprint*, arXiv:2304.07327.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *Preprint*, arXiv:2402.10373.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *Preprint*, arXiv:1910.09700.

Ariel N. Lee, Cole J. Hunter, and Nataniel Ruiz. 2024. Platypus: Quick, cheap, and powerful refinement of llms. *Preprint*, arXiv:2308.07317.

Ariel N. Lee, Cole J. Hunter, Nataniel Ruiz, Bleys Goodson, Wing Lian, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Openorcaplatypus: Llama2-13b model instruct-tuned on filtered openorcav1 gpt-4 dataset and merged with divergent stem and logic dataset model.

Junbum Lee. 2023. llama-2-ko-7b (revision 4a9993e).

Junbum Lee. 2024a. Llama-3-koen.

Junbum Lee. 2024b. Yi-ko-6b (revision 205083a).

Jungwon Lee and Seungjun Ahn. 2024. Llama-3-instruction-constructionsafety-layertuning.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. 2020. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *Preprint*, arXiv:2005.01643.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, and 48 others. 2023a. Starcoder: may the source be with you! *Preprint*, arXiv:2305.06161.

Shenggui Li, Hongxin Liu, Zhengda Bian, Jiarui Fang, Haichen Huang, Yuliang Liu, Boxiang Wang, and Yang You. 2023b. Colossal-ai: A unified deep learning system for large-scale parallel training. *Preprint*, arXiv:2110.14883.

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023c. Textbooks are all you need ii: phi-1.5 technical report. *Preprint*, arXiv:2309.05463.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023d. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Wing Lian, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023a. Openorca_preview1: A llama-13b model fine-tuned on small portion of openorcav1 dataset.

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023b. Jackalope 7b: Mistral-7b model multi-turn chat tuned on filtered openorcav1 gpt-4 dataset.

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023c. Llongorca7b: Llama2-7b model instruct-tuned for long context on filtered openorcav1 gpt-4 dataset.

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023d. Mistralorca: Mistral-7b model instruct-tuned on filtered openorcav1 gpt-4 dataset.

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023e. Mistralslimorca: Mistral-7b model instruct-tuned on filtered, corrected, openorcav1 gpt-4 dataset.

Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023f. Slimorca: An open dataset of gpt-4 augmented flan reasoning traces, with verification.

Wing Lian, Guan Wang, Bleys Goodson, Eugene Pentland, Austin Cook, Chanvichet Vong, "Teknium", and Nathan Hoos. 2023g. Slimorca dedup: A deduplicated subset of slimorca.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022a. Truthfulqa: Measuring how models mimic human falsehoods. *Preprint*, arXiv:2109.07958.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022b. Few-shot learning with multilingual language models. *Preprint*, arXiv:2112.10668.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2025. World model on million-length video and language with blockwise ringattention. *Preprint*, arXiv:2402.08268.

Hong Liu, Sang Michael Xie, Zhiyuan Li, and Tengyu Ma. 2023a. Same pre-training loss, better downstream: Implicit bias matters for language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 22208–22242. PMLR.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Zihan Liu, Wei Ping, Rajarshi Roy, Peng Xu, Chankyu Lee, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Chatqa: Surpassing gpt-4 on conversational qa and rag. *Preprint*, arXiv:2401.10225.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. *Preprint*, arXiv:2301.13688.

Anton Lozhkov, Raymond Li, Loubna Ben Allal, Federico Cassano, Joel Lamy-Poirier, Nouamane Tazi, Ao Tang, Dmytro Pykhtar, Jiawei Liu, Yuxiang Wei, Tianyang Liu, Max Tian, Denis Kocetkov, Arthur Zucker, Younes Belkada, Zijian Wang, Qian Liu, Dmitry Abulkhanov, Indraneil Paul, and 47 others. 2024. Starcoder 2 and the stack v2: The next generation. *Preprint*, arXiv:2402.19173.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Jianguang Lou, Chongyang Tao, Xiubo Geng, Qingwei Lin, Shifeng Chen, Yansong Tang, and Dongmei Zhang. 2025. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. *Preprint*, arXiv:2308.09583.

Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang. 2023. Wizardcoder: Empowering code large language models with evol-instruct. *Preprint*, arXiv:2306.08568.

Xingtai Lv, Ning Ding, Yujia Qin, Zhiyuan Liu, and Maosong Sun. 2023. Parameter-efficient weight ensembling facilitates task-level knowledge transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 270–282, Toronto, Canada. Association for Computational Linguistics.

Manuel Romero. 2023. llama-2-coder-7b (revision d30d193).

Pankaj Mathur. 2023a. orca_mini_v2_7b: An explain tuned llama-7b model on uncensored wizardlm, alpaca, & dolly datasets.

Pankaj Mathur. 2023b. orca_mini_v3_13b: An orca style llama2-70b model.

Pankaj Mathur. 2023c. orca_mini_v3_7b: An explain tuned llama2-7b model.

Rimon Melamed, Lucas Hurley McCabe, Tanay Wakhare, Yejin Kim, H. Howie Huang, and Enric Boix-Adserà. 2024. Prompts have evil twins. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 46–74, Miami, Florida, USA. Association for Computational Linguistics.

Sean Meyn and Richard L. Tweedie. 2009. *Markov Chains and Stochastic Stability*, 2nd edition. Cambridge Mathematical Library. Cambridge University Press.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *Preprint*, arXiv:1809.02789.

Xu Ming. 2023. textgen: Implementation of language model finetune.

Arindam Mitra, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah. 2023. Orca 2: Teaching small language models how to reason. *Preprint*, arXiv:2311.11045.

Koh Mitsuda, Xinqi Chen, Toshiaki Wakatsuki, and Kei Sawada. 2024. rinna/llama-3-youko-8b.

MosaicML NLP Team. 2023a. Introducing mpt-30b: Raising the bar for open-source foundation models. Accessed: 2023-06-22.

MosaicML NLP Team. 2023b. Introducing mpt-7b: A new standard for open-source, commercially usable llms. Accessed: 2023-03-28.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2025. Generative representational instruction tuning. *Preprint*, arXiv:2402.09906.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4. *Preprint*, arXiv:2306.02707.

Xavier Murias. 2023. Cybertron: Uniform neural alignment.

Nexusflow.ai team. 2023. Nexusraven-v2: Surpassing gpt-4 for zero-shot function calling.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, Hang Zhang, and Lidong Bing. 2024. Seallms – large language models for southeast asia. *Preprint*, arXiv:2312.00738.

Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. 2023. Codegen: An open large language model for code with multi-turn program synthesis. *Preprint*, arXiv:2203.13474.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *Preprint*, arXiv:2303.13375.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, SpeakLeash Team, and Cyfronet Team. 2024a. Introducing bielik-7b-v0.1: Polish language model. Accessed: 2024-04-01.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Sebastian Kondracki, SpeakLeash Team, and Cyfronet Team. 2024b. Introducing bielik-7b-instruct-v0.1: Instruct polish language model. Accessed: 2024-04-01.

Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2024c. Bielik 7b v0.1: A polish language model – development, insights, and evaluation. *Preprint*, arXiv:2410.18565.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Momose Oyama, Sho Yokoi, and Hidetoshi Shimodaira. 2023. Norm of word embedding encodes information gain. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2108–2130, Singapore. Association for Computational Linguistics.

Ankit Pal and Malaikannan Sankarasubbu. 2024a. Gemini goes to med school: Exploring the capabilities of multimodal large language models on medical challenge problems & hallucinations. *Preprint*, arXiv:2402.07023.

Ankit Pal and Malaikannan Sankarasubbu. 2024b. Openbiollms: Advancing open-source large language models for healthcare and life sciences.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. 2024. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *Preprint*, arXiv:2402.13228.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R. Bowman. 2022. Bbq: A hand-built bias benchmark for question answering. *Preprint*, arXiv:2110.08193.

Leonid Pekelis, Michael Feil, Forrest Moret, Mark Huang, and Tiffany Peng. 2024. Llama 3 gradient: A series of long context models.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data, and web data only. *Preprint*, arXiv:2306.01116.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023a. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2023b. Yarn: Efficient context window extension of large language models. *Preprint*, arXiv:2309.00071.

Nikhil Pinnaparaju, Reshinth Adithyan, Duy Phung, Jonathan Tow, James Baicoianu, and Nathan Cooper. 2024. Stable code 3b.

Kunat Pipatanakul, Phatrasek Jirabovonvisut, Potsawee Manakul, Sittipong Sripaisarnmongkol, Ruangsak Patomwong, Pathomporn Chokchainant, and Kasima Tharnpipitchai. 2023. Typhoon: Thai large language models. *Preprint*, arXiv:2312.13951.

Marco Polignano, Pierpaolo Basile, and Giovanni Semeraro. 2024. Advanced natural-based interaction for the italian language: Llamantino-3-anita. *Preprint*, arXiv:2405.07101.

Ronak Pradeep, Sahel Sharifymoghaddam, and Jimmy Lin. 2023. Rankvicuna: Zero-shot listwise document reranking with open-source large language models. *Preprint*, arXiv:2309.15088.

Ofir Press, Noah A. Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. *Preprint*, arXiv:2108.12409.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Preprint*, arXiv:2305.18290.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, and 7 others. 2024. Code llama: Open foundation models for code. *Preprint*, arXiv:2308.12950.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. *Preprint*, arXiv:1804.09301.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. *Preprint*, arXiv:2310.03668.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *Preprint*, arXiv:1907.10641.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *Preprint*, arXiv:1904.09728.

Akira Sasaki, Masato Hirakawa, Shintaro Horie, and Tomoaki Nakamura. 2023a. Elyza-japanese-llama-2-7b.

Akira Sasaki, Masato Hirakawa, Shintaro Horie, Tomoaki Nakamura, Sam Passaglia, and Daisuke Oba. 2023b. Elyza-japanese-llama-2-13b.

Ryoma Sato. 2022. Word tour: One-dimensional word embeddings via the traveling salesman problem. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2166–2172. Association for Computational Linguistics.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained models for the japanese language. *Preprint*, arXiv:2404.01657.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Pratyusha Sharma, Jordan T. Ash, and Dipendra Misra. 2023. The truth is in there: Improving reasoning in language models with layer-selective rank reduction. *Preprint*, arXiv:2312.13558.

Noam Shazeer. 2019. Fast transformer decoding: One write-head is all you need. *Preprint*, arXiv:1911.02150.

Noam Shazeer. 2020. Glu variants improve transformer. *Preprint*, arXiv:2002.05202.

Weijia Shi, Sewon Min, Maria Lomeli, Chunting Zhou, Margaret Li, Gergely Szilvasy, Rich James, Xi Victoria Lin, Noah A. Smith, Luke Zettlemoyer, Scott Yih, and Mike Lewis. 2024. In-context pretraining: Language modeling beyond document boundaries. *Preprint*, arXiv:2310.10638.

Hidetoshi Shimodaira. 1993. A model search technique based on confidence set and map of models [モデルの信頼集合と地図によるモデル探索] (in japanese). *Proceedings of the Institute of Statistical Mathematics*, 41(2):131–147.

Hidetoshi Shimodaira. 2001. Multiple comparisons of log-likelihoods and combining nonnested models with applications to phylogenetic tree selection. *Communications in Statistics-Theory and Methods*, 30(8-9):1751–1772.

Hidetoshi Shimodaira and Ying Cao. 1998. A graphical technique for model selection diagnosis. *The Institute of Statistical Mathematics Research Memorandum*, 680.

Hidetoshi Shimodaira and Masami Hasegawa. 2005. *Assessing the Uncertainty in Phylogenetic Inference*, pages 463–493. Springer New York, New York, NY.

Hidetoshi Shimodaira and Yoshikazu Terada. 2019. Selective inference for testing trees and edges in phylogenetics. *Frontiers in ecology and evolution*, 7:174.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2020. Megatron-lm: Training multi-billion parameter language models using model parallelism. *Preprint*, arXiv:1909.08053.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, and 11 others. 2022. Large language models encode clinical knowledge. *Preprint*, arXiv:2212.13138.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Shashank Sonkar, Naiming Liu, Debshila Basu Mallick, and Richard G. Baraniuk. 2023. Class: A design framework for building intelligent tutoring systems based on learning science principles. *Preprint*, arXiv:2305.13272.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 432 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Preprint*, arXiv:2206.04615.

Stability AI Language Team. 2024. Stable lm 2 1.6b.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. 2023a. Roformer: Enhanced transformer with rotary position embedding. *Preprint*, arXiv:2104.09864.

Yixuan Su, Tian Lan, and Deng Cai. 2023b. Openalpaca: A fully open-source instruction-following model based on openllama.

Shivchander Sudalairaj, Abhishek Bhandwaldar, Aldo Pareja, Kai Xu, David D. Cox, and Akash Srivastava. 2024. Lab: Large-scale alignment for chatbots. *Preprint*, arXiv:2403.01081.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. *Preprint*, arXiv:2306.11695.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model.

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. Ul2: Unifying language learning paradigms. *Preprint*, arXiv:2205.05131.

Yi Tay, Jason Wei, Hyung Won Chung, Vinh Q. Tran, David R. So, Siamak Shakeri, Xavier Garcia, Huaixiu Steven Zheng, Jinfeng Rao, Aakanksha Chowdhery, Denny Zhou, Donald Metzler, Slav Petrov, Neil Houlsby, Quoc V. Le, and Mostafa Dehghani. 2022. Transcending scaling laws with 0.1% extra compute. *Preprint*, arXiv:2210.11399.

"Teknium", Charles Goddard, "interstellarninja", "theemozilla", "karan4d", and "huemin_art". 2024a. Hermes-2-theta-llama-3-8b.

"Teknium", "interstellarninja", "theemozilla", "karan4d", and "huemin_art". 2024b. Hermes-2-pro-llama-3-8b.

"Teknium", "theemozilla", "karan4d", and "huemin_art". 2024c. Nous hermes 2 mistral 7b dpo.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. 2025. Improving pretraining data using perplexity correlations. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Migel Tissera. 2023. Synthia-7b-v1.3: Synthetic intelligent agent.

Together Computer. 2023. Redpajama-data: An open source recipe to reproduce llama training dataset.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jonathan Tow. 2023. Stablelm alpha v2 models.

Jonathan Tow, Marco Bellagente, Dakota Mahan, and Carlos Riquelme. 2023. Stablelm 3b 4e1t.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Bram Vanroy. 2024. Geitje 7b ultra: A conversational model for dutch. *Preprint*, arXiv:2412.04092.

Gaël Varoquaux, Lars Buitinck, Gilles Louppe, Olivier Grisel, Fabian Pedregosa, and Andreas Mueller. 2015. Scikit-learn: Machine learning without learning the machinery. *GetMobile Mob. Comput. Commun.*, 19(1):29–33.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Leandro von Werra, Alex Havrilla, Max reciprocated, Jonathan Tow, Aman cat state, Duy V. Phung, Louis Castricato, Shahbuland Matiana, Alan, Ayush Thakur, Alexey Bukhtiyarov, aaronrmm, Fabrizio Milo, Daniel, Daniel King, Dong Shin, Ethan Kim, Justin Wei, Manuel Romero, and 10 others. 2023. CarperAI/trlx: v0.6.0: LLaMa (Alpaca), Benchmark Util, T5 ILQL, Tests.

Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX.

Ben Wang and Aran Komatsuzaki. 2021. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.

Guan Wang, Sijie Cheng, Qiying Yu, and Changling Liu. 2023a. OpenChat: Advancing Open-source Language Models with Imperfect Data.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024a. Openchat: Advancing open-source language models with mixed-quality data. *Preprint*, arXiv:2309.11235.

Guan Wang, Bleys Goodson, Wing Lian, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023b. Openorcaxopenchatpreview2: Llama2-13b model instruct-tuned on filtered openorcav1 gpt-4 dataset.

Shenzhi Wang, Yaowei Zheng, Guoyin Wang, Shiji Song, and Gao Huang. 2024b. Llama3-8b-chinese-chat (revision 6622a23).

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023c. How far can camels go? exploring the state of instruction tuning on open resources. *Preprint*, arXiv:2306.04751.

Zihan Wang, Deli Chen, Damai Dai, Runxin Xu, Zhuoshu Li, and Y. Wu. 2024c. Let the expert stick to his last: Expert-specialized fine-tuning for sparse architectural large language models. *Preprint*, arXiv:2407.01906.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *Preprint*, arXiv:2203.05482.

Haoyuan Wu, Haisheng Zheng, Zhuolun He, and Bei Yu. 2024. Parameter-efficient sparsity crafting from dense to mixture-of-experts for instruction tuning on general tasks. *Preprint*, arXiv:2401.02731.

Mengzhou Xia, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen. 2024. Sheared llama: Accelerating language model pre-training via structured pruning. *Preprint*, arXiv:2310.06694.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. 2023. Lm-cocktail: Resilient tuning of language models via model merging. *Preprint*, arXiv:2311.13534.

Huajian Xin, Daya Guo, Zhihong Shao, Zhizhou Ren, Qihao Zhu, Bo Liu, Chong Ruan, Wenda Li, and Xiaodan Liang. 2024a. Deepseek-prover: Advancing theorem proving in llms through large-scale synthetic data. *Preprint*, arXiv:2405.14333.

Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. 2024b. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search. *Preprint*, arXiv:2408.08152.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023a. Wizardlm: Empowering large language models to follow complex instructions. *Preprint*, arXiv:2304.12244.

Canwen Xu, Daya Guo, Nan Duan, and Julian McAuley. 2023b. Baize: An open-source chat model with parameter-efficient tuning on self-chat data. *Preprint*, arXiv:2304.01196.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024a. A paradigm shift in machine translation: Boosting translation performance of large language models. *Preprint*, arXiv:2309.11674.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024b. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *Preprint*, arXiv:2401.08417.

Xwin-LM Team. 2023. Xwin-lm.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin Raffel, and Mohit Bansal. 2023a. Ties-merging: Resolving interference when merging models. *Preprint*, arXiv:2306.01708.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. 2023b. Ties-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, volume 36, pages 7093–7115. Curran Associates, Inc.

Hiroaki Yamagiwa, Yusuke Takase, and Hidetoshi Shimodaira. 2024. Axis tour: Word tour determines the order of axes in ica-transformed embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 477–506. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024a. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024b. Mentallama: Interpretable mental health analysis on social media with large language models. *Preprint*, arXiv:2309.13567.

Ping Yang, Junjie Wang, Ruyi Gan, Xinyu Zhu, Lin Zhang, Ziwei Wu, Xinyu Gao, Jiaxing Zhang, and Tetsuya Sakai. 2022. Zero-shot learners for natural language understanding via a unified multiple choice perspective. *Preprint*, arXiv:2210.08590.

Nicolas Yax, Pierre-Yves Oudeyer, and Stefano Palminteri. 2025. PhyloLM: Inferring the phylogeny of large language models and predicting their performances in benchmarks. In *The Thirteenth International Conference on Learning Representations*.

Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024a. Language models are super mario: Absorbing abilities from homologous models as a free lunch. *Preprint*, arXiv:2311.03099.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T. Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. 2024b. Metamath: Bootstrap your own mathematical questions for large language models. *Preprint*, arXiv:2309.12284.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhu Chen. 2024b. Mammoth2: Scaling instructions from the web. *Preprint*, arXiv:2405.03548.

YuLan-Team. 2023. Yulan-chat: An open-source bilingual chatbot.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. *Preprint*, arXiv:1910.07467.

Chen Zhang, Dawei Song, Zheyu Ye, and Yan Gao. 2024a. Towards the law of capacity gap in distilling language models. *Preprint*, arXiv:2311.07052.

Ge Zhang, Xinrun Du, Bei Chen, Yiming Liang, Tongxu Luo, Tianyu Zheng, Kang Zhu, Yuyang Cheng, Chunpu Xu, Shuyue Guo, Haoran Zhang, Xingwei Qu, Junjie Wang, Ruibin Yuan, Yizhi Li, Zekun Wang, Yudong Liu, Yu-Hsuan Tsai, Fengji Zhang, and 3 others. 2024b. Cmmmu: A chinese massive multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2401.11944.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, Jianheng Zhuo, Qi Yang, Yongfeng Huang, Xiayu Li, Yanghan Wu, Junyu Lu, Xinyu Zhu, Weifeng Chen, Ting Han, and 6 others. 2022a. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *CoRR*, abs/2209.02970.

Kaiyan Zhang, Ning Ding, Biqing Qi, Sihang Zeng, Haoxin Li, Xuekai Zhu, Zhang-Ren Chen, and Bowen Zhou. 2024c. Ultramedical: Building specialized generalists in biomedicine.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022b. Opt: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Wenxuan Zhang, Sharifah Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2023. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Preprint*, arXiv:2306.05179.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *Preprint*, arXiv:1804.06876.

Tianyu Zhao, Akio Kaga, and Kei Sawada. 2023a. rinna/youri-7b.

Tianyu Zhao and Kei Sawada. 2023. rinna/japanese-gpt-neox-3.6b.

Tianyu Zhao, Toshiaki Wakatsuki, Akio Kaga, Koh Mitsuda, and Kei Sawada. 2023b. rinna/bilingual-gpt-neox-4b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

Xinyu Zhou, Delong Chen, Samuel Cahyawijaya, Xufeng Duan, and Zhenguang Cai. 2025. Linguistic minimal pairs elicit linguistic similarity in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6866–6888, Abu Dhabi, UAE. Association for Computational Linguistics.

Banghua Zhu, Evan Frick, Tianhao Wu, Hanlin Zhu, Karthik Ganesan, Wei-Lin Chiang, Jian Zhang, and Jiantao Jiao. 2023a. Starling-7b: Improving llm helpfulness & harmlessness with rlaif.

Banghua Zhu, Hiteshi Sharma, Felipe Vieira Frujeri, Shi Dong, Chenguang Zhu, Michael I. Jordan, and Jiantao Jiao. 2023b. Fine-tuning language models with advantage-induced policy alignment. *Preprint*, arXiv:2306.02231.

Sally Zhu, Ahmed M Ahmed, Rohith Kuditipudi, and Percy Liang. 2025. Independence tests for language models.

Richard Zhuang, Tianhao Wu, Zhaojin Wen, Andrew Li, Jiantao Jiao, and Kannan Ramchandran. 2025. EmbedLLM: Learning compact representations of large language models. In *The Thirteenth International Conference on Learning Representations*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Related Work

In recent years, research on comparing large language models (LLMs) has gained attention. This section provides an overview of existing studies from three perspectives: model parameters, activations[13], and model outputs.

Comparison of model parameters. One approach to comparing LLMs is to analyze their parameters. Zhu et al. (2025) proposed a statistical framework for evaluating parameter similarity between different models and introduced a method for determining whether these models were trained independently. Horwitz et al. (2025) compare weights to complement undocumented model relationships on Hugging Face. Additionally, Yadav et al. (2023b) focused on parameter changes due to task adaptation, specifically analyzing task vectors[14]. They proposed a method to mitigate interference when integrating task vectors from different models. Specifically, by reducing redundant numerical components and adjusting for conflicting signs, their approach enables effective model merging.

Comparison of activations. Comparisons of LLMs based on activations have also been studied. Zhou et al. (2025) quantified the similarity between LLMs by measuring the cosine similarity of activation differences for linguistic minimal pairs. In particular, they used datasets such as BLiMP (Warstadt et al., 2020) and showed that model similarity is significantly influenced by the pre-training dataset.

Comparison of model outputs. Several approaches compare LLMs based on their outputs. Lv et al. (2023) proposed a method for computing coefficients in parameter ensembling by providing the same input text to two models and comparing the softmax probability distributions at each token. Specifically, they used KL divergence and summed the results to derive appropriate coefficients. Furthermore, Yax et al. (2025) proposed a similarity metric based on the conditional probabilities of LLMs and introduced a method for calculating the phylogenetic distance between different models. Zhuang et al. (2025) propose a framework for vector representations of language models based on their performance on downstream tasks. Additionally, a method has been proposed for measuring differences in conditional probabilities based on prompts using KL divergence (Melamed et al., 2024).

B Double Centering

We confirm the notation and computational operations. The matrices $\boldsymbol{L}$, $\boldsymbol{\Xi}$, and $\boldsymbol{Q}$ are all of size $K \times N$, and their elements are denoted by $\ell_{is}, \xi_{is}, q_{is}$, respectively. In particular, we have $\ell_{is} = \ell_i(x_s)$. First, row-wise centering of $\boldsymbol{L}$ is performed by subtracting the mean log-likelihood $\bar{\ell}_i$ of each model from each row $(\ell_{i1}, \ldots, \ell_{iN})$, resulting in $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)^\top$. Next, column-wise centering of $\boldsymbol{\Xi}$ is performed by subtracting the coordinate component $\bar{\xi}_s$ of the mean vector $\bar{\boldsymbol{\xi}}$ from each column $(\xi_{1s}, \ldots, \xi_{Ks})^\top$, yielding $\boldsymbol{Q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K)^\top$. Thus, this process involves *double centering*, where column-wise centering follows row-wise centering. Notably, even after column-wise centering, the row-wise mean of $\boldsymbol{Q}$

[13]In general, "activations" refer to the intermediate outputs of Transformer models (e.g., the residual stream or neurons) (Bereska and Gavves, 2024).

[14]The difference in parameters before and after fine-tuning is referred to as a task vector, and arithmetic operations on these vectors are effective (Ilharco et al., 2023).

remains zero:

$$\begin{aligned}\frac{1}{N}\sum_{s=1}^{N} q_{is} &= \frac{1}{N}\sum_{s=1}^{N}(\xi_{is}-\bar{\xi}_s)\\ &= \frac{1}{N}\sum_{s=1}^{N}\xi_{is} - \frac{1}{NK}\sum_{i=1}^{K}\sum_{s=1}^{N}\xi_{is}\\ &= 0-0=0. \end{aligned} \tag{6}$$

The column-wise centering can be interpreted as follows. In the $\boldsymbol{\xi}$-coordinate system, the mean vector $\bar{\boldsymbol{\xi}}$ of the K model coordinates $\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_K$ can be regarded as representing an "average model." By redefining this average model as the new origin, we obtain the $\boldsymbol{q}$-coordinate system. From the definition $\boldsymbol{q}_i = \boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}$, its mean satisfies

$$\sum_{i=1}^{K} \boldsymbol{q}_i / K = \mathbf{0}.$$

The row-wise centering can be interpreted as follows. Let $\mathbf{1}_N = (1,\dots,1)^\top \in \mathbb{R}^N$. Since $\bar{\ell}_i = \mathbf{1}_N^\top \boldsymbol{\ell}_i / N$, we have $\boldsymbol{\xi}_i = \boldsymbol{\ell}_i - \bar{\ell}_i \mathbf{1}_N$. Thus,

$$\mathbf{1}_N^\top \boldsymbol{\xi}_i = \mathbf{1}_N^\top \boldsymbol{\ell}_i - \bar{\ell}_i N = 0,$$

and furthermore, $\mathbf{1}_N^\top \bar{\boldsymbol{\xi}} = 0$. From this, equation (6) is actually trivial, as

$$\mathbf{1}_N^\top \boldsymbol{q}_i = \mathbf{1}^\top(\boldsymbol{\xi}_i - \bar{\boldsymbol{\xi}}) = 0-0=0.$$

The row-wise centering implies that $\boldsymbol{\xi}_1,\dots,\boldsymbol{\xi}_K$ and $\boldsymbol{q}_1,\dots,\boldsymbol{q}_K$ lie in the subspace orthogonal to $\mathbf{1}_N$.

C Effect of Errors in Model Coordinates

We analyze the impact of additive errors ϵ_{is} in the log-likelihood vector components ℓ_{is} for $i = 1,\dots,K$ and $s = 1,\dots,N$. Denoting the true values with an asterisk as ℓ^*_{is}, the observed values can be expressed as

$$\ell_{is} = \ell^*_{is} + \epsilon_{is}.$$

We decompose the error as follows:

$$\epsilon_{is} = a + b_i + c_s + d_{is},$$

where we assume, without loss of generality, the constraints

$$\sum_{i=1}^{K} b_i = \sum_{s=1}^{N} c_s = \sum_{i=1}^{K} d_{is} = \sum_{s=1}^{N} d_{is} = 0.$$

Here, a, b_i, and c_s are bias terms, while d_{is} represents interaction terms. Using simple calculations, we obtain:

$$\bar{\ell}_i = \frac{1}{N}\sum_{s=1}^{N} \ell_{is} = \bar{\ell}^*_i + a + b_i,$$

$$\xi_{is} = \ell_{is} - \bar{\ell}_i = \xi^*_{is} + c_s + d_{is},$$

$$\bar{\xi}_s = \frac{1}{K}\sum_{i=1}^{K} \xi_{is} = \bar{\xi}^*_s + c_s,$$

$$q_{is} = \xi_{is} - \bar{\xi}_s = q^*_{is} + d_{is}.$$

Additionally, for the differences between two models, which are crucial for the model map, we obtain:

$$\ell_{is} - \ell_{js} = \ell^*_{is} - \ell^*_{js} + b_i - b_j + d_{is} - d_{js},$$

$$\xi_{is} - \xi_{js} = \xi^*_{is} - \xi^*_{js} + d_{is} - d_{js},$$

$$q_{is} - q_{js} = q^*_{is} - q^*_{js} + d_{is} - d_{js}.$$

Thus, the terms affected by the error are $\bar{\ell}_i$, which is influenced by $a + b_i$, and $\ell_{is} - \ell_{js}$, which is affected by $b_i + d_{is}$, meaning it is influenced by the bias terms. However, in the centered values $\xi_{is} - \xi_{js}$ and $q_{is} - q_{js}$, only the interaction term d_{is} contributes to the error.

D Theory of Model Coordinates for Text Probability Distributions

In this section, we prove the main results of Section 2, namely (2) and (3). Our discussion applies not only to text probability distributions but also more generally to any setting where i.i.d. observations $x_1,\dots,x_N \sim p_i(x)$ are available. Compared to the previous study that proposed model maps (Shimodaira, 1993; Shimodaira and Cao, 1998), we conduct a more precise analysis in this paper. Specifically, while the previous study provides only a brief evaluation of the approximation, we present a more transparent discussion based on the properties of the exponential family of distributions. In the next section, as the starting point of our discussion, we construct a *super model* that includes the K models p_i, $i = 1,\dots,K$, as submodels. This model is introduced as a mathematical tool to rigorously prove the main theorem of this paper, and we do not compute it numerically in practice.

D.1 Exponential family of distributions

We first consider a model in the exponential family of distributions parameterized by a K-dimensional parameter $\boldsymbol{\theta} \in \mathbb{R}^K$:

$$p(x; \boldsymbol{\theta}) = p_0(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{b}(x) - \psi(\boldsymbol{\theta})). \quad (7)$$

Here, the function $\boldsymbol{b}(x) = (b_1(x), \ldots, b_K(x))^\top$ will be defined later using the K models. The normalization constant is given by

$$\begin{aligned} Z(\boldsymbol{\theta}) &= \sum_{x \in \mathcal{X}} p_0(x) \exp(\boldsymbol{\theta}^\top \boldsymbol{b}(x)), \\ \psi(\boldsymbol{\theta}) &= \log Z(\boldsymbol{\theta}), \end{aligned}$$

which ensures that $\sum_{x \in \mathcal{X}} p(x; \boldsymbol{\theta}) = 1$. For $\boldsymbol{\theta} = \mathbf{0}$, where $\mathbf{0} = (0, \ldots, 0)^\top$, we obtain

$$p(x; \mathbf{0}) = p_0(x).$$

To associate the K models with (7), we define $\boldsymbol{b}(x)$. For a constant $\lambda > 0$, we set

$$\lambda b_i(x) := \ell_i(x) - \ell_0(x). \quad (8)$$

The constant λ is an order parameter introduced for theoretical convenience, and in our theoretical framework, we assume that $b_i(x)$ is of constant order and that λ is sufficiently small[15]. Thus, we essentially assume that $|\ell_i(x) - \ell_0(x)| = O_p(\lambda)$ is sufficiently small, implying that each model p_i provides a good approximation of the true generative model p_0. In the proof of the main theorem, we consider the asymptotic theory as $\lambda \to 0$, retaining terms up to $O(\lambda^2)$ while ignoring those of $O(\lambda^3)$.

A one-hot vector is defined as $\boldsymbol{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)^\top \in \mathbb{R}^K$ for $i = 1, \ldots, K$, where only the i-th element is 1. Then, setting $\boldsymbol{\theta} = \lambda \boldsymbol{e}_i$ gives

$$p(x; \lambda \boldsymbol{e}_i) = p_i(x). \quad (9)$$

Indeed, substituting (8) into (7) yields

$$\begin{aligned} & p(x; \lambda \boldsymbol{e}_i) \\ =& p_0(x) \exp(\lambda \boldsymbol{e}_i^\top \boldsymbol{b}(x) - \psi(\lambda \boldsymbol{e}_i)) \\ =& p_0(x) \exp(\ell_i(x) - \ell_0(x) - \psi(\lambda \boldsymbol{e}_i)) \\ =& p_0(x)(p_i(x)/p_0(x)) \exp(-\psi(\lambda \boldsymbol{e}_i)) \\ =& p_i(x) \exp(-\psi(\lambda \boldsymbol{e}_i)) \\ =& p_i(x), \end{aligned}$$

where $\psi(\lambda \boldsymbol{e}_i) = 0$.

[15] If numerical computation were to be performed, λ could be set to any arbitrary value (e.g., $\lambda = 1$).

D.2 Properties of the exponential family of distributions

This section outlines some well-known basic properties of the exponential family of distributions, which have been established in the literature (Barndorff-Nielsen, 2014; Efron, 1978, 2022; Amari, 1982). We define the expectation and covariance matrix of $\boldsymbol{b}(x)$ as follows:

$$\begin{aligned} \boldsymbol{\eta}(\boldsymbol{\theta}) &:= \mathop{\mathbb{E}}_{x \sim p(\boldsymbol{\theta})} \big(\boldsymbol{b}(x)\big) = \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) p(x; \boldsymbol{\theta}), \\ G(\boldsymbol{\theta}) &:= \mathop{\mathbb{E}}_{x \sim p(\boldsymbol{\theta})} \big\{(\boldsymbol{b}(x) - \boldsymbol{\eta}(\boldsymbol{\theta}))(\boldsymbol{b}(x) - \boldsymbol{\eta}(\boldsymbol{\theta}))^\top\big\} \\ &= \mathop{\mathrm{Var}}_{x \sim p(\boldsymbol{\theta})} \big(\boldsymbol{b}(x)\big). \end{aligned}$$

Here, the elements of $\boldsymbol{\eta}(\boldsymbol{\theta})$ are expectations given by $\mathbb{E}_{x \sim p(\boldsymbol{\theta})}(b_i(x))$, and the elements of $G(\boldsymbol{\theta})$ are covariances given by $G_{ij}(\boldsymbol{\theta}) = \mathrm{Cov}_{x \sim p(\boldsymbol{\theta})}(b_i(x), b_j(x))$. These quantities can be expressed in terms of $\psi(\boldsymbol{\theta})$ as follows:

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad (10)$$

$$G(\boldsymbol{\theta}) = \frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}. \quad (11)$$

We now derive these two equations. First, from

$$\frac{\partial Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) p_0(x) e^{\boldsymbol{\theta}^\top \boldsymbol{b}(x)},$$

we obtain

$$\begin{aligned} \frac{\partial \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} =& \frac{\partial \log Z}{\partial \boldsymbol{\theta}} = \frac{1}{Z(\boldsymbol{\theta})} \frac{\partial Z}{\partial \boldsymbol{\theta}} \\ =& \frac{1}{Z(\boldsymbol{\theta})} \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) p_0(x) e^{\boldsymbol{\theta}^\top \boldsymbol{b}(x)} \\ =& \sum_{x \in \mathcal{X}} \boldsymbol{b}(x) p(x; \boldsymbol{\theta}) = \boldsymbol{\eta}(\boldsymbol{\theta}). \end{aligned}$$

Thus, we have established (10). Next, using

$$\begin{aligned} \frac{\partial p(x; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} =& \Big(\boldsymbol{b}(x) - \frac{\partial \psi}{\partial \boldsymbol{\theta}}\Big) p(x; \boldsymbol{\theta}) \\ =& (\boldsymbol{b}(x) - \boldsymbol{\eta}(\boldsymbol{\theta})) p(x; \boldsymbol{\theta}), \end{aligned}$$

we obtain

$$\begin{aligned}\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top} &= \frac{\partial\boldsymbol{\eta}(\boldsymbol{\theta})^\top}{\partial\boldsymbol{\theta}}\\ &=\frac{\partial}{\partial\boldsymbol{\theta}}\sum_{x\in\mathcal{X}}\boldsymbol{b}(x)^\top p(x;\boldsymbol{\theta})\\ &=\sum_{x\in\mathcal{X}}(\boldsymbol{b}(x)-\boldsymbol{\eta}(\boldsymbol{\theta}))\boldsymbol{b}(x)^\top p(x;\boldsymbol{\theta})\\ &=\sum_{x\in\mathcal{X}}(\boldsymbol{b}(x)-\boldsymbol{\eta}(\boldsymbol{\theta}))(\boldsymbol{b}(x)-\boldsymbol{\eta}(\boldsymbol{\theta}))^\top p(x;\boldsymbol{\theta})\\ &=G(\boldsymbol{\theta}).\end{aligned}$$

Thus, we have established (11).

D.3 Approximation of the KL divergence

The Kullback-Leibler (KL) divergence between the models $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\theta}')$ at parameter values $\boldsymbol{\theta},\boldsymbol{\theta}'\in\mathbb{R}^K$ is given by

$$\begin{aligned}&\mathrm{KL}(p(\boldsymbol{\theta}),p(\boldsymbol{\theta}'))\\ =&\sum_{x\in\mathcal{X}}p(x;\boldsymbol{\theta})\log\frac{p(x;\boldsymbol{\theta})}{p(x;\boldsymbol{\theta}')}\\ =&\sum_{x\in\mathcal{X}}p(x;\boldsymbol{\theta})\{(\boldsymbol{\theta}-\boldsymbol{\theta}')^\top\boldsymbol{b}(x)-\psi(\boldsymbol{\theta})+\psi(\boldsymbol{\theta}')\}\\ =&(\boldsymbol{\theta}-\boldsymbol{\theta}')^\top\boldsymbol{\eta}(\boldsymbol{\theta})-\psi(\boldsymbol{\theta})+\psi(\boldsymbol{\theta}'). \end{aligned}\tag{12}$$

Here, we assume that the parameter values $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are sufficiently close to $\mathbf{0}$. In particular, we assume $\|\boldsymbol{\theta}\|=O(\lambda)$ and $\|\boldsymbol{\theta}'\|=O(\lambda)$. Substituting (10) and (11) into the Taylor expansion of $\psi(\boldsymbol{\theta})$ gives

$$\begin{aligned}&\psi(\boldsymbol{\theta}')\\ =&\psi(\boldsymbol{\theta})+\frac{\partial\psi}{\partial\boldsymbol{\theta}^\top}(\boldsymbol{\theta}'-\boldsymbol{\theta})\\ &+\frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{\theta})^\top\frac{\partial^2\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}(\boldsymbol{\theta}'-\boldsymbol{\theta})\\ &+O(\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|^3)\\ =&\psi(\boldsymbol{\theta})+\boldsymbol{\eta}(\boldsymbol{\theta})^\top(\boldsymbol{\theta}'-\boldsymbol{\theta})\\ &+\frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{\theta})^\top G(\boldsymbol{\theta})(\boldsymbol{\theta}'-\boldsymbol{\theta})+O(\lambda^3). \end{aligned}\tag{13}$$

Substituting (13) into (12) gives

$$\begin{aligned}&\mathrm{KL}(p(\boldsymbol{\theta}),p(\boldsymbol{\theta}'))\\ &=\frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{\theta})^\top G(\boldsymbol{\theta})(\boldsymbol{\theta}'-\boldsymbol{\theta})+O(\lambda^3). \end{aligned}\tag{14}$$

This corresponds to eq. (9) of Oyama et al. (2023). Here, the equation holds approximately by ignoring higher-order terms of $O(\lambda^3)$. For more details, refer to Amari (1982, p. 369) and Efron (2022, p. 35). More generally, $G(\boldsymbol{\theta})$ represents the Fisher information metric, and (14) holds for a wide class of probability models (Amari, 1998) with $\|\boldsymbol{\theta}'-\boldsymbol{\theta}\|=O(\lambda)$. Furthermore, since $G(\boldsymbol{\theta})=G(\mathbf{0})+O(\|\boldsymbol{\theta}\|)=G(\mathbf{0})+O(\lambda)$, we obtain

$$\begin{aligned}&\mathrm{KL}(p(\boldsymbol{\theta}),p(\boldsymbol{\theta}'))\\ &=\frac{1}{2}(\boldsymbol{\theta}'-\boldsymbol{\theta})^\top G(\mathbf{0})(\boldsymbol{\theta}'-\boldsymbol{\theta})+O(\lambda^3). \end{aligned}\tag{15}$$

D.4 The variance representation of the KL divergence

Substituting $p_i=p(\lambda\boldsymbol{e}_i)$ into (15) gives

$$\begin{aligned}&2\mathrm{KL}(p_i,p_j)=2\mathrm{KL}(p(\lambda\boldsymbol{e}_i),p(\lambda\boldsymbol{e}_j))\\ &=\lambda^2(\boldsymbol{e}_i-\boldsymbol{e}_j)^\top G(\mathbf{0})(\boldsymbol{e}_i-\boldsymbol{e}_j)+O(\lambda^3). \end{aligned}\tag{16}$$

Here, we have

$$\begin{aligned}&\boldsymbol{e}_i^\top G(\mathbf{0})\boldsymbol{e}_j=G_{ij}(\mathbf{0})\\ &=\mathop{\mathbb{E}}_{x\sim p_0}\Big\{(b_i(x)-\eta_i(\mathbf{0}))(b_j(x)-\eta_j(\mathbf{0}))\Big\}. \end{aligned}\tag{17}$$

Next, we substitute (17) and (8) into the right-hand side of (16) to derive an alternative expression for the KL divergence:

$$\begin{aligned}&\lambda^2(\boldsymbol{e}_i-\boldsymbol{e}_j)^\top G(\mathbf{0})(\boldsymbol{e}_i-\boldsymbol{e}_j)\\ =&\lambda^2\mathop{\mathbb{E}}_{x\sim p_0}\Big[\Big\{(b_i(x)-\eta_i(\mathbf{0}))\\ &\qquad-(b_j(x)-\eta_j(\mathbf{0}))\Big\}^2\Big]\\ =&\mathop{\mathbb{E}}_{x\sim p_0}\Big[\Big\{(\ell_i(x)-\ell_0(x))-(\ell_j(x)-\ell_0(x))\\ &\qquad-\mathop{\mathbb{E}}_{x'\sim p_0}\Big((\ell_i(x')-\ell_0(x'))\\ &\qquad\qquad-(\ell_j(x')-\ell_0(x'))\Big)\Big\}^2\Big]\\ =&\mathop{\mathbb{E}}_{x\sim p_0}\Big[\Big\{\ell_i(x)-\ell_j(x)\\ &\qquad-\mathop{\mathbb{E}}_{x'\sim p_0}\Big(\ell_i(x')-\ell_j(x')\Big)\Big\}^2\Big]\\ =&\mathop{\mathrm{Var}}_{x\sim p_0}\Big(\ell_i(x)-\ell_j(x)\Big).\end{aligned}$$

Finally, substituting this result into (16) yields

$$2\mathrm{KL}(p_i,p_j)=\mathop{\mathrm{Var}}_{x\sim p_0}\Big(\ell_i(x)-\ell_j(x)\Big)+O(\lambda^3). \tag{18}$$

This establishes (2). Furthermore, since $|\ell_i(x)-\ell_j(x)|=O_p(\lambda)$, the magnitude of (18) is $O(\lambda^2)$.

D.5 Estimation of the KL divergence

If the expected value $\mathbb{E}_{x\sim p_0}(f(x))$ of a function $f(x)$ exists and is bounded, then by the law of large numbers, the sample mean[16] converges to the expected value as $N \to \infty$, and we have

$$\mathop{\mathbb{E}}_{x\sim D}(f(x)) = \mathop{\mathbb{E}}_{x\sim p_0}(f(x)) + O_p(N^{-1/2}).$$

Applying this to (18) and ignoring the terms of order $O_p(\lambda^3 + \lambda^2 N^{-1/2})$, we obtain the following approximation:

$$2\mathrm{KL}(p_i, p_j) \approx \mathop{\mathrm{Var}}_{x\sim D}\Big(\ell_i(x) - \ell_j(x)\Big). \quad (19)$$

We define the coordinates $\boldsymbol{\xi}_i \in \mathbb{R}^N$ of model p_i as

$$\boldsymbol{\xi}_i = (\xi_{i1}, \ldots, \xi_{iN})^\top$$

with

$$\xi_{is} := \ell_i(x_s) - \mathop{\mathbb{E}}_{x\sim D}(\ell_i(x))$$

for $s = 1, \ldots, N$. From (19), we obtain

$$\begin{aligned} 2\mathrm{KL}(p_i, p_j) \approx& \frac{1}{N}\sum_{s=1}^{N}(\xi_{is} - \xi_{js})^2 \\ =& \frac{1}{N}\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2. \end{aligned} \quad (20)$$

Since

$$\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|^2 = \|(\boldsymbol{q}_i + \bar{\boldsymbol{\xi}}) - (\boldsymbol{q}_j + \bar{\boldsymbol{\xi}})\|^2 = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2,$$

this establishes (3).

D.6 Relationships among the three types of model coordinates

Let $\mathbf{1}_N = (1, \ldots, 1)^\top \in \mathbb{R}^N$. From the definitions of the $\boldsymbol{\xi}$-coordinate system and the $\boldsymbol{q}$-coordinate system, we have

$$\begin{aligned} \boldsymbol{\xi}_i &= \boldsymbol{q}_i + \bar{\boldsymbol{\xi}}, \\ \boldsymbol{\ell}_i &= \boldsymbol{\xi}_i + \bar{\ell}_i \mathbf{1}_N \\ &= \boldsymbol{q}_i + \bar{\ell}_i \mathbf{1}_N + \bar{\boldsymbol{\xi}}. \end{aligned} \quad (21)$$

Additionally, equation (6) in Appendix B can be rewritten as

$$\mathbf{1}_N^\top \boldsymbol{q}_i = 0. \quad (22)$$

[16] $\mathbb{E}_{x\sim D}(f(x)) = \frac{1}{N}\sum_{x\in D} f(x) = \frac{1}{N}\sum_{s=1}^{N} f(x_s)$ represents the sample mean, and $\mathrm{Var}_{x\sim D}(\cdot)$ represents the sample variance.

Thus, we obtain

$$\begin{aligned} &\|\boldsymbol{\ell}_i - \boldsymbol{\ell}_j\|^2 \\ =&\|(\boldsymbol{q}_i - \boldsymbol{q}_j) + (\bar{\ell}_i - \bar{\ell}_j)\mathbf{1}_N\|^2 \\ =&\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 + N(\bar{\ell}_i - \bar{\ell}_j)^2 \\ &\quad + 2(\bar{\ell}_i - \bar{\ell}_j)\mathbf{1}_N^\top(\boldsymbol{q}_i - \boldsymbol{q}_j) \\ =&\|\boldsymbol{q}_i - \boldsymbol{q}_j\|^2 + N(\bar{\ell}_i - \bar{\ell}_j)^2, \end{aligned}$$

where (21) and (22) are used in the first and last equations, respectively. This establishes (4).

Moreover, since $\bar{\ell}_i = \mathbf{1}_N^\top \boldsymbol{\ell}_i / N$, it is straightforward that the component of the $\boldsymbol{\ell}$-coordinate system in the $\mathbf{1}_N$ direction is given by

$$(\mathbf{1}_N/\sqrt{N})^\top \boldsymbol{\ell}_i = \sqrt{N}\bar{\ell}_i.$$

D.7 Additional notes on modifying the underlying generative model

We examine the effect of changing the true generative model $p_0 = p(\mathbf{0})$ that produces the data D. For clarity, we continue to use the same p_0 as before in constructing $p(x;\boldsymbol{\theta})$ as described in Section D.1. We then introduce a new parameter value $\boldsymbol{\theta}^*$. We assume that each element x_s of the dataset D is generated independently from

$$x_1, \ldots, x_N \sim p(x;\boldsymbol{\theta}^*), \quad (23)$$

and that $\|\boldsymbol{\theta}^*\| = O(\lambda)$. In other words, the true generative model remains sufficiently close to the K models, satisfying $\|\boldsymbol{\theta}_i - \boldsymbol{\theta}^*\| = O(\lambda)$ for $i = 1, \ldots, K$.

First, consider replacing $G(\boldsymbol{\theta})$ in (14) of Section D.3 with $G(\boldsymbol{\theta}^*)$. Because $G(\boldsymbol{\theta}^*) = G(\boldsymbol{\theta}) + O(\lambda)$, we obtain

$$\begin{aligned} &\mathrm{KL}(p(\boldsymbol{\theta}), p(\boldsymbol{\theta}')) \\ &= \frac{1}{2}(\boldsymbol{\theta}' - \boldsymbol{\theta})^\top G(\boldsymbol{\theta}^*)(\boldsymbol{\theta}' - \boldsymbol{\theta}) + O(\lambda^3). \end{aligned} \quad (24)$$

We are generalizing the discussion in the previous sections, and indeed, if we set $\boldsymbol{\theta}^* = \mathbf{0}$ in (24), then (15) is recovered.

Now substitute $p_i = p(\lambda \boldsymbol{e}_i)$ into (24), yielding

$$\begin{aligned} &2\,\mathrm{KL}(p_i, p_j) = 2\,\mathrm{KL}\big(p(\lambda\boldsymbol{e}_i), p(\lambda\boldsymbol{e}_j)\big) \\ &= \lambda^2(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top G(\boldsymbol{\theta}^*)(\boldsymbol{e}_i - \boldsymbol{e}_j) + O(\lambda^3), \end{aligned} \quad (25)$$

which is a generalization of (16) in Section D.4. Using the definition of $\boldsymbol{G}(\boldsymbol{\theta})$, we have

$$\boldsymbol{e}_i^\top G(\boldsymbol{\theta}^*)\boldsymbol{e}_j = G_{ij}(\boldsymbol{\theta}^*) = \mathop{\mathrm{Cov}}_{x\sim p(\boldsymbol{\theta}^*)}\big(b_i(x), b_j(x)\big). \quad (26)$$

Substituting (26) and (8) into the right-hand side of (25) yields

$$\begin{aligned}
&\lambda^2(\boldsymbol{e}_i-\boldsymbol{e}_j)^\top G(\boldsymbol{\theta}^*)(\boldsymbol{e}_i-\boldsymbol{e}_j)\\
&=\lambda^2\Big\{\underset{x\sim p(\boldsymbol{\theta}^*)}{\mathrm{Var}}\big(b_i(x)\big)+\underset{x\sim p(\boldsymbol{\theta}^*)}{\mathrm{Var}}\big(b_j(x)\big)\\
&\qquad-2\underset{x\sim p(\boldsymbol{\theta}^*)}{\mathrm{Cov}}\big(b_i(x),b_j(x)\big)\Big\}\\
&=\lambda^2\underset{x\sim p(\boldsymbol{\theta}^*)}{\mathrm{Var}}\big(b_i(x)-b_j(x)\big)\\
&=\underset{x\sim p(\boldsymbol{\theta}^*)}{\mathrm{Var}}\big(\ell_i(x)-\ell_j(x)\big).
\end{aligned}$$

Finally, substituting this back into (25) gives

$$2\,\mathrm{KL}(p_i,p_j)=\underset{x\sim p(\boldsymbol{\theta}^*)}{\mathrm{Var}}\big(\ell_i(x)-\ell_j(x)\big)+O(\lambda^3), \tag{27}$$

which is a generalization of (2).

Hence, the estimators for KL divergence in Section D.5, specifically (19) and (20), and also (3) in Section 2, remain valid even when the texts are generated by (23). Since (27) holds for any $\boldsymbol{\theta}^*$ with $\|\boldsymbol{\theta}^*\|=O(\lambda)$, this result also applies when the true data-generating model is $p(\boldsymbol{\theta}^*)=p_0$, or, for instance, one of the models $p(\boldsymbol{\theta}^*)=p_i$, or a mixture of the K models, $\boldsymbol{\theta}^*=\sum_{i=1}^K\alpha_i\lambda\boldsymbol{e}_i$ with $\alpha_i=O(1)$. Therefore, this method is robust to contamination in the dataset D (e.g., when the text corpus used for pre-training a model is included in D), as the estimation of KL divergence via the squared Euclidean distance in the $\boldsymbol{\xi}$-coordinate system or the $\boldsymbol{q}$-coordinate system remains relatively unaffected.

E Mapping Language Models into the Space of Token Probability Distributions

In Section 2, we discussed model maps based on the probability distributions $p_i(x)$ of texts generated by language models. This approach requires computing probabilities for a large number of texts in the dataset $D=(x_1,\ldots,x_N)$, leading to high computational costs. To mitigate this issue, we focus on the fact that a text $x=(y_1,\ldots,y_n)$ is a sequence of tokens. Instead of using text probabilities, we discuss model maps based on the conditional probability distributions of token generation, $p_i(y_t|y^{t-1})$. In this approach, model coordinates are computed using only a single text x. A limitation of this approach is that it can only be used for comparing models that share the same tokenizer. Furthermore, the current estimation method ignores the variance of the expected log-likelihood ratio of conditional probabilities, resulting in a rough approximation. Thus, the estimated values should be regarded only as reference values rather than precise measurements.

E.1 Model coordinates

For a text $x=(y_1,\ldots,y_n)$, the coordinates of model p_i

$$\boldsymbol{\zeta}_i=(\zeta_{i1},\ldots,\zeta_{in})^\top\in\mathbb{R}^n$$

are defined as

$$\zeta_{it}:=\log p_i(y_t|y^{t-1})-\ell_i(x)/n \tag{28}$$

for $t=1,\ldots,n$. This is centered for each i and for each text, satisfying $\sum_{t=1}^n\zeta_{it}=0$.

E.2 Kullback-Leibler divergence

The KL divergence for next-token generation in language models, where $y_t\sim p_i(y_t|y^{t-1})$, is given by

$$\mathrm{KL}(p_i(y_t|y^{t-1}),p_j(y_t|y^{t-1}))= \tag{29}$$

$$\sum_{y_t\in\mathcal{V}}p_i(y_t|y^{t-1})\log\frac{p_i(y_t|y^{t-1})}{p_j(y_t|y^{t-1})}. \tag{30}$$

We apply the results for text probability distributions from Section 2 and Appendix D to the conditional probability distributions of token generation. The equation corresponding to (2) is

$$2\mathrm{KL}(p_i(y_t|y^{t-1}),p_j(y_t|y^{t-1}))\approx \underset{y_t\sim p_0(y_t|y^{t-1})}{\mathrm{Var}}\left\{\log\frac{p_i(y_t|y^{t-1})}{p_j(y_t|y^{t-1})}\right\}. \tag{31}$$

The squared Euclidean distance in the $\boldsymbol{\zeta}$-coordinate system provides an estimate of the sum of (31) over all tokens in the text x:

$$\|\boldsymbol{\zeta}_i-\boldsymbol{\zeta}_j\|^2$$

$$\approx 2\sum_{t=1}^n\mathrm{KL}(p_i(y_t|y^{t-1}),p_j(y_t|y^{t-1})) \tag{32}$$

$$\approx 2\mathrm{KL}(p_i,p_j). \tag{33}$$

The proof is provided in Appendix F. To justify the estimation in (32), we assume the following:

$$\underset{y_t\sim p_0(y_t|y^{t-1})}{\mathbb{E}}\left\{\log\frac{p_i(y_t|y^{t-1})}{p_j(y_t|y^{t-1})}\right\} \tag{34}$$

takes a constant value independent of t. In reality, this assumption is not entirely correct, and the degree of variation affects the accuracy of the approximation in (32)[17]. On the other hand, the approximation in (33) holds more generally and is demonstrated in Appendix F.5.

F Theory of Model Coordinates for Token Probability Distributions

In this section, we provide a more detailed explanation of the content discussed in Appendix E. We extend the discussion of text probability distributions in Appendix D to the case of conditional probability distributions for token generation.

F.1 Exponential family of distributions

We apply the same setting as for $p_i(x)$ in Section D to the conditional probability distributions of tokens:

$$y_t \sim p_i(y_t|y^{t-1}), \quad t = 1, \dots, n.$$

The exponential family of distributions incorporating K models, corresponding to (7), is given here as

$$\begin{aligned} p(y_t|y^{t-1}; \boldsymbol{\theta}) &:= p_0(y_t|y^{t-1}) \\ &\exp(\boldsymbol{\theta}^\top \boldsymbol{b}(y_t|y^{t-1}) - \psi(\boldsymbol{\theta}|y^{t-1})). \end{aligned} \tag{35}$$

The setting (8), which associates the K models with (35), is given here as

$$\begin{aligned} &\lambda b_i(y_t|y^{t-1}) := \\ &\quad \log p_i(y_t|y^{t-1}) - \log p_0(y_t|y^{t-1}). \end{aligned} \tag{36}$$

Thus, we have

$$p_i(y_t|y^{t-1}) = p(y_t|y^{t-1}; \lambda \boldsymbol{e}_i)$$

for $i = 1, \dots, K$.

F.2 The variance representation of the KL divergence

The KL divergence is given by (30). Applying the result for the model $p(x; \boldsymbol{\theta})$ in (18) to the token-level conditional distribution model $p(y_t|y^{t-1}; \boldsymbol{\theta})$, we obtain

$$\begin{aligned} &2\mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1})) = \\ &\operatorname*{Var}_{y_t \sim p_0(y_t|y^{t-1})} \left\{ \log \frac{p_i(y_t|y^{t-1})}{p_j(y_t|y^{t-1})} \right\} + O(\lambda^3). \end{aligned} \tag{37}$$

[17]Since this assumption does not affect (3), there is no concern regarding the use of model maps based on text probabilities.

F.3 Two additional assumptions

To estimate the KL divergence from a single text x, two additional assumptions are required, as described below. Such assumptions were not necessary when estimating the KL divergence from the dataset D in Appendix D. In reality, these two assumptions are not strictly satisfied, and the discrepancy between these assumptions and reality affects the accuracy of the KL divergence approximation.

Assumption 1: We assume that the probability distribution of y_t depends only on the past k tokens, denoted as $y_{t-k}^{t-1} = (y_{t-k}, y_{t-k+1}, \dots, y_{t-1})$. That is,

$$p_i(y_t|y^{t-1}) = p_i(y_t|y_{t-k}^{t-1}),$$

which allows us to regard y_{t-k}^t as the state of a Markov chain. More generally, we use the notation y^k to represent a state variable. We consider a function f of the state variable y^k. Furthermore, we assume that this Markov chain is positive Harris recurrent, has a stationary distribution π, and that f is absolutely integrable, i.e.,

$$\mathop{\mathbb{E}}_{y^k \sim \pi}(|f(y^k)|) < \infty.$$

Then, by the strong law of large numbers for Markov chains (Meyn and Tweedie, 2009, Theorem 17.0.1 (i)), in the limit $n \to \infty$,

$$\frac{1}{n} \sum_{t=1}^n f(y_{t-k}^t) \to \mathop{\mathbb{E}}_{y^k \sim \pi}(f(y^k)) \quad \text{a.s.}$$

For simplicity in notation and discussion, we assume that $y_{-k+1}, \dots, y_0$ are appropriately defined. Since the Markov chain converges to π, we also have

$$\frac{1}{n} \sum_{t=1}^n f(y_{t-k}^t) \to \frac{1}{n} \sum_{t=1}^n \mathop{\mathbb{E}}_{y^t \sim p_0}(f(y_{t-k}^t)) \tag{38}$$

almost surely as $n \to \infty$.

Assumption 2:

$$\mathop{\mathbb{E}}_{y_t \sim p_0(y_t|y^{t-1})} \log \frac{p_i(y_t|y^{t-1})}{p_j(y_t|y^{t-1})} = c \tag{39}$$

for some $c \in \mathbb{R}$ that can depend on the indices i and j but not on t. In other words, (39) takes a constant value independent of t.

F.4 Estimation of the KL divergence

Define

$$h(y^t) := \log \frac{p_i(y_t|y^{t-1})}{p_j(y_t|y^{t-1})}.$$

From Assumption 1, $h(y^t)$ can be written in the form $h(y^t) = f_1(y^t_{t-k})$ for some f_1, so applying (38), for sufficiently large n, we obtain

$$\frac{1}{n}\sum_{t=1}^{n} h(y^t) \approx \frac{1}{n}\sum_{t=1}^{n} \mathop{\mathbb{E}}_{y^t\sim p_0}(h(y^t)).$$

Applying (39) to the right-hand side gives

$$\frac{1}{n}\sum_{t=1}^{n} h(y^t) \approx c.$$

Next, since $(h(y^t) - c)^2$ can be written in the form of $f_2(y^t_{t-k})$ for some function f_2, applying (38) again yields

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^{n}(h(y^t) - c)^2\\ \approx&\frac{1}{n}\sum_{t=1}^{n} \mathop{\mathbb{E}}_{y^{t-1}\sim p_0}\Big\{\mathop{\mathrm{Var}}_{y_t\sim p_0(y_t|y^{t-1})}(h(y^t))\Big\}\\ \approx&\frac{1}{n}\sum_{t=1}^{n} \mathop{\mathrm{Var}}_{y_t\sim p_0(y_t|y^{t-1})}(h(y^t)).\end{aligned}$$

In the final equation, we applied (38) using the fact that $\mathrm{Var}_{y_t\sim p_0(y_t|y^{t-1})}(h(y^t)) = f_3(y^t_{t-k})$ for some f_3. Using (37), we obtain

$$\sum_{t=1}^{n}(h(y^t) - c)^2 \approx 2\sum_{t=1}^{n}\mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1})). \tag{40}$$

Meanwhile, the components of the model coordinate $\boldsymbol{\zeta}_i$ are given by

$$\zeta_{it} = \log p_i(y_t|y^{t-1}) - c_i$$

where

$$c_i = \frac{1}{n}\sum_{t=1}^{n}\log p_i(y_t|y^{t-1}).$$

Since

$$\zeta_{it} - \zeta_{jt} = h(y^t) - (c_i - c_j)$$

with $c_i - c_j \approx c$, equation (40) can be rewritten as

$$\|\boldsymbol{\zeta}_i - \boldsymbol{\zeta}_j\|^2 \approx 2\sum_{t=1}^{n}\mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1})).$$

Thus, (32) is established.

F.5 Connecting the KL divergence of token and text probability distributions

Here, we fix the sequence length of the text $x = (y_1, \dots, y_n)$ as n, i.e., we set $\mathcal{X} = \mathcal{V}^n$. For notational simplicity, we define

$$g_i(y^t) = \log p_i(y_t|y^{t-1}).$$

Noting that

$$p_i(x) = \prod_{t=1}^{n} p_i(y_t|y^{t-1}) = \prod_{t=1}^{n} e^{g_i(y^t)},$$

we obtain

$$\begin{aligned}&\mathrm{KL}(p_i, p_j)\\ &= \sum_{x\in\mathcal{X}}\prod_{t'=1}^{n} e^{g_i(y^{t'})}\sum_{t=1}^{n}(g_i(y^t) - g_j(y^t))\\ &= \sum_{t=1}^{n}\sum_{y^t\in\mathcal{V}^t}\prod_{t'=1}^{t} e^{g_i(y^{t'})}(g_i(y^t) - g_j(y^t))\\ &= \sum_{t=1}^{n}\sum_{y^{t-1}\in\mathcal{V}^{t-1}}\prod_{t'=1}^{t-1} e^{g_i(y^{t'})}\\ &\qquad \sum_{y_t\in\mathcal{V}} e^{g_i(y^t)}(g_i(y^t) - g_j(y^t))\\ &= \sum_{t=1}^{n}\sum_{y^{t-1}\in\mathcal{V}^{t-1}} p_i(y^{t-1})\\ &\qquad \mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1}))\\ &= \sum_{t=1}^{n}\mathop{\mathbb{E}}_{y^{t-1}\sim p_i}\Big\{\mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1}))\Big\}\\ &= \mathop{\mathbb{E}}_{x\sim p_i}\Big\{\sum_{t=1}^{n}\mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1}))\Big\}.\end{aligned}$$

Thus, for sufficiently large n, by the strong law of large numbers for Markov chains, we obtain

$$\mathrm{KL}(p_i, p_j) \approx \sum_{t=1}^{n}\mathrm{KL}(p_i(y_t|y^{t-1}), p_j(y_t|y^{t-1})). \tag{41}$$

This corresponds to (33). Assumption 1 from Appendix F.3 is used in (41), but Assumption 2 is not needed in the discussion of this subsection.

G Details of Experiments

G.1 Information obtained via the Hugging Face Hub API

We used the Hugging Face Hub API to retrieve information about each language model's tags, the

Model type	Models
llama-1	69
llama-2	223
llama-3	62
llama	217
mistral	232
gpt_neox	54
deepseek	26
gptj	19
gemma	18
opt	15
bloom	12
falcon	11
qwen2	10
mixtral	9
mpt	6
stablelm	6
gpt_neo	3
phi	3
gpt_bigcode	3
phi3	3
xglm	3
rwkv	3
starcoder2	2
olmo	2
camelidae	2
codegen	2
deci	1
recurrent_gemma	1
stablelm_alpha	1
Total	1,018

Table 4: Number of models by model type.

date the model was created, the number of downloads over the past 30 days, and the model's configuration details. All of this information is current as of February 1, 2025.

Among the model tags, we specifically used llama2, llama-2, license:llama2, llama3, llama-3, and license:llama3 to determine the model type (llama-1, llama-2, or llama-3). Furthermore, to identify language models that were pre-trained on the Pile in Section 4, we employed tags such as dataset:eleutherai/pile, dataset:eleutherai/the_pile, dataset:eleutherai/the_pile_deduplicated, and arxiv:2101.00027.

G.2 How the model type was determined

In principle, we used the value of model_type in the config retrieved from the Hugging Face Hub API as the model type. However, out of the 1,018 language models we examined, there were 587 whose config model_type was listed as llama. For these, we used the following procedure to determine whether they were the original Llama (llama-1), Llama-2, or Llama-3; if we were able to identify which version they were, we reclassified them as llama-1, llama-2, or llama-3 accordingly.

1. We checked the tags assigned to each model. Of these, 136 models that included any of llama2, llama-2, or license:llama2 were classified as llama-2. Similarly, 39 models that included any of llama3, llama-3, or license:llama3 were classified as llama-3.

2. For the remaining 412 models whose classification was not determined by tags alone, we used the creation date and the model name (converted to lowercase) to make a decision. First, 69 models that were created prior to July 18, 2023 (the Llama-2 release date) were classified as llama-1. Next, 88 models whose lowercase model name contained either llama2 or llama-2 were classified as llama-2. Among those whose lowercase model name contained llama3 or llama-3, 22 models whose creation date was after April 18, 2024 (the Llama-3 release date) were classified as llama-3.

3. After following the steps above, the 217 models that could not be classified were left as llama.

Furthermore, any model whose name prior to the slash (/) was deepseek-ai was defined as deepseek. In addition, even though abacusai/Llama-3-Smaug-8B was tagged with license:llama2, we manually reclassified it as llama-3.

Table 4 shows the number of models classified into each model type.

G.3 Basic information on the dataset

The dataset used in our experiments consists of a total of 10,000 texts, which are divided into 17 text categories. Table 5 shows the number of texts in each category.

Text category	Texts
Pile-CC	2,353
PubMed Central	1,763
ArXiv	1,172
Github	925
FreeLaw	837
StackExchange	712
Wikipedia (en)	567
USPTO Backgrounds	487
PubMed Abstracts	464
Gutenberg (PG-19)	251
DM Mathematics	151
EuroParl	83
HackerNews	67
Ubuntu IRC	54
PhilPapers	51
NIH ExPorter	41
Enron Emails	22
Total	10,000

Table 5: Number of texts in each text category.

To assign colors to the text categories, we first compute the average text embedding for each category in the Pile using `simcse-roberta-large` (Gao et al., 2021b). Next, we calculate a tour over the 17 average embedding vectors by solving the traveling salesman problem[18]. The TSP is solved using the nearest neighbor method to generate an initial tour, which is then refined using a 2-opt improvement procedure, and Euclidean distance is used as the metric. Based on the adjacency relationships along this tour, we segment the hue circle at equal intervals and color each category accordingly.

G.4 Standard scores

In the three experiments described below, we use standardized values[19], or Z-score normalization, of both the log-likelihood and the benchmark scores calculated for the K language models.

- In Figures 1 and 11, where we define each language model's primary text category, we use the average log-likelihood for each category, standardized across all models.

[18]Inspired by Sato (2022); Yamagiwa et al. (2024).

[19]By subtracting the mean from each value and dividing by the standard deviation, the data is transformed to have a mean of 0 and a variance of 1.

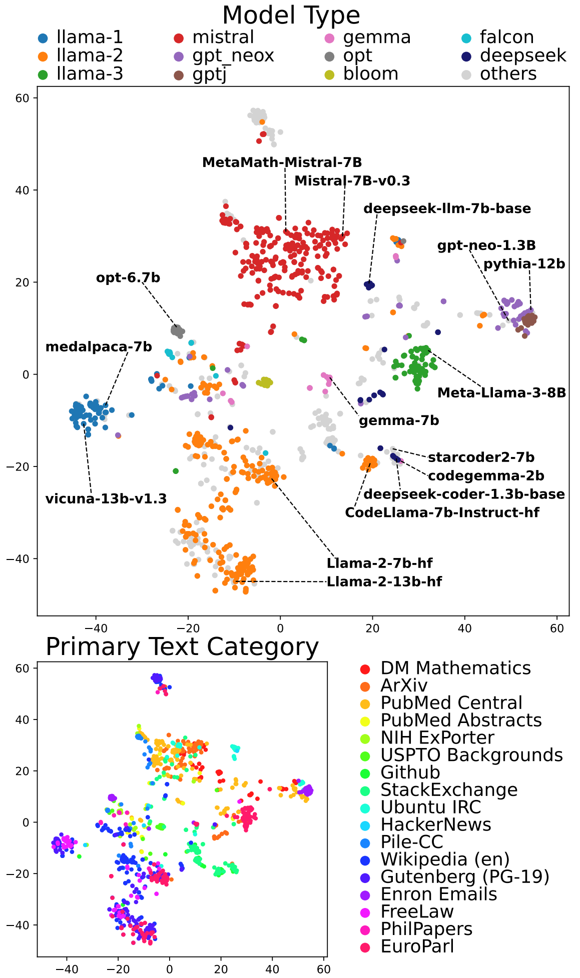

Figure 11: Model maps obtained by double centered log-likelihood matrix $\boldsymbol{Q}$. These maps correspond to Figure 1. (Top) Colors indicate model types. (Bottom) Colors indicate the text category in which each model attains the highest standardized log-likelihood score among 17 categories.

- In Section 4, to determine each language model's primary task, we standardize each task's score across the K models.

- Furthermore, in the data leakage detection described in Section 4, we use the difference between the standardized mean log-likelihood and the standardized 6-TaskMean score as the indicator.

G.5 Hierarchical clustering settings

Figure 3 displays the double-centered log-likelihood matrix $\boldsymbol{Q}$, with hierarchical clustering applied to both its rows and columns. We implemented the clustering using SciPy (Virtanen et al., 2020). Distance matrices were computed using `scipy.spatial.distance.pdist`,

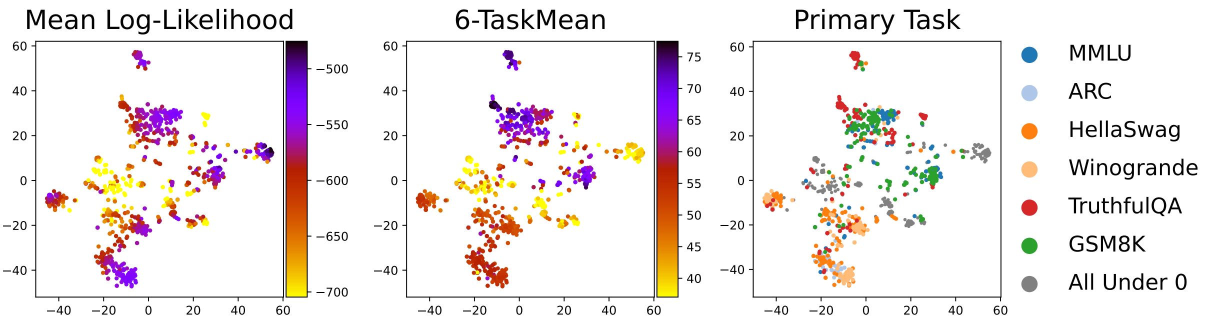

Figure 12: Model maps obtained by double centered log-likelihood matrix $\boldsymbol{Q}$. These maps correspond to Figure 5. These maps are illustrating model performance. From left to right, the panels show each model's mean log-likelihood, 6-TaskMean score, and the "primary task," which refers to the task where each model achieves the highest standardized score among the six tasks, color-coded accordingly. The color bar is clipped at the 10th percentile for mean log-likelihood and 6-TaskMean, with darker colors indicating better performance. In the primary task panel, models with standardized scores below zero across all six tasks are labeled as "All Under 0."

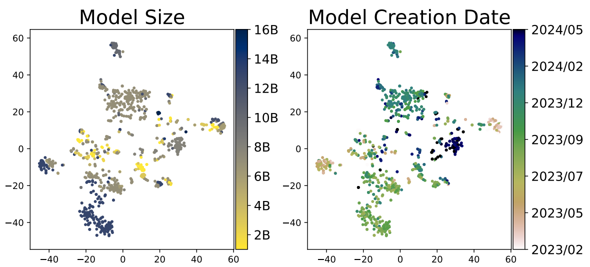

Figure 13: Model maps obtained by double centered log-likelihood matrix $\boldsymbol{Q}$ color-coded by (Left) model size and (Right) model creation date.

and clustering was performed using `scipy.cluster.hierarchy.linkage`. For clustering models, we used `sqeuclidean` as the metric and `median` as the linkage method. For clustering texts, we used `correlation` as the metric and `average` as the linkage method. For the hierarchical clustering shown in Fig. 4, which presents a dendrogram of 100 language models, we used `sqeuclidean` as the metric and `median` as the linkage method. The vertical axis of the dendrogram uses the symmetric logarithmic scale (with a linear threshold of 250) implemented in `matplotlib`. To ensure that the values on the vertical axis correspond to the Kullback-Leibler divergence, we used the $\boldsymbol{q}$-coordinates divided by $\sqrt{2N}$.

H Additional Model Maps

In this section, we present additional model maps, including a figure that lists all the model names and a map obtained through dimensionality reduction of the double-centered log-likelihood matrix $\boldsymbol{Q}$.

H.1 Model map via the double centered log-likelihood matrix

In the main text, we use a model map generated by dimensionality reduction of the log-likelihood matrix $\boldsymbol{L}$. Here, in Figs. 11, 12 and 13, we present model maps obtained by dimensionality reduction of the double-centered log-likelihood matrix $\boldsymbol{Q}$.

H.2 Model map with model names

We present figures that display the model names corresponding to each point on the model maps defined in Section 4. For both the log-likelihood matrix $\boldsymbol{L}$ and the double-centered log-likelihood matrix $\boldsymbol{Q}$, we provide two types of maps: one colored by model type and another colored by primary text category. Figures 14 and 15 show the maps obtained by applying dimensionality reduction to $\boldsymbol{L}$, while Figures 16 and 17 show the maps obtained using $\boldsymbol{Q}$.

I Table of Nearest Neighbor Models

Table 6 presents the top 10 nearest neighbors among the 1,018 language models for each of the models highlighted in the top panel of Figure 1. Each sub-table corresponds to a specific model of interest, listing its nearest neighbors in descending order of the KL divergence. From this table, we can see that similar models tend to cluster together, exhibiting relatively small KL divergence values. Additionally, the values in parentheses denote the KL divergence measured in bits per byte of text. This conversion makes it easier to interpret the information cost per byte when comparing predictive distributions of different models.

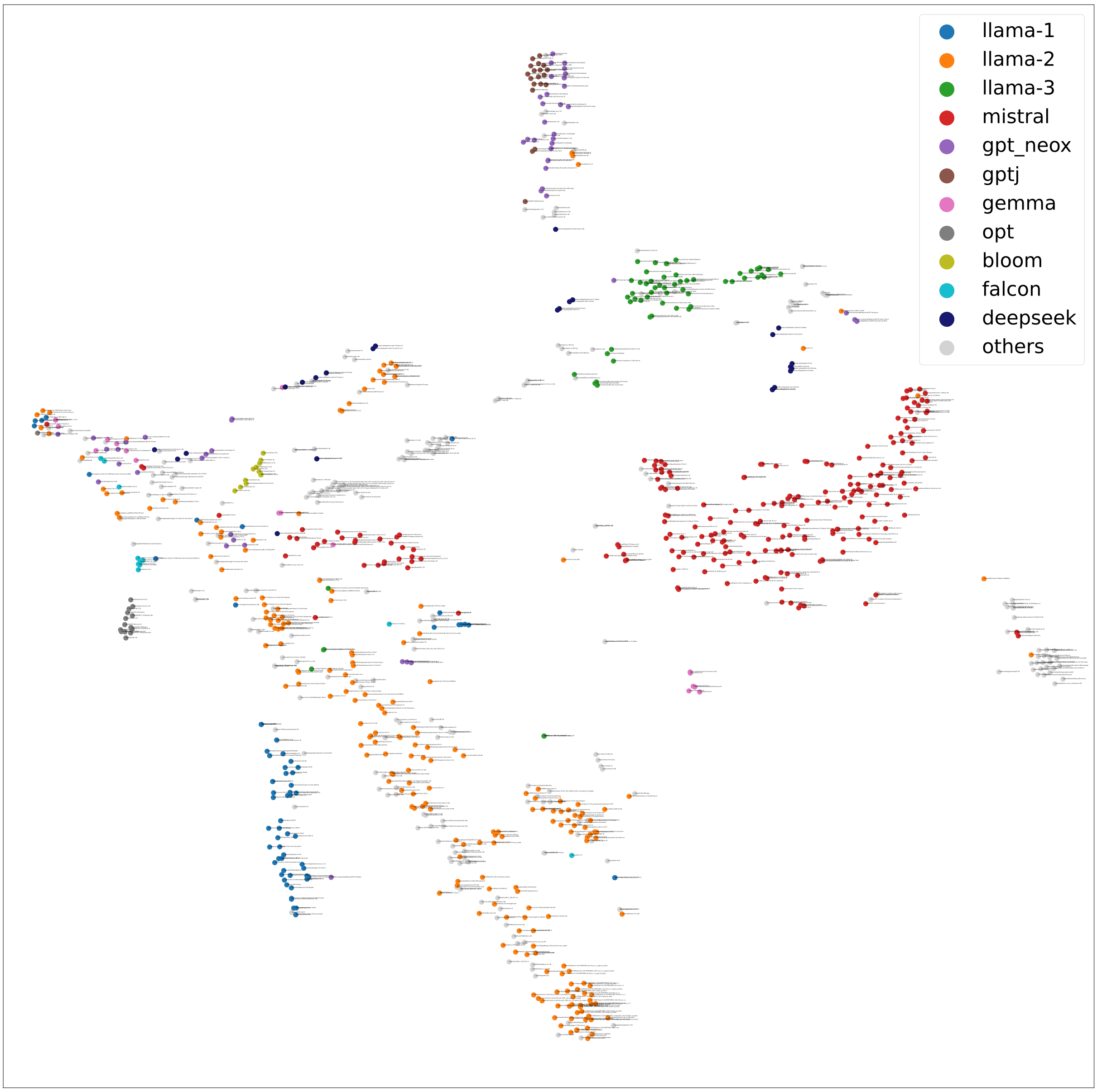

Figure 14: Model map obtained by dimensionality reduction of the log-likelihood matrix $\boldsymbol{L}$. Each point on the model map is labeled with the corresponding model name.

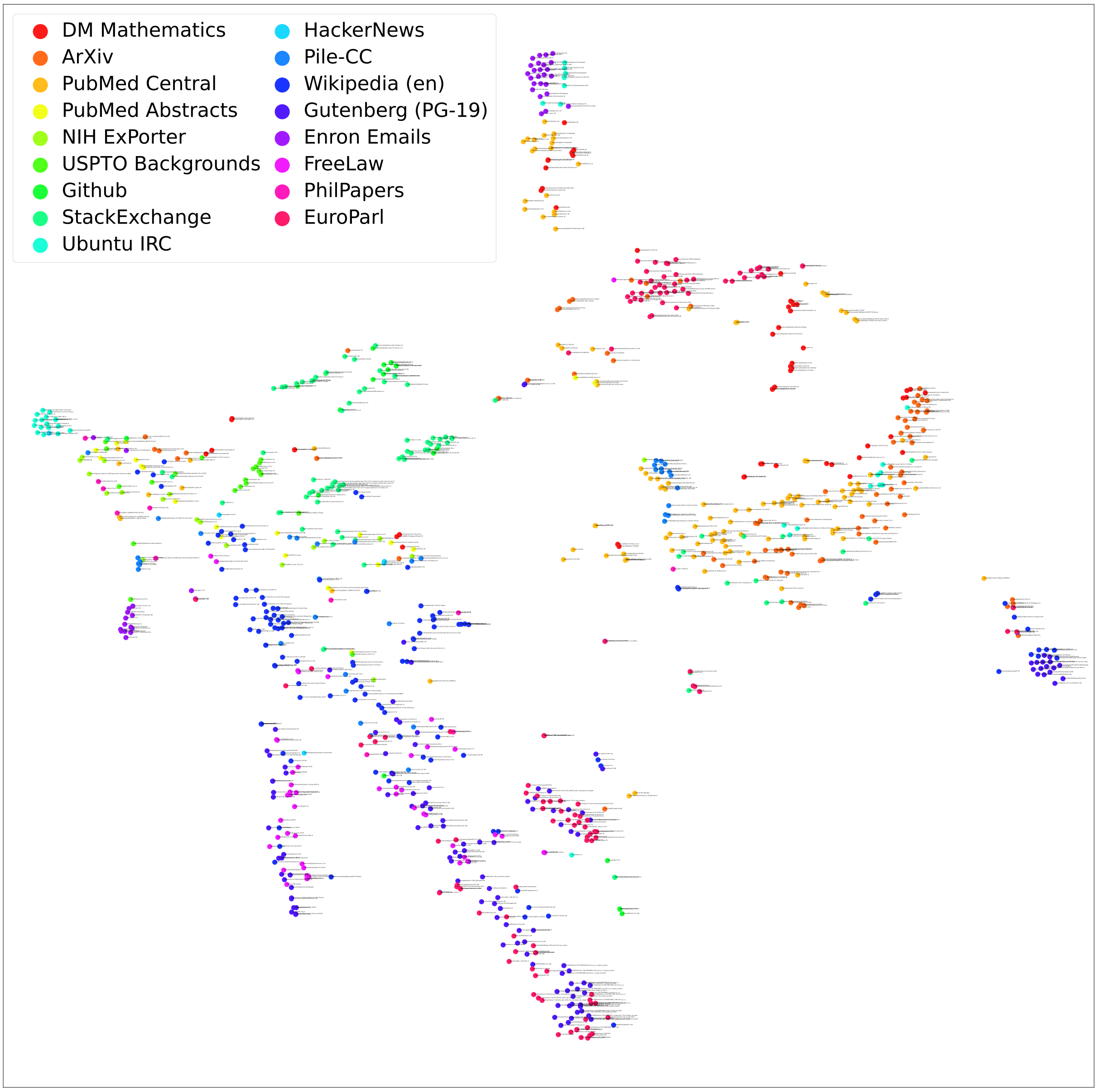

Figure 15: Model map obtained by dimensionality reduction of the log-likelihood matrix $\boldsymbol{L}$. Each point on the model map is labeled with the corresponding model name. Colors indicate the model's "primary text category," the text category where the model achieves the highest standardized log-likelihood score among 17 categories.

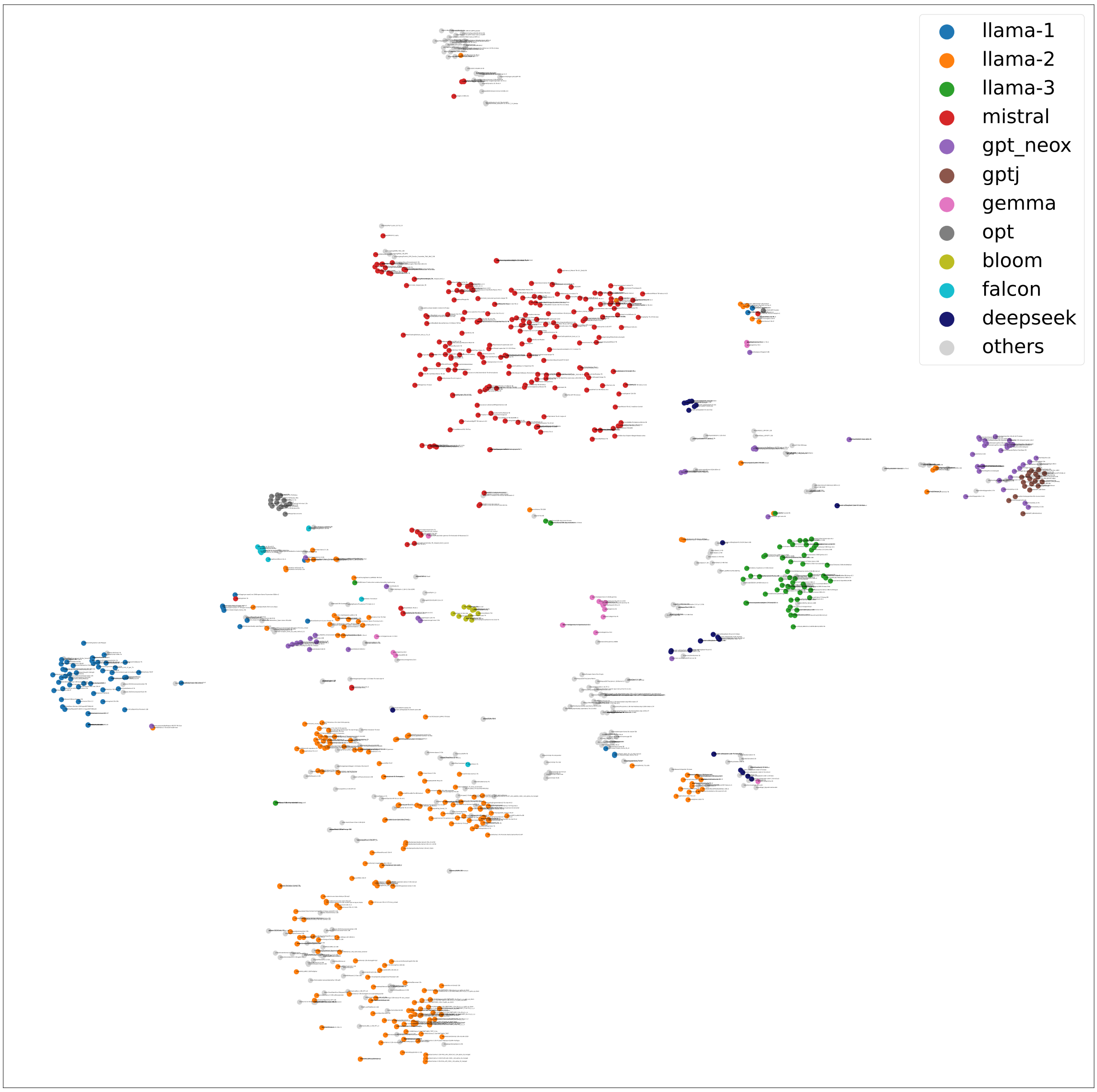

Figure 16: Model map obtained by dimensionality reduction of the double-centered log-likelihood matrix $\boldsymbol{Q}$. Each point on the model map is labeled with the corresponding model name.

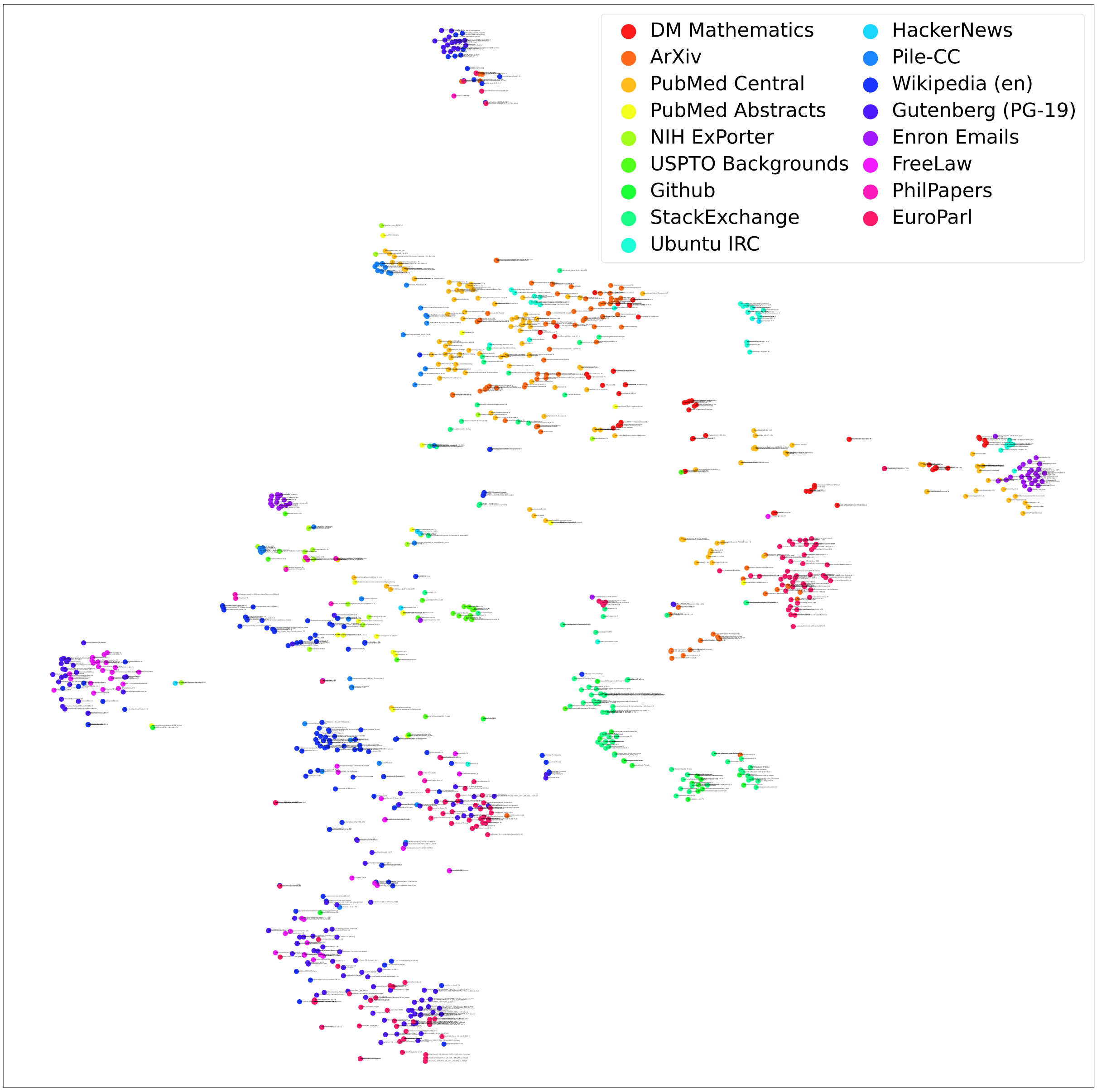

Figure 17: Model map obtained by dimensionality reduction of the double-centered log-likelihood matrix $\boldsymbol{Q}$. Each point on the model map is labeled with the corresponding model name. Colors indicate the model's "primary text category," the text category where the model achieves the highest standardized log-likelihood score among 17 categories.

bigcode/starcoder2-7b	KL [bpb]	codellama/CodeLlama-7b-Instruct-hf	KL [bpb]
bigcode/starcoder2-3b	1.14	codellama/CodeLlama-7b-hf	0.0715
stabilityai/stable-code-3b	2.38	NousResearch/CodeLlama-7b-hf	0.0715
deepseek-ai/deepseek-coder-6.7b-base	2.50	codellama/CodeLlama-13b-Instruct-hf	0.206
EleutherAI/llemma_7b	2.81	TheBloke/CodeLlama-13B-Instruct-fp16	0.206
deepseek-ai/deepseek-coder-7b-base-v1.5	3.01	Nexusflow/NexusRaven-V2-13B	0.218
deepseek-ai/DeepSeek-Coder-V2-Lite-Base	3.03	codellama/CodeLlama-13b-hf	0.236
deepseek-ai/deepseek-coder-7b-instruct-v1.5	3.03	NousResearch/CodeLlama-13b-hf	0.236
deepseek-ai/deepseek-coder-6.7b-instruct	3.18	OpenAssistant/codellama-13b-oasst-sft-v10	0.326
deepseek-ai/DeepSeek-Coder-V2-Lite-Instruct	3.19	HiTZ/GoLLIE-7B	0.335
meta-math/MetaMath-Llemma-7B	3.24	WhiteRabbitNeo/WhiteRabbitNeo-13B-v1	0.386
deepseek-ai/deepseek-coder-1.3b-base	**KL [bpb]**	**deepseek-ai/deepseek-llm-7b-base**	**KL [bpb]**
deepseek-ai/deepseek-coder-1.3b-instruct	0.610	deepseek-ai/deepseek-moe-16b-base	0.194
deepseek-ai/deepseek-coder-6.7b-instruct	0.761	deepseek-ai/deepseek-llm-7b-chat	0.364
deepseek-ai/deepseek-coder-6.7b-base	0.892	deepseek-ai/deepseek-moe-16b-chat	0.368
bigcode/starcoderbase-1b	1.185	deepseek-ai/DeepSeek-V2-Lite	0.455
bigcode/starcoderbase-7b	1.855	deepseek-ai/ESFT-vanilla-lite	0.456
NTQAI/Nxcode-CQ-7B-orpo	1.940	deepseek-ai/DeepSeek-V2-Lite-Chat	0.868
Qwen/CodeQwen1.5-7B-Chat	1.954	mistralai/Mistral-7B-Instruct-v0.1	1.356
bigcode/starcoder2-3b	2.757	statking/zephyr-7b-sft-full-orpo	1.616
Salesforce/codegen-6B-multi	2.929	Severian/ANIMA-Phi-Neptune-Mistral-7B	1.692
google/codegemma-2b	3.139	sethuiyer/Medichat-Llama3-8B	1.718
EleutherAI/gpt-neo-1.3B	**KL [bpb]**	**EleutherAI/pythia-12b**	**KL [bpb]**
EleutherAI/gpt-neo-2.7B	0.325	matsuo-lab/weblab-10b	0.207
EleutherAI/pythia-1.4b	0.429	h2oai/h2ogpt-oig-oasst1-256-6_9b	0.237
EleutherAI/pythia-1b-deduped	0.613	Salesforce/codegen-6B-nl	0.260
HWERI/pythia-1.4b-deduped-sharegpt	0.625	EleutherAI/gpt-j-6b	0.277
beaugogh/pythia-1.4b-deduped-sharegpt	0.625	TehVenom/Dolly_Malion-6b	0.344
EleutherAI/pythia-1.4b-deduped	0.663	TehVenom/PPO_Shygmalion-6b	0.346
PygmalionAI/metharme-1.3b	0.666	TehVenom/Dolly_Shygmalion-6b-Dev_V8P2	0.351
RWKV/rwkv-raven-1b5	0.761	TehVenom/PPO_Pygway-V8p4_Dev-6b	0.352
databricks/dolly-v2-3b	0.780	TehVenom/GPT-J-Pyg_PPO-6B-Dev-V8p4	0.364
EleutherAI/pythia-2.8b-deduped	0.872	TehVenom/PPO_Shygmalion-V8p4_Dev-6b	0.364
facebook/opt-6.7b	**KL [bpb]**	**google/codegemma-2b**	**KL [bpb]**
KoboldAI/OPT-6.7B-Erebus	0.00102	deepseek-ai/deepseek-coder-1.3b-instruct	2.46
KoboldAI/OPT-6.7B-Nerybus-Mix	0.117	bigcode/starcoderbase-1b	2.69
KoboldAI/OPT-6B-nerys-v2	0.185	deepseek-ai/deepseek-coder-1.3b-base	3.14
facebook/opt-13b	0.292	bigcode/gpt_bigcode-santacoder	3.92
KoboldAI/OPT-13B-Nerybus-Mix	0.341	deepseek-ai/deepseek-coder-6.7b-instruct	3.97
KoboldAI/OPT-13B-Erebus	0.353	Qwen/CodeQwen1.5-7B-Chat	4.09
KoboldAI/OPT-13B-Nerys-v2	0.406	NTQAI/Nxcode-CQ-7B-orpo	4.11
facebook/opt-2.7b	0.486	Salesforce/codegen-6B-multi	4.24
KoboldAI/OPT-2.7B-Nerybus-Mix	0.634	bigcode/starcoderbase-7b	4.44
KoboldAI/OPT-2.7B-Erebus	0.663	deepseek-ai/deepseek-coder-6.7b-base	4.87
google/gemma-7b	**KL [bpb]**	**lmsys/vicuna-13b-v1.3**	**KL [bpb]**
SeaLLMs/SeaLLM-7B-v2.5	0.367	TheBloke/stable-vicuna-13B-HF	0.290
VAGOsolutions/SauerkrautLM-Gemma-7b	0.383	Yhyu13/chimera-inst-chat-13b-hf	0.292
lemon-mint/gemma-ko-7b-instruct-v0.62	0.511	junelee/wizard-vicuna-13b	0.325
google/gemma-2b	0.748	TheBloke/wizard-vicuna-13B-HF	0.325
google/codegemma-7b	0.885	TheBloke/UltraLM-13B-fp16	0.326
PathFinderKR/Waktaverse-Llama-3-KO-8B-Instruct	1.209	NousResearch/Nous-Hermes-13b	0.328
FlagAlpha/Llama3-Chinese-8B-Instruct	1.214	project-baize/baize-v2-13b	0.347
FairMind/Llama-3-8B-4bit-UltraChat-Ita	1.236	TheBloke/guanaco-13B-HF	0.347
migtissera/Llama-3-8B-Synthia-v3.5	1.251	openaccess-ai-collective/minotaur-13b-fixed	0.349
Orenguteng/Llama-3-8B-Lexi-Uncensored	1.253	openaccess-ai-collective/wizard-mega-13b	0.353
medalpaca/medalpaca-7b	**KL [bpb]**	**meta-llama/Llama-2-13b-hf**	**KL [bpb]**
TheBloke/guanaco-7B-HF	0.424	TaylorAI/Flash-Llama-13B	1.82e-16
eachadea/vicuna-7b-1.1	0.499	TheBloke/Llama-2-13B-fp16	3.6e-06
TehVenom/Pygmalion-Vicuna-1.1-7b	0.548	StudentLLM/Alpagasus-2-13b-QLoRA-merged	0.00668
lmsys/vicuna-7b-v1.3	0.556	CHIH-HUNG/llama-2-13b-FINETUNE2_TEST_2.2w	0.0113
jphme/orca_mini_v2_ger_7b	0.569	garage-bAInd/Platypus2-13B	0.0154
ajibawa-2023/Uncensored-Jordan-7B	0.570	CHIH-HUNG/llama-2-13b-dolphin_5w	0.0213
bofenghuang/vigogne-7b-instruct	0.580	CHIH-HUNG/llama-2-13b-OpenOrca_5w	0.0222
TheBloke/airoboros-7b-gpt4-fp16	0.582	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r16-gate_up_down-test1	0.0256
TheBloke/tulu-7B-fp16	0.628	CHIH-HUNG/llama-2-13b-FINETUNE4_addto15k_4.5w-r16-gate_up_down	0.0258
TehVenom/Pygmalion_AlpacaLora-7b	0.640	CHIH-HUNG/llama-2-13b-FINETUNE4_compare15k_4.5w-r16-gate_up_down	0.0288
meta-llama/Llama-2-7b-hf	**KL [bpb]**	**meta-llama/Meta-Llama-3-8B**	**KL [bpb]**
ibranze/araproje-llama2-7b-hf	0	Undi95/Meta-Llama-3-8B-hf	4.56e-06
TheTravellingEngineer/llama2-7b-chat-hf-v4	0	dfurman/Llama-3-8B-Orpo-v0.1	0.0104
TheTravellingEngineer/llama2-7b-chat-hf-v2	0	migtissera/Tess-2.0-Llama-3-8B	0.0428
TaylorAI/Flash-Llama-7B	0	freewheelin/free-llama3-dpo-v0.2	0.101
yeen214/test_llama2_7b	0	jondurbin/bagel-8b-v1.0	0.164
NewstaR/Starlight-7B	4.57e-06	migtissera/Llama-3-8B-Synthia-v3.5	0.190
Delcos/Mistral-Pygmalion-7b	0.0220	nvidia/Llama3-ChatQA-1.5-8B	0.191
elliotthwang/elliott_Llama-2-7b-hf	0.0246	ruslanmv/Medical-Llama3-8B	0.206
garage-bAInd/Platypus2-7B	0.0338	FairMind/Llama-3-8B-4bit-UltraChat-Ita	0.296
Lazycuber/L2-7b-Base-Guanaco-Uncensored	0.0346	NousResearch/Hermes-2-Theta-Llama-3-8B	0.330
meta-math/MetaMath-Mistral-7B	**KL [bpb]**	**mistralai/Mistral-7B-v0.3**	**KL [bpb]**
Weyaxi/MetaMath-NeuralHermes-2.5-Mistral-7B-Linear	0.0639	MaziyarPanahi/Mistral-7B-v0.3	3.64e-16
Weyaxi/MetaMath-Tulpar-7b-v2-Slerp	0.121	mistral-community/Mistral-7B-v0.2	0.0107
Weyaxi/MetaMath-OpenHermes-2.5-neural-chat-v3-3-Slerp	0.150	unsloth/mistral-7b-v0.2	0.0107
Q-bert/Bumblebee-7B	0.158	mistralai/Mistral-7B-v0.1	0.0327
OpenPipe/mistral-ft-optimized-1227	0.163	Cartinoe5930/Llama2_init_Mistral	0.0442
Toten5/Marcoroni-neural-chat-7B-v2	0.169	Locutusque/Hercules-3.1-Mistral-7B	0.0476
ignos/Mistral-T5-7B-v1	0.170	migtissera/Synthia-7B-v3.0	0.0496
Weyaxi/MetaMath-Chupacabra-7B-v2.01-Slerp	0.170	uukuguy/speechless-zephyr-code-functionary-7b	0.0530
Q-bert/Optimus-7B	0.175	uukuguy/zephyr-7b-alpha-dare-0.85	0.0530
Weyaxi/MetaMath-NeuralHermes-2.5-Mistral-7B-Ties	0.189	crumb/apricot-wildflower-20	0.0607

Table 6: Top 10 nearest neighbors among the 1,018 language models for each model labeled in the top panel of Fig. 1. The values indicate the KL divergence measured in bits per byte (bpb), as defined in Section 3.5. These are computed using formula (3) in Section 2, by multiplying the original KL divergence by 0.001484.

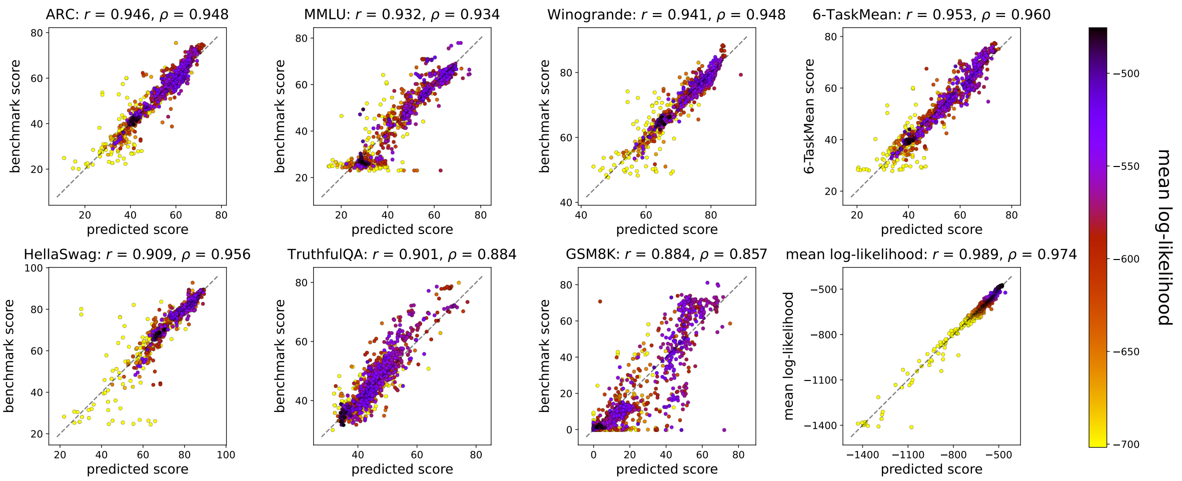

Figure 18: Scatter plots of predicted scores versus benchmark scores for test sets across the six benchmark tasks (the dashed line indicates the identity line). Additionally, results for predicting 6-TaskMean (identical to Fig. 8) and the mean log-likelihood are also shown. Each point is color-coded by the mean log-likelihood, with higher mean log-likelihood values generally corresponding to higher task scores. For better visualization, the color bar range is clipped to the 10th–100th percentile.

J Details of Weight Interpolation

In this section, we describe the experimental details concerning the relationship between model coordinates and model weights, as introduced in Section 6.3.

Constructing the weight grid. Let p_0 denote the base model, and p_1 and p_2 denote the fine-tuned models derived from p_0. We denote the weight parameter vectors of these models as W_0, W_1, and W_2, respectively. To construct the weight grid, we merged the model weights using the following linear operation:

$$W_{\alpha,\beta} = W_0 + \alpha(W_1 - W_0) + \beta(W_2 - W_0), \quad (42)$$

where the merge ratios $\alpha, \beta \in \mathbb{R}$ were chosen from 36 evenly spaced combinations within the interval $[0, 1]$: $\{0.0, 0.2, 0.6, 0.8, 1.0\}$. The original models W_0, W_1, and W_2 correspond to $W_{0,0}$, $W_{1,0}$, and $W_{0,1}$, respectively. When $\alpha + \beta \leq 1$, the operation corresponds to linear interpolation between models. Even among models with the same architecture, the sizes of the embedding/unembedding matrices may differ. In such cases, we truncated or reshaped the weight parameters to match the base model.

Computing model coordinates. For each composed model $p_{\alpha,\beta}$ with weight $W_{\alpha,\beta}$, we computed the model coordinates $\boldsymbol{q}_{\alpha,\beta}$ following the method described in Section 2. The tokenizer of the base model was used to ensure consistency. The text data consists of all 10,000 texts prepared in Section 3.1. Additionally, the deterministic algorithms option[20] was enabled in the implementation to ensure reproducibility.

Selection of models. We conducted experiments using two base models: `Llama-2-7b-hf` and `mistralai/Mistral-7B-v0.1`. For each base model p_0, we selected the two most downloaded fine-tuned models available on Hugging Face. Specifically, when p_0 = `Llama-2-7b-hf`, we set p_1 = `vicuna-7b-v1.5` and p_2 = `Llama-2-7b-chat-hf`. When p_0 = `mistralai/Mistral-7B-v0.1`, we set p_1 = `HuggingFaceH4/zephyr-7b-beta` and p_2 = `mistralai/Mistral-7B-Instruct-v0.1`.

Visualization. Figures 10 and 19 show the linearly merged models, visualized in both weight space and log-likelihood space. The corners of the grid are labeled with their corresponding (α, β) values.

In the left panel, we visualized $W_{\alpha,\beta}$ in the weight space. Since the dimensionality of $W_{\alpha,\beta}$, i.e., the number of model parameters, is extremely high, we employed a 2D projection method using the norms of the difference vectors $r_1 = \|W_1 - W_0\|_2$, $r_2 = \|W_2 - W_0\|_2$, and the angle between them, $\phi = \arccos((W_1 - W_0)^\top (W_2 - W_0)/r_1 r_2)$. Each point was placed at $(\alpha r_1 + \beta r_2 \cos\phi, \beta r_2 \sin\phi)$.

In the right panel, we visualized the model coordinates $\boldsymbol{q}_{\alpha,\beta}$ by Principal Component Analysis

[20] `torch.use_deterministic_algorithms(True)`

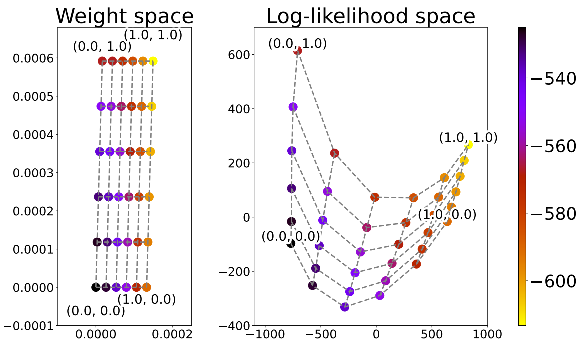

Figure 19: Visualization of 36 language models obtained by linearly interpolating pretrained model weights based on `mistralai/Mistral-7B-v0.1`. Each point is color-coded according to its mean log-likelihood. (Left) Models in the weight parameter space. (Right) Models in the log-likelihood space, represented by the $\boldsymbol{q}$-coordinate system.

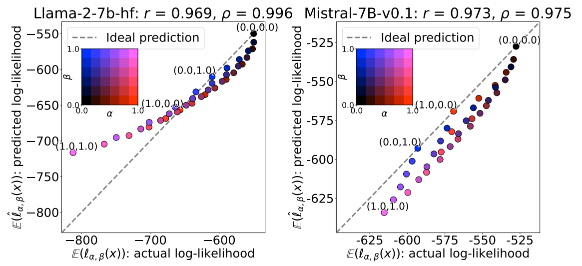

Figure 20: Scatter plots comparing the actual and predicted mean log-likelihoods for 36 interpolated models derived from `Llama-2-7b-hf` (left) and `Mistral-7B-v0.1` (right). Each point corresponds to a unique combination of merge ratios (α, β) and is color-coded accordingly. The dashed line indicates ideal prediction. Strong correlations are observed for both model sets: Llama-2 shows $r = 0.969$, $\rho = 0.996$, and Mistral shows $r = 0.973$, $\rho = 0.975$. A 2D color map of (α, β) values is shown as an inset.

(PCA). Each model $p_{\alpha,\beta}$ was mapped onto the $\boldsymbol{q}$-coordinate system to analyze the structure of the interpolated models.

Estimating the mean log-likelihood of interpolated models. To assess the potential of predicting task performance from log-likelihood vectors without explicitly computing them for every interpolated model, we evaluated whether the mean log-likelihoods can be estimated by linearly interpolating those of the base models. Let $\ell_{\alpha,\beta}(x)$ denote the log-likelihood of input x computed by the merged model $p_{\alpha,\beta}$. Let ℓ_0, ℓ_1, and ℓ_2 be the log-likelihoods from the base and fine-tuned models p_0, p_1, and p_2, respectively. We define the linearly interpolated log-likelihood as:

$$\hat{\ell}_{\alpha,\beta} = \ell_0 + \alpha(\ell_1 - \ell_0) + \beta(\ell_2 - \ell_0). \quad (43)$$

Figure 20 shows a strong correlation between the predicted and actual mean log-likelihoods. This result suggests that log-likelihood vectors can be approximated by linear interpolation, enabling efficient prediction of model performance without computing log-likelihoods for every model.

K Details of Model Performance Prediction

This section provides additional details on the prediction of benchmark scores using model coordinates, as discussed in Section 5.

K.1 Details of ridge regression

As described in Section 5.2, ridge regression requires setting a regularization strength parameter, α. To determine α from $\{10^1, \ldots, 10^9\}$, we performed a five-fold cross-validation within each training dataset[21]. As a post-processing step, we clipped the predicted scores to the range $[0, 100]$.

For the setting where the target variable $\boldsymbol{v} \in \mathbb{R}^K$ was replaced with the mean log-likelihood $(\bar{\ell}_1, \ldots, \bar{\ell}_K) \in \mathbb{R}^K$, we searched for α within $\{10^{-4}, \ldots, 10^4\}$ and did not apply clipping as a post-processing step.

K.2 Details of prediction results

Figure 18 shows scatter plots of predicted scores and actual benchmark scores for each benchmark task, as well as for 6-TaskMean and the mean log-likelihood. As in Fig. 8, the scatter plots show strong correlations for all six benchmark tasks and for the mean log-likelihood.

To account for randomness, we ran five different data splits based on model types when predicting each benchmark score. As explained in Section 5, the final predicted score was the average of these five runs. Additionaly, for each split, we computed the correlation coefficients between the predicted scores and the actual benchmark scores. Table 7 presents their mean and standard deviation, and shows a similar trend as Table 2.

K.3 Results from different settings

To investigate potential data leakage due to model types, Table 8 shows the correlation coefficients

[21] The training dataset consisted of four out of the five folds obtained by splitting the dataset for each benchmark task.

	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average	mean log-likelihood
Pearson's r	0.941 ± 0.002	0.904 ± 0.005	0.924 ± 0.006	0.889 ± 0.013	0.934 ± 0.005	0.864 ± 0.019	0.947 ± 0.005	0.987 ± 0.006
Spearman's ρ	0.944 ± 0.005	0.951 ± 0.003	0.926 ± 0.005	0.875 ± 0.003	0.943 ± 0.009	0.841 ± 0.013	0.956 ± 0.004	0.971 ± 0.006

Table 7: Group 5-fold based on model types using $\boldsymbol{Q}$: Mean and standard deviation of the correlation coefficients between predicted and actual benchmark scores. Coefficients were obtained by ridge regression under five data splits based on model types. Results for predicting 6-TaskMean and the mean log-likelihood are also included.

	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average	mean log-likelihood
Pearson's r	0.968 ± 0.001	0.939 ± 0.004	0.960 ± 0.002	0.952 ± 0.001	0.960 ± 0.003	0.931 ± 0.001	0.973 ± 0.001	0.994 ± 0.001
Spearman's ρ	0.972 ± 0.001	0.970 ± 0.002	0.966 ± 0.001	0.929 ± 0.001	0.969 ± 0.001	0.891 ± 0.004	0.976 ± 0.001	0.990 ± 0.000

Table 8: Random 5-fold using $\boldsymbol{Q}$: Mean and standard deviation of the correlation coefficients, obtained as in Table 7, but using five random data splits instead of splits based on model types.

	ARC	HellaSwag	MMLU	TruthfulQA	Winogrande	GSM8K	Average	mean log-likelihood
Pearson's r	0.939 ± 0.002	0.903 ± 0.007	0.923 ± 0.006	0.890 ± 0.013	0.935 ± 0.006	0.863 ± 0.018	0.945 ± 0.005	1.000 ± 0.000
Spearman's ρ	0.943 ± 0.005	0.952 ± 0.003	0.925 ± 0.005	0.875 ± 0.003	0.944 ± 0.008	0.841 ± 0.013	0.954 ± 0.004	1.000 ± 0.000

Table 9: Group 5-fold based on model types using $\boldsymbol{L}$: Mean and standard deviation of the correlation coefficients, obtained as in Table 7, but using the matrix $\boldsymbol{L}$ instead of the matrix $\boldsymbol{Q}$ in the ridge regression of (5) in Section 5.

using five random data splits. These results demonstrate higher correlation coefficients across all tasks compared to those obtained using splits based on model types (Table 7), suggesting that randomly splitting models may unintentionally simplify the prediction task due to leakage from model types.

As shown in (5) in Section 5, we performed ridge regression using the double-centered log-likelihood matrix $\boldsymbol{Q}$, derived from the log-likelihood matrix $\boldsymbol{L}$. To assess the effect of replacing $\boldsymbol{Q}$ with $\boldsymbol{L}$, we repeated the analysis using $\boldsymbol{L}$ and report the resulting means and standard deviations of the correlation coefficients in Table 9. The results obtained with $\boldsymbol{L}$ (Table 9) show nearly the same trend as those with $\boldsymbol{Q}$ (Table 7). Note, however, that the mean of each row of $\boldsymbol{L}$ equals the mean log-likelihood itself, so this specific target can be reproduced exactly. Consequently, the mean correlation coefficients are 1.000 for both Pearson's r and Spearman's ρ.

L Model List

Table 10 lists the 1,018 models used in this study, sorted alphabetically by their names. The BibTeX entries cited for each model were determined through the following procedure.

First, we extracted the BibTeX entries available in each model's Hugging Face model card[22]. If the BibTeX entry was missing a year of publication, we filled it in with the model's creation date[23]. Additionally, we generated BibTeX entries using the arXiv IDs found in the model card tags by querying the arXiv API[24]. This process resulted in a set of BibTeX entries for each model.

Next, we manually checked pairs of different BibTeX entries where the title similarity[25] was high, or the authors matched, to determine whether they corresponded to the same source. This step allowed us to create groups of BibTeX entries that were considered identical.

Then, for each BibTeX group, we selected a representative entry as follows. Within each group, the entry most frequently cited by the models was chosen as the representative. If multiple candidates met this criterion, we prioritized BibTeX entries generated from arXiv IDs when available. If no such entry existed, we selected the one with the longest string.

Finally, we replaced each model's BibTeX entry with the representative entry from its corresponding group. Any selected BibTeX entry that contained typos or formatting errors was manually corrected based on compilation errors. If the author information was incomplete, we corrected it manually by checking the source.

Note that for `google/codegemma-2b` and `deepseek-ai/deepseek-llm-7b-base` in Table 1, as well as for `deepseek-ai/deepseek-coder-1.3b-base` and `mistralai/Mistral-7B-v0.3` in Table 6, we

[22] https://github.com/huggingface/huggingface_hub.

[23] https://github.com/sciunto-org/python-bibtexparser.

[24] https://github.com/lukasschwab/arxiv.py.

[25] We used Python's `difflib.SequenceMatcher`.

manually prepared the BibTeX entries for citation based on their respective sources. We also used the same BibTeX entries for all other models that were considered to be of the same type.

Table 10: List of 1018 models. "ID" denotes the alphabetical index; "Model Name" denotes the name of the model; "Model Type" denotes the classification defined in this paper; "Size" denotes the size of the model (B: billion); "Date" denotes the date of model creation; "DLs" denotes the total number of downloads; $\bar{\ell}_i$ denotes the mean log-likelihood; "Task" denotes the mean of the 6 benchmark scores (i.e., 6-TaskMean).

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
1	aaditya/Llama3-OpenBioLLM-8B (Singhal et al., 2022, 2023; Nori et al., 2023; Rafailov et al., 2024; Pal and Sankarasubbu, 2024b,a)	llama-3	8B	2024-04-20	9308	-590.14	54.06
2	aari1995/germeo-7b-laser	mistral	7B	2024-01-09	2351	-686.18	62.82
3	abacusai/bigstral-12b-32k	mistral	12B	2024-03-06	5216	-632.41	62.17
4	abacusai/Giraffe-13b-32k-v3	llama-2	13B	2023-12-06	1214	-546.12	57.24
5	abacusai/Llama-3-Smaug-8B (Pal et al., 2024)	llama-3	8B	2024-04-19	12873	-565.82	64.61
6	abacusai/Slerp-CM-mist-dpo	mistral	7B	2024-01-03	4030	-553.48	73.10
7	abhinand/Llama-3-OpenBioMed-8B-slerp-v0.3	llama-3	8B	2024-05-02	2713	-558.51	62.08
8	abhinand/tamil-llama-7b-base-v0.1 (Balachandran, 2023)	llama-2	7B	2023-11-08	1382	-877.00	44.52
9	abhinand/tamil-llama-7b-instruct-v0.1 (Balachandran, 2023)	llama-2	7B	2023-11-08	3507	-708.97	45.52
10	abhishekchohan/mistral-7B-forest-dpo	mistral	7B	2024-01-21	1961	-561.51	63.28
11	abhishekchohan/Yi-9B-Forest-DPO-v1.0	llama	9B	2024-03-18	2755	-528.97	64.11
12	acrastt/Bean-3B	llama	3B	2023-09-02	1191	-592.51	40.18
13	acrastt/Griffin-3B	llama	3B	2023-08-18	1198	-584.63	41.13
14	acrastt/Marx-3B	llama	3B	2023-08-15	2006	-603.32	41.71
15	acrastt/Marx-3B-V2	llama	3B	2023-08-22	1234	-609.79	42.08
16	acrastt/OmegLLaMA-3B	llama	3B	2023-08-26	1201	-640.98	38.28
17	acrastt/Puma-3B	llama	3B	2023-08-16	1198	-584.54	41.02
18	acrastt/RedPajama-INCITE-Chat-Instruct-3B-V1	gpt_neox	2B	2023-07-27	1202	-565.14	39.23
19	adamo1139/Mistral-7B-AEZAKMI-v1	mistral	7B	2023-11-27	1199	-596.56	54.92
20	adonlee/LLaMA_2_13B_SFT_v0	llama	13B	2023-10-03	1268	-556.59	57.31
21	adonlee/LLaMA_2_13B_SFT_v1	llama	13B	2023-11-06	1265	-546.75	63.04
22	aerdincdal/CBDDO-LLM-8B-Instruct-v1	llama	8B	2024-05-02	4114	-613.44	56.94
23	ahnyeonchan/OpenOrca-AYT-13B	llama-2	13B	2023-09-07	1184	-569.10	–
24	AIChenKai/TinyLlama-1.1B-Chat-v1.0-x2-MoE	mixtral	1B	2024-01-03	1217	-619.71	36.98
25	AIJUUD/juud-Mistral-7B	mistral	7B	2024-01-31	1312	-564.68	61.72
26	AIJUUD/juud-Mistral-7B-dpo (Lacoste et al., 2019)	mistral	7B	2024-02-07	3217	-567.00	60.89
27	ajibawa-2023/carl-7b	llama	7B	2023-07-22	1212	-599.09	46.16
28	ajibawa-2023/Code-13B	llama	13B	2023-12-08	1181	-604.28	54.81
29	ajibawa-2023/Python-Code-13B	llama	13B	2023-11-11	1201	-582.71	53.61
30	ajibawa-2023/SlimOrca-13B	llama	13B	2023-11-27	1177	-608.08	60.39
31	ajibawa-2023/Uncensored-Frank-13B	llama	13B	2023-09-14	1192	-577.46	55.64
32	ajibawa-2023/Uncensored-Frank-7B	llama	7B	2023-09-14	1178	-654.45	47.90
33	ajibawa-2023/Uncensored-Jordan-13B	llama	13B	2023-10-23	1174	-578.48	56.27
34	ajibawa-2023/Uncensored-Jordan-7B	llama	7B	2023-10-23	1174	-636.79	49.95
35	akjindal53244/Mistral-7B-v0.1-Open-Platypus	mistral	7B	2023-10-05	1286	-543.69	58.92
36	AlekseyKorshuk/pygmalion-6b-vicuna-chatml	gptj	6B	2023-06-22	1169	-564.90	42.08
37	alignment-handbook/zephyr-7b-sft-full	mistral	7B	2023-11-09	11605	-569.88	57.52
38	allbyai/ToRoLaMa-7b-v1.0 (Do et al., 2023)	llama-2	7B	2023-12-19	1185	-732.57	47.87
39	allenai/OLMo-1B-hf (Touvron et al., 2023a; Groeneveld et al., 2024)	olmo	1B	2024-04-12	15221	-619.51	36.78
40	allenai/OLMo-7B-hf (Touvron et al., 2023a; Groeneveld et al., 2024)	olmo	6B	2024-04-12	5407	-554.02	43.36
41	allknowingroger/MultiverseEx26-7B-slerp	mistral	7B	2024-03-30	3034	-599.87	76.80
42	alnrg2arg/blockchainlabs_7B_merged_test2_4	mistral	7B	2024-01-17	1455	-577.39	75.28
43	alnrg2arg/blockchainlabs_7B_merged_test2_4_prune	mistral	7B	2024-01-18	1939	-673.45	57.91
44	aloobun/bun_mistral_7b_v2	mistral	7B	2023-12-20	1305	-558.93	59.76
45	aloobun/falcon-1b-cot-t2	falcon	1B	2024-01-07	2161	-735.39	28.56
46	aloobun/open-llama-3b-v2-elmv3	llama	3B	2023-12-08	1207	-604.18	41.14
47	andreaskoepf/llama2-13b-megacode2_min100	llama-2	13B	2023-08-14	1162	-571.36	56.92
48	Andron00e/YetAnother_Open-Llama-3B-LoRA	llama	3B	2023-07-21	1163	-646.46	–
49	argilla/notus-7b-v1	mistral	7B	2023-11-16	7660	-564.14	60.22
50	Artples/L-MChat-Small	phi	2B	2024-04-11	2793	-640.49	63.14
51	Artples/L-MChat-7b	mistral	7B	2024-04-02	9634	-557.41	69.57
52	Aspik101/StableBeluga-13B-instruct-PL-lora_unload	llama-2	13B	2023-08-04	1244	-544.64	56.24
53	Aspik101/vicuna-13b-v1.5-PL-lora_unload	llama-2	13B	2023-08-03	1263	-566.80	55.24
54	Aspik101/WizardVicuna-Uncensored-3B-instruct-PL-lora_unload	llama-2	3B	2023-08-07	1161	-626.59	39.95
55	AtAndDev/CapybaraMarcoroni-7B	mistral	7B	2024-01-03	1162	-541.33	70.32
56	athirdpath/NSFW_DPO_Noromaid-7b	mistral	7B	2023-12-12	1280	-546.50	61.59
57	augmxnt/shisa-base-7b-v1	mistral	7B	2023-11-19	1188	-648.62	51.64
58	augmxnt/shisa-gamma-7b-v1	mistral	7B	2023-12-23	151426	-590.25	55.50
59	augmxnt/shisa-7b-v1 (Jain et al., 2023; Rafailov et al., 2024)	mistral	7B	2023-11-27	1195	-643.26	55.01
60	Austism/chronos-hermes-13b-v2	llama-2	13B	2023-08-02	1225	-585.56	56.10
61	automerger/YamshadowExperiment28-7B	mistral	7B	2024-03-18	3152	-595.65	76.86
62	Azazelle/Argetsu	mistral	7B	2023-12-30	1208	-560.37	69.64
63	Azazelle/Dumb-Maidlet	mistral	7B	2023-12-30	1201	-548.86	68.34
64	Azazelle/Half-NSFW_Noromaid-7b	mistral	7B	2023-12-29	1208	-547.63	62.32
65	Azazelle/Maylin-7b	mistral	7B	2024-01-04	1198	-575.94	70.26
66	Azazelle/Silicon-Medley	mistral	7B	2023-12-29	1200	-566.66	69.49
67	Azazelle/SlimMelodicMaid	mistral	7B	2023-12-30	1205	-577.28	69.70
68	Azazelle/smol_bruin-7b	mistral	7B	2023-12-29	1207	-572.80	71.05
69	Azazelle/Tippy-Toppy-7b	mistral	7B	2024-01-03	1197	-560.94	69.58
70	Azazelle/xDAN-SlimOrca	mistral	7B	2023-12-29	1206	-575.20	68.04
71	Azazelle/Yuna-7b-Merge	mistral	7B	2024-01-05	1201	-580.65	71.46
72	Azure99/blossom-v1-3b	bloom	3B	2023-07-29	1228	-641.52	36.90
73	Azure99/blossom-v2-llama2-7b	llama-2	7B	2023-09-06	1241	-572.17	51.71
74	Azure99/blossom-v2-3b	bloom	3B	2023-08-08	1237	-654.44	35.98
75	Azure99/blossom-v3-mistral-7b	mistral	7B	2023-11-20	1319	-565.31	62.95
76	Azure99/blossom-v3_1-mistral-7b	mistral	7B	2023-11-27	1320	-567.17	62.53
77	Azure99/blossom-v4-mistral-7b	mistral	7B	2023-12-26	1315	-550.37	63.61
78	BarryFutureman/NeuralTurdusVariant1-7B	mistral	7B	2024-01-22	1170	-592.94	74.83

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
79	beaugogh/Llama2-7b-openorca-mc-v1	llama-2	7B	2023-08-20	1177	-613.56	52.24
80	beaugogh/Llama2-7b-openorca-mc-v2-dpo	llama-2	7B	2023-10-06	1173	-608.74	52.32
81	beaugogh/Llama2-7b-sharegpt4	llama-2	7B	2023-08-12	1183	-660.22	51.05
82	beaugogh/pythia-1.4b-deduped-sharegpt	gpt_neox	1B	2023-07-25	1293	-557.93	35.11
83	BEE-spoke-data/Mixtral-GQA-400m-v2	mixtral	2B	2023-12-20	1182	-794.25	28.45
84	beomi/KoAlpaca-KoRWKV-6B	rwkv	6B	2023-06-02	2288	-1099.15	28.57
85	beomi/KoAlpaca-Polyglot-5.8B	gpt_neox	6B	2023-03-16	4198	-1309.35	29.46
86	beomi/KoRWKV-6B	rwkv	6B	2023-05-26	2127	-1148.78	28.19
87	beomi/llama-2-ko-7b (Lee, 2023)	llama-2	6B	2023-07-20	5226	-907.04	45.32
88	beomi/Yi-Ko-6B (Lee, 2024b)	llama	6B	2023-11-30	4261	-589.43	50.27
89	beowolx/CodeNinja-1.0-OpenChat-7B	mistral	7B	2023-12-20	6300	-555.93	67.40
90	berkeley-nest/Starling-LM-7B-alpha (Zhu et al., 2023a,b)	mistral	7B	2023-11-25	18366	-571.63	67.05
91	bhavinjawade/SOLAR-10B-OrcaDPO-Jawade	llama	10B	2024-01-06	1224	-563.19	74.27
92	bigcode/gpt_bigcode-santacoder	gpt_bigcode	1B	2023-04-06	40198	-861.42	28.49
93	bigcode/starcoderbase-1b (Shazeer, 2019; Dao et al., 2022; Bavarian et al., 2022; Li et al., 2023a)	gpt_bigcode	1B	2023-07-03	5247	-734.99	30.06
94	bigcode/starcoderbase-7b (Shazeer, 2019; Dao et al., 2022; Bavarian et al., 2022; Li et al., 2023a)	gpt_bigcode	7B	2023-07-26	3566	-652.23	33.75
95	bigcode/starcoder2-3b (Beltagy et al., 2020; Dao et al., 2022; Bavarian et al., 2022; Ainslie et al., 2023; Lozhkov et al., 2024)	starcoder2	3B	2023-11-29	445316	-620.70	39.25
96	bigcode/starcoder2-7b (Beltagy et al., 2020; Dao et al., 2022; Bavarian et al., 2022; Ainslie et al., 2023; Lozhkov et al., 2024)	starcoder2	7B	2024-02-20	11834	-584.03	42.95
97	bigscience/bloomz-3b (Muennighoff et al., 2023)	bloom	3B	2022-10-08	7975	-701.42	37.03
98	bigscience/bloomz-7b1 (Muennighoff et al., 2023)	bloom	7B	2022-09-27	12152	-652.34	42.21
99	bigscience/bloomz-7b1-mt (Muennighoff et al., 2023)	bloom	7B	2022-09-28	2427	-653.08	42.14
100	bigscience/bloom-1b1 (Shoeybi et al., 2020; Dettmers et al., 2022; Press et al., 2022)	bloom	1B	2022-05-19	11054	-687.97	32.47
101	bigscience/bloom-1b7 (Shoeybi et al., 2020; Dettmers et al., 2022; Press et al., 2022)	bloom	1B	2022-05-19	40840	-657.05	33.98
102	bigscience/bloom-3b (Shoeybi et al., 2020; Dettmers et al., 2022; Press et al., 2022)	bloom	3B	2022-05-19	13914	-633.79	36.07
103	bigscience/bloom-7b1 (Shoeybi et al., 2020; Dettmers et al., 2022; Press et al., 2022)	bloom	7B	2022-05-19	21428	-604.20	39.18
104	BioMistral/BioMistral-7B (Labrak et al., 2024)	mistral	7B	2024-02-14	11349	-652.04	52.33
105	BioMistral/BioMistral-7B-DARE (Yadav et al., 2023a; Labrak et al., 2024; Yu et al., 2024a)	mistral	7B	2024-02-05	1209	-586.37	57.03
106	BlueNipples/TimeCrystal-l2-13B	llama-2	13B	2023-11-11	1282	-567.72	59.26
107	bofenghuang/vigogne-2-13b-instruct	llama-2	13B	2023-07-26	1203	-534.88	55.14
108	bofenghuang/vigogne-2-7b-chat	llama-2	7B	2023-07-29	1170	-558.60	52.45
109	bofenghuang/vigogne-2-7b-instruct	llama-2	7B	2023-07-20	1243	-555.11	52.02
110	bofenghuang/vigogne-7b-instruct	llama-1	7B	2023-03-22	1167	-570.61	47.76
111	bofenghuang/vigostral-7b-chat	mistral	7B	2023-09-29	4118	-535.15	59.18
112	BramVanroy/GEITje-7B-ultra (Vanroy, 2024)	mistral	7B	2024-01-27	1597	-674.95	52.61
113	Brouz/Slerpeno	llama	13B	2023-09-08	1201	-545.39	56.59
114	budecosystem/boomer-1b	llama	1B	2023-10-03	1267	-846.13	28.44
115	CalderaAI/13B-BlueMethod	llama-1	13B	2023-07-07	1193	-569.51	54.12
116	CalderaAI/13B-Legerdemain-L2	llama-2	13B	2023-08-03	1200	-551.69	55.13
117	CalderaAI/13B-Ouroboros	llama	13B	2023-07-20	1197	-1133.56	49.54
118	CalderaAI/13B-Thorns-l2	llama	13B	2023-09-06	1207	-580.08	54.72
119	Cartinoe5930/Llama2_init_Mistral	llama-2	7B	2024-01-16	1174	-528.82	60.98
120	castorini/rank_vicuna_7b_v1_fp16 (Touvron et al., 2023b; Pradeep et al., 2023)	llama-2	7B	2023-09-27	1165	-720.69	44.36
121	ceadar-ie/FinanceConnect-13B (CeADAR, 2023)	llama	13B	2023-11-28	1270	-661.62	49.34
122	Changgil/K2S3-SOLAR-11b-v1.0	llama	10B	2024-03-03	2290	-811.23	36.67
123	Changgil/k2s3_test_24001	llama-2	13B	2024-02-14	2346	-561.61	56.67
124	chargoddard/storytime-13b	llama-2	13B	2023-09-22	1166	-599.45	56.64
125	chickencaesar/llama2-platypus-llama2-chat-13B-hf	llama-2	13B	2023-09-28	1216	-536.92	54.11
126	CHIH-HUNG/llama-2-13b-dolphin_20w	llama-2	13B	2023-08-29	1187	-533.73	55.06
127	CHIH-HUNG/llama-2-13b-dolphin_5w	llama-2	13B	2023-08-25	1175	-533.03	55.53
128	CHIH-HUNG/llama-2-13b-FINETUNE2_TEST_2.2w	llama-2	13B	2023-09-04	1174	-530.90	53.20
129	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r16-gate_up_down	llama-2	13B	2023-09-20	1171	-533.48	54.32
130	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r4-gate_up_down	llama-2	13B	2023-09-19	1166	-535.81	53.35
131	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r4-q_k_v_o	llama-2	13B	2023-09-19	1170	-536.50	53.62
132	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r8-gate_up_down	llama-2	13B	2023-09-20	1165	-533.51	53.71
133	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r8-q_k_v_o	llama-2	13B	2023-09-20	1170	-535.00	53.99
134	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r8-q_k_v_o_gate_up_down	llama-2	13B	2023-09-24	1165	-536.24	53.43
135	CHIH-HUNG/llama-2-13b-FINETUNE4_addto15k_4.5w-r16-gate_up_down	llama-2	13B	2023-10-08	1161	-531.46	54.88
136	CHIH-HUNG/llama-2-13b-FINETUNE4_compare15k_4.5w-r16-gate_up_down	llama-2	13B	2023-10-08	1169	-532.00	53.94
137	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r16-gate_up_down	llama-2	13B	2023-09-22	1165	-533.56	53.52
138	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r16-gate_up_down-test1	llama-2	13B	2023-10-07	1165	-531.27	53.66
139	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r16-q_k_v_o	llama-2	13B	2023-09-22	1164	-534.36	53.68
140	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r4-gate_up_down	llama-2	13B	2023-09-21	1174	-537.05	53.48
141	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r4-q_k_v_o_gate_up_down	llama-2	13B	2023-09-25	1163	-540.47	53.23
142	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r8-gate_up_down	llama-2	13B	2023-09-21	1171	-533.65	53.58
143	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r8-q_k_v_o	llama-2	13B	2023-09-21	1171	-533.89	54.06
144	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r8-q_k_v_o_gate_up_down	llama-2	13B	2023-09-25	1165	-537.12	52.88
145	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r16-gate_up_down	llama-2	13B	2023-10-04	1167	-534.88	53.44
146	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r16-q_k_v_o	llama-2	13B	2023-10-04	1161	-533.84	54.63
147	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r4-q_k_v_o	llama-2	13B	2023-10-01	1172	-536.02	53.32
148	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r4-q_k_v_o_gate_up_down	llama-2	13B	2023-10-02	1163	-541.64	53.38
149	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r8-gate_up_down	llama-2	13B	2023-10-03	1163	-534.71	54.02
150	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r8-q_k_v_o	llama-2	13B	2023-10-02	1167	-534.88	54.64
151	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r8-q_k_v_o_gate_up_down	llama-2	13B	2023-10-04	1164	-540.68	53.69
152	CHIH-HUNG/llama-2-13b-OpenOrca_5w	llama-2	13B	2023-08-24	1178	-533.15	55.80
153	CHIH-HUNG/llama-2-13b-Open_Platypus_and_ccp_2.6w-3_epoch	llama-2	13B	2023-09-05	1165	-533.19	53.80

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
154	chinoll/Yi-6b-200k-dpo	llama	6B	2023-12-01	1187	-573.79	51.93
155	circulus/Llama-2-13b-orca-v1	llama-2	13B	2023-08-01	1228	-540.82	57.05
156	circulus/Llama-2-7b-orca-v1	llama-2	7B	2023-08-01	1235	-565.19	53.56
157	clibrain/Llama-2-7b-ft-instruct-es	llama-2	7B	2023-08-09	2062	-559.81	49.63
158	CobraMamba/mamba-gpt-3b-v4 (chiliu, 2023)	llama	3B	2023-09-05	1183	-605.83	41.24
159	codellama/CodeLlama-13b-hf (Rozière et al., 2024)	llama-2	13B	2023-08-24	8432	-582.07	43.35
160	codellama/CodeLlama-13b-Instruct-hf (Rozière et al., 2024)	llama-2	13B	2023-08-24	25248	-579.86	45.82
161	codellama/CodeLlama-13b-Python-hf (Rozière et al., 2024)	llama-2	13B	2023-08-24	2537	-587.68	37.00
162	codellama/CodeLlama-7b-hf (Rozière et al., 2024)	llama-2	6B	2023-08-24	66040	-597.05	39.81
163	codellama/CodeLlama-7b-Instruct-hf (Rozière et al., 2024)	llama-2	6B	2023-08-24	133523	-598.06	40.05
164	codellama/CodeLlama-7b-Python-hf (Rozière et al., 2024)	llama-2	6B	2023-08-24	5138	-605.60	36.42
165	cognitivecomputations/dolphin-2.2.1-mistral-7b	mistral	7B	2023-10-30	7292	-546.22	65.01
166	cognitivecomputations/dolphin-2.6-mistral-7b	mistral	7B	2023-12-27	1418	-559.17	64.92
167	cognitivecomputations/dolphin-2.6-mistral-7b-dpo	mistral	7B	2023-12-31	1236	-565.34	67.20
168	cognitivecomputations/dolphin-2.6-mistral-7b-dpo-laser (Sharma et al., 2023)	mistral	7B	2024-01-01	2070	-558.72	67.28
169	cognitivecomputations/dolphin-2.9-llama3-8b	llama-3	8B	2024-04-20	130977	-640.39	65.92
170	cognitivecomputations/dolphin-2.9.1-llama-3-8b	llama-3	8B	2024-05-10	7575	-647.01	66.23
171	cognitivecomputations/dolphin-2.9.1-yi-1.5-9b	llama	8B	2024-05-18	4880	-598.79	68.92
172	cognitivecomputations/Llama-3-8B-Instruct-abliterated-v2	llama-3	8B	2024-05-09	8728	-581.93	66.00
173	cognitivecomputations/openchat-3.5-0106-laser	mistral	7B	2024-01-27	6155	-554.46	69.46
174	cognitivecomputations/TinyDolphin-2.8-1.1b	llama	1B	2024-01-21	1678	-683.06	36.34
175	cognitivecomputations/WestLake-7B-v2-laser	mistral	7B	2024-01-26	3980	-591.19	74.78
176	ContextualAI/archangel_sft-kto_llama13b (Ethayarajh et al., 2023)	llama	13B	2023-12-03	1214	-542.49	52.87
177	cookinai/BruinHermes	mistral	7B	2023-12-17	1193	-551.19	73.42
178	cookinai/CatMacaroni14	mistral	7B	2023-12-31	1186	-552.22	72.68
179	Corianas/gpt-j-6B-Dolly	gptj	6B	2023-03-28	1173	-502.52	39.60
180	crumb/apricot-wildflower-20	mistral	7B	2023-12-19	1179	-531.99	59.74
181	cyberagent/open-calm-7b (Andonian et al., 2021)	gpt_neox	7B	2023-05-15	24168	-1002.50	28.21
182	cypienai/cymist2-v01-SFT (Lacoste et al., 2019)	mistral	7B	2024-05-12	2754	-628.53	51.71
183	cypienai/cymist-2-v02-SFT (Lacoste et al., 2019)	mistral	7B	2024-05-22	2717	-537.73	62.57
184	Dampish/Dante-2.8B	gpt_neox	2B	2023-05-09	1237	-671.96	–
185	Dampish/StellarX-4B-V0 (Black et al., 2022)	gpt_neox	4B	2023-05-27	1268	-722.64	37.31
186	Dampish/StellarX-4B-V0.2 (Black et al., 2022)	gpt_neox	4B	2023-06-03	1245	-866.96	36.15
187	Danielbrdz/Barcenas-Llama3-8b-ORPO	llama-3	8B	2024-04-29	12814	-557.50	72.50
188	Danielbrdz/Barcenas-13b	llama-2	13B	2023-09-09	1178	-558.81	55.83
189	Danielbrdz/Barcenas-3b	llama-2	3B	2023-11-15	1167	-554.05	41.74
190	Danielbrdz/Barcenas-7b	llama	7B	2023-08-25	1175	-603.96	50.87
191	Danielbrdz/CodeBarcenas-7b	llama-2	7B	2023-09-03	1174	-656.68	40.09
192	danielpark/gorani-100k-llama2-13b-instruct	llama-2	13B	2023-10-04	1176	-1410.79	29.69
193	databricks/dolly-v2-12b (Conover et al., 2023)	gpt_neox	12B	2023-04-11	3860	-566.95	39.46
194	databricks/dolly-v2-3b (Conover et al., 2023)	gpt_neox	3B	2023-04-13	21443	-556.97	–
195	databricks/dolly-v2-7b (Conover et al., 2023)	gpt_neox	7B	2023-04-13	10069	-555.60	39.24
196	DBCMLAB/Llama-3-instruction-constructionsafety-layertuning (AI@Meta, 2024; Lee, 2024a; Lee and Ahn, 2024)	llama-3	8B	2024-05-22	2712	-658.58	56.32
197	Deathsquad10/TinyLlama-Remix	llama	1B	2023-11-26	1213	-812.12	34.00
198	Deathsquad10/TinyLlama-repeat	llama	1B	2024-01-06	1191	-622.61	37.09
199	Deathsquad10/TinyLlama-1.1B-Remix-V.2	llama	1B	2024-01-05	1192	-670.53	34.91
200	Deci/DeciLM-7B-instruct (DeciAI Research Team, 2023)	deci	7B	2023-12-10	9769	-570.53	63.19
201	DeepMount00/Llama-3-8b-Ita	llama-3	8B	2024-05-01	179401	-578.66	73.65
202	DeepMount00/Mistral-Ita-7b	mistral	7B	2023-11-08	2927	-579.30	58.26
203	deepseek-ai/DeepSeek-Coder-V2-Lite-Base (Dai et al., 2024)	deepseek	16B	2024-06-14	13642	-527.73	–
204	deepseek-ai/DeepSeek-Coder-V2-Lite-Instruct (Dai et al., 2024)	deepseek	16B	2024-06-14	142758	-569.70	–
205	deepseek-ai/deepseek-coder-1.3b-base (Guo et al., 2024)	deepseek	1B	2023-10-28	48960	-718.82	–
206	deepseek-ai/deepseek-coder-1.3b-instruct (Guo et al., 2024)	deepseek	1B	2023-10-29	28984	-785.33	32.40
207	deepseek-ai/deepseek-coder-6.7b-base (Guo et al., 2024)	deepseek	6B	2023-10-23	38348	-647.51	40.87
208	deepseek-ai/deepseek-coder-6.7b-instruct (Guo et al., 2024)	deepseek	6B	2023-10-29	28828	-686.56	43.57
209	deepseek-ai/deepseek-coder-7b-base-v1.5 (Guo et al., 2024)	deepseek	7B	2024-01-25	895	-570.98	–
210	deepseek-ai/deepseek-coder-7b-instruct-v1.5 (Guo et al., 2024)	deepseek	6B	2024-01-25	28928	-604.47	50.89
211	deepseek-ai/deepseek-llm-7b-base (DeepSeek-AI et al., 2024a)	deepseek	7B	2023-11-29	19180	-552.39	–
212	deepseek-ai/deepseek-llm-7b-chat (DeepSeek-AI et al., 2024a)	deepseek	7B	2023-11-29	34271	-594.23	59.38
213	deepseek-ai/deepseek-math-7b-base (Shao et al., 2024)	deepseek	7B	2024-02-05	37614	-566.93	57.61
214	deepseek-ai/deepseek-math-7b-instruct (Shao et al., 2024)	deepseek	7B	2024-02-05	8352	-593.77	51.48
215	deepseek-ai/deepseek-math-7b-rl (Shao et al., 2024)	deepseek	6B	2024-02-05	1633	-607.25	49.54
216	deepseek-ai/deepseek-moe-16b-base (Dai et al., 2024)	deepseek	16B	2024-01-08	13924	-544.49	–
217	deepseek-ai/deepseek-moe-16b-chat (Dai et al., 2024)	deepseek	16B	2024-01-09	8852	-578.23	–
218	deepseek-ai/DeepSeek-Prover-V1 (Xin et al., 2024a)	deepseek	7B	2024-08-16	114	-791.21	–
219	deepseek-ai/DeepSeek-Prover-V1.5-Base (Xin et al., 2024b)	deepseek	7B	2024-08-15	230	-554.13	–
220	deepseek-ai/DeepSeek-Prover-V1.5-RL (Xin et al., 2024b)	deepseek	7B	2024-08-15	12223	-672.91	–
221	deepseek-ai/DeepSeek-Prover-V1.5-SFT (Xin et al., 2024b)	deepseek	7B	2024-08-15	6459	-670.56	–
222	deepseek-ai/DeepSeek-R1-Distill-Llama-8B (DeepSeek-AI et al., 2025)	deepseek	8B	2025-01-20	199313	-693.69	–
223	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI et al., 2025)	deepseek	14B	2025-01-20	139489	-605.82	–
224	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B (DeepSeek-AI et al., 2025)	deepseek	2B	2025-01-20	290094	-854.97	–
225	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B (DeepSeek-AI et al., 2025)	deepseek	7B	2025-01-20	201118	-758.13	–
226	deepseek-ai/DeepSeek-V2-Lite (DeepSeek-AI et al., 2024b)	deepseek	16B	2024-05-15	25474	-533.54	–
227	deepseek-ai/DeepSeek-V2-Lite-Chat (DeepSeek-AI et al., 2024b)	deepseek	16B	2024-05-15	18056	-588.92	–
228	deepseek-ai/ESFT-vanilla-lite (Wang et al., 2024c)	deepseek	16B	2024-07-04	270	-550.46	–
229	Delcos/Mistral-Pygmalion-7b	llama-2	7B	2023-10-09	1185	-556.74	51.02
230	dfurman/Llama-3-8B-Orpo-v0.1	llama-3	8B	2024-04-26	5146	-515.13	64.67
231	dhmeltzer/Llama-2-13b-hf-ds_eli5_1024_r_64_alpha_16_merged	llama-2	13B	2023-09-14	1188	-542.35	54.16
232	dhmeltzer/Llama-2-13b-hf-ds_wiki_1024_full_r_64_alpha_16_merged	llama-2	13B	2023-09-14	1198	-536.93	52.94
233	dhmeltzer/Llama-2-13b-hf-eli5-wiki-1024_r_64_alpha_16_merged	llama-2	13B	2023-09-14	1192	-538.52	53.57
234	dhmeltzer/llama-7b-SFT_eli5_wiki65k_1024_r_64_alpha_16_merged	llama	6B	2023-08-25	1257	-568.38	50.00
235	diffnamehard/Psyfighter2-Noromaid-ties-13B	llama	13B	2023-12-28	1215	-569.13	59.47
236	digitous/Alpacino13b	llama-1	13B	2023-04-13	1182	-547.35	52.39
237	digitous/13B-Chimera	llama-1	13B	2023-05-23	1359	-562.36	54.92
238	Doctor-Shotgun/CalliopeDS-L2-13B (Yadav et al., 2023a)	llama-2	13B	2023-09-16	1428	-568.72	56.34

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
239	Doctor-Shotgun/CalliopeDS-v2-L2-13B	llama-2	13B	2023-09-28	1250	-572.65	57.12
240	DopeorNope/You_can_cry_Snowman-13B	llama	13B	2023-12-27	1203	-577.68	69.46
241	dotvignesh/perry-7b	llama	7B	2023-09-28	1171	-618.90	49.55
242	dreamgen/WizardLM-2-7B (Xu et al., 2023a; Luo et al., 2023, 2025)	mistral	7B	2024-04-16	2533	-641.93	63.75
243	dvruette/oasst-pythia-12b-flash-attn-5000-steps	gpt_neox	12B	2023-03-12	1175	-602.37	40.73
244	dvruette/oasst-pythia-12b-pretrained-sft	gpt_neox	12B	2023-04-03	1170	-552.58	41.48
245	dvruette/oasst-pythia-12b-reference	gpt_neox	12B	2023-04-03	1174	-557.88	40.33
246	eachadea/vicuna-13b-1.1	llama-1	13B	2023-04-13	1336	-646.61	53.29
247	eachadea/vicuna-7b-1.1	llama-1	7B	2023-04-13	1388	-613.88	50.37
248	eldogbbhed/Peagle-9b	mistral	8B	2024-03-10	10180	-581.92	73.30
249	EleutherAI/gpt-j-6b (Gao et al., 2020; Wang, 2021; Wang and Komatsuzaki, 2021; Su et al., 2023a)	gptj	6B	2022-03-02	241435	-479.14	40.10
250	EleutherAI/gpt-neo-1.3B (Gao et al., 2020; Black et al., 2021)	gpt_neo	1B	2022-03-02	243332	-545.00	33.58
251	EleutherAI/gpt-neo-2.7B (Gao et al., 2020; Black et al., 2021)	gpt_neo	2B	2022-03-02	205503	-520.30	36.20
252	EleutherAI/llemma_7b (Azerbayev et al., 2024)	llama-2	7B	2023-09-12	4671	-554.43	48.75
253	EleutherAI/polyglot-ko-12.8b (Kim et al., 2022; Su et al., 2023a; Ko et al., 2023)	gpt_neox	13B	2022-10-14	2941	-967.65	33.33
254	EleutherAI/pythia-1b-deduped (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	1B	2023-02-14	17983	-551.07	32.78
255	EleutherAI/pythia-12b (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	12B	2023-02-28	9192	-475.42	38.82
256	EleutherAI/pythia-12b-deduped (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	12B	2023-02-27	9737	-477.81	39.70
257	EleutherAI/pythia-1.4b (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	1B	2023-02-09	22152	-531.68	34.75
258	EleutherAI/pythia-1.4b-deduped (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	1B	2023-02-09	12730	-534.52	35.00
259	EleutherAI/pythia-2.8b-deduped (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	2B	2023-02-10	11649	-507.42	36.72
260	EleutherAI/pythia-6.9b-deduped (Gao et al., 2020; Biderman et al., 2022, 2023)	gpt_neox	6B	2023-02-25	9824	-490.70	39.30
261	elinas/chronos-13b-v2	llama-2	13B	2023-08-02	1944	-595.38	55.25
262	elliotthwang/elliott_Llama-2-7b-hf	llama-2	6B	2023-10-09	1181	-553.66	50.20
263	elyza/ELYZA-japanese-Llama-2-13b (Sasaki et al., 2023b; Touvron et al., 2023b)	llama-2	13B	2023-12-25	1200	-587.25	56.14
264	elyza/ELYZA-japanese-Llama-2-13b-instruct (Sasaki et al., 2023b; Touvron et al., 2023b)	llama-2	13B	2023-12-25	1710	-594.98	54.72
265	elyza/ELYZA-japanese-Llama-2-7b (Touvron et al., 2023b; Sasaki et al., 2023a)	llama-2	7B	2023-08-28	2356	-630.66	48.70
266	elyza/ELYZA-japanese-Llama-2-7b-fast (Touvron et al., 2023b; Sasaki et al., 2023a)	llama-2	7B	2023-08-28	1766	-640.71	47.67
267	elyza/ELYZA-japanese-Llama-2-7b-fast-instruct (Touvron et al., 2023b; Sasaki et al., 2023a)	llama-2	7B	2023-08-28	2387	-629.72	49.15
268	elyza/ELYZA-japanese-Llama-2-7b-instruct (Touvron et al., 2023b; Sasaki et al., 2023a)	llama-2	7B	2023-08-28	6296	-625.45	49.78
269	Enno-Ai/ennodata-raw-pankajmathur-13b-peft	llama	13B	2023-09-29	1215	-613.07	55.40
270	ericzzz/falcon-rw-1b-chat	falcon	1B	2023-12-05	1319	-722.18	37.37
271	ericzzz/falcon-rw-1b-instruct-openorca	falcon	1B	2023-11-24	1585	-752.98	37.63
272	euclaise/falcon_1b_stage1	falcon	1B	2023-09-15	2119	-734.28	37.25
273	euclaise/falcon_1b_stage2	falcon	1B	2023-09-17	4016	-717.40	37.59
274	euclaise/Ferret_7B	mistral	7B	2023-10-28	1172	-634.24	53.87
275	Eurdem/Defne_llama3_2x8B	mixtral	13B	2024-05-10	9076	-571.19	71.73
276	Expert68/llama2_13b_instructed_version2	llama-2	13B	2023-10-14	1184	-561.35	55.41
277	facebook/opt-iml-max-1.3b (Iyer et al., 2023)	opt	1B	2023-01-26	8604	-712.07	35.21
278	facebook/opt-13b (Brown et al., 2020; Zhang et al., 2022b)	opt	13B	2022-05-11	19130	-625.38	40.06
279	facebook/opt-1.3b (Brown et al., 2020; Zhang et al., 2022b)	opt	1B	2022-05-11	139758	-700.64	34.60
280	facebook/opt-2.7b (Brown et al., 2020; Zhang et al., 2022b)	opt	2B	2022-05-11	61450	-673.03	36.74
281	facebook/opt-6.7b (Brown et al., 2020; Zhang et al., 2022b)	opt	6B	2022-05-11	44909	-641.13	39.08
282	facebook/xglm-1.7B (Lin et al., 2022b)	xglm	1B	2022-03-02	2388	-723.91	31.42
283	facebook/xglm-4.5B (Lin et al., 2022b)	xglm	5B	2022-03-02	1436	-669.92	34.31
284	facebook/xglm-7.5B (Lin et al., 2022b)	xglm	7B	2022-03-02	4654	-663.87	36.38
285	failspy/Meta-Llama-3-8B-Instruct-abliterated-v3	llama-3	8B	2024-05-20	9975	-572.33	67.27
286	failspy/Phi-3-medium-4k-instruct-abliterated-v3	phi3	13B	2024-05-22	5430	-569.05	70.12
287	FairMind/Llama-3-8B-4bit-UltraChat-Ita	llama-3	8B	2024-05-03	5014	-543.07	61.54
288	FairMind/Phi-3-mini-4k-instruct-bnb-4bit-Ita	mistral	4B	2024-05-02	2746	-634.71	66.61
289	fblgit/LUNA-SOLARkrautLM-Instruct	llama	10B	2023-12-22	1180	-582.98	73.79
290	fblgit/una-cybertron-7b-v2-bf16 (Murias, 2023)	mistral	7B	2023-12-02	1440	-568.27	69.67
291	fblgit/UNA-POLAR-10.7B-InstructMath-v2	llama	10B	2024-01-02	1180	-560.34	74.07
292	feidfoe/Metamath-reproduce-7b	llama-2	7B	2023-11-24	1198	-724.03	55.81
293	FelixChao/CodeLlama13B-Finetune-v1	llama	13B	2023-09-13	1188	-647.40	47.19
294	FelixChao/llama2-13b-math1.1	llama-2	13B	2023-08-12	1181	-602.88	54.18
295	FelixChao/llama2-13b-math1.2	llama-2	13B	2023-08-15	1197	-607.07	54.19
296	FinancialSupport/saiga-7b	mistral	7B	2023-12-28	3902	-558.71	64.51
297	fireballoon/baichuan-vicuna-chinese-7b	llama-1	7B	2023-06-18	1210	-702.08	46.06
298	FlagAlpha/Llama2-Chinese-7b-Chat	llama-2	7B	2023-07-23	1383	-627.19	51.13
299	FlagAlpha/Llama3-Chinese-8B-Instruct	llama-3	8B	2024-04-23	1978	-557.17	63.50
300	Fredithefish/Guanaco-3B-Uncensored-v2	gpt_neox	2B	2023-08-27	2189	-632.46	38.98
301	FreedomIntelligence/AceGPT-7B	llama	7B	2023-09-14	2982	-565.76	49.47
302	FreedomIntelligence/phoenix-inst-chat-7b	bloom	7B	2023-04-11	1296	-693.44	43.03
303	freewheelin/free-llama3-dpo-v0.2 (Kim et al., 2024b)	llama-3	8B	2024-05-09	3896	-499.17	62.69
304	gagan3012/MetaModelv2	llama	10B	2024-01-03	1203	-560.16	74.24
305	gagan3012/MetaModelv3	llama	10B	2024-01-05	1207	-561.14	74.39
306	garage-bAInd/Platypus2-13B (Hu et al., 2022; Touvron et al., 2023b; Lee et al., 2024)	llama	13B	2023-08-05	3716	-533.92	54.89
307	garage-bAInd/Platypus2-7B (Hu et al., 2022; Touvron et al., 2023b; Lee et al., 2024)	llama	6B	2023-08-22	6198	-554.08	49.97
308	GeneZC/MiniChat-1.5-3B (Jain et al., 2023; Rafailov et al., 2024; Zhang et al., 2024a)	llama	3B	2023-11-26	1601	-599.51	50.23
309	GeneZC/MiniChat-2-3B (Jain et al., 2023; Rafailov et al., 2024; Zhang et al., 2024a)	llama	3B	2023-12-27	4167	-607.10	51.49
310	GeneZC/MiniChat-3B (Zhang et al., 2024a)	llama	3B	2023-11-11	1633	-585.38	45.31

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
311	GeneZC/MiniMA-2-3B (Zhang et al., 2024a)	llama	3B	2023-12-27	1492	-526.37	44.75
312	GeneZC/MiniMA-3B (Zhang et al., 2024a)	llama	3B	2023-11-11	1494	-537.57	41.44
313	georgesung/llama2_7b_chat_uncensored	llama-2	7B	2023-07-20	2602	-568.34	49.67
314	ghost-x/ghost-7b-alpha	mistral	7B	2024-04-13	4854	-662.59	57.65
315	glaiveai/glaive-coder-7b	llama-2	7B	2023-09-17	1222	-692.31	41.56
316	google/codegemma-2b (CodeGemma Team et al., 2024)	gemma	2B	2024-03-21	8798	-793.92	32.19
317	google/codegemma-7b (CodeGemma Team et al., 2024)	gemma	8B	2024-03-21	3779	-556.79	56.73
318	google/codegemma-7b-it (CodeGemma Team et al., 2024)	gemma	8B	2024-03-21	11227	-1040.98	58.28
319	google/gemma-1.1-7b-it (Joshi et al., 2017; Zhao et al., 2018; Mihaylov et al., 2018; Zellers et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Bisk et al., 2019; Sap et al., 2019; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Austin et al., 2021; Parrish et al., 2022; Zhong et al., 2023; Srivastava et al., 2023; Gemini Team et al., 2024)	gemma	8B	2024-03-26	19835	-1355.60	60.09
320	google/gemma-2b (Joshi et al., 2017; Mihaylov et al., 2018; Zhao et al., 2018; Rudinger et al., 2018; Zellers et al., 2019; Bisk et al., 2019; Sap et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Gehman et al., 2020; Austin et al., 2021; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Dhamala et al., 2021; Parrish et al., 2022; Lin et al., 2022a; Hartvigsen et al., 2022; Srivastava et al., 2023; Zhong et al., 2023; Gemini Team et al., 2024)	gemma	2B	2024-02-08	256798	-581.35	46.51
321	google/gemma-2b-it (Joshi et al., 2017; Mihaylov et al., 2018; Zhao et al., 2018; Rudinger et al., 2018; Zellers et al., 2019; Bisk et al., 2019; Sap et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Gehman et al., 2020; Austin et al., 2021; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Dhamala et al., 2021; Parrish et al., 2022; Lin et al., 2022a; Hartvigsen et al., 2022; Srivastava et al., 2023; Zhong et al., 2023; Gemini Team et al., 2024)	gemma	2B	2024-02-08	103851	-886.70	42.75
322	google/gemma-7b (Joshi et al., 2017; Mihaylov et al., 2018; Zhao et al., 2018; Rudinger et al., 2018; Zellers et al., 2019; Bisk et al., 2019; Sap et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Gehman et al., 2020; Austin et al., 2021; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Dhamala et al., 2021; Parrish et al., 2022; Lin et al., 2022a; Hartvigsen et al., 2022; Srivastava et al., 2023; Zhong et al., 2023; Dettmers et al., 2023; Gemini Team et al., 2024)	gemma	8B	2024-02-08	72047	-539.12	63.75
323	google/gemma-7b-it (Joshi et al., 2017; Mihaylov et al., 2018; Zhao et al., 2018; Rudinger et al., 2018; Zellers et al., 2019; Bisk et al., 2019; Sap et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Gehman et al., 2020; Austin et al., 2021; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Dhamala et al., 2021; Parrish et al., 2022; Lin et al., 2022a; Hartvigsen et al., 2022; Srivastava et al., 2023; Zhong et al., 2023; Gemini Team et al., 2024)	gemma	8B	2024-02-13	66776	-1327.43	53.56
324	google/recurrentgemma-2b-it (Joshi et al., 2017; Mihaylov et al., 2018; Zhao et al., 2018; Rudinger et al., 2018; Zellers et al., 2019; Bisk et al., 2019; Sap et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Gehman et al., 2020; Hendrycks et al., 2021b; Austin et al., 2021; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Dhamala et al., 2021; Parrish et al., 2022; Lin et al., 2022a; Hartvigsen et al., 2022; Srivastava et al., 2023; Zhong et al., 2023; De et al., 2024; Griffin Team et al., 2024)	recurrent_gemma	2B	2024-04-08	4046	-1033.09	40.86
325	gradientai/Llama-3-8B-Instruct-Gradient-1048k (Ding et al., 2023; Peng et al., 2023b; Pekelis et al., 2024; AI@Meta, 2024; Liu et al., 2025)	llama-3	8B	2024-04-29	6854	-564.98	59.84
326	gradientai/Llama-3-8B-Instruct-262k (Ding et al., 2023; Peng et al., 2023b; Pekelis et al., 2024; AI@Meta, 2024; Liu et al., 2025)	llama-3	8B	2024-04-25	12536	-555.17	60.26
327	GritLM/GritLM-7B (Muennighoff et al., 2025)	mistral	7B	2024-02-11	72991	-564.58	61.41
328	Gryphe/MythoLogic-L2-13b	llama	13B	2023-08-03	1198	-570.07	56.19
329	Gryphe/MythoMax-L2-13b	llama	13B	2023-08-10	14308	-583.80	56.00
330	Gryphe/MythoMix-L2-13b	llama	13B	2023-08-08	1174	-566.44	56.31
331	guardrail/llama-2-7b-guanaco-instruct-sharded	llama-2	6B	2023-07-21	1320	-647.09	50.58
332	gywy/llama2-13b-chinese-v2	llama-2	13B	2023-08-22	1163	-719.40	49.58
333	habanoz/tinyllama-oasst1-top1-instruct-full-lr1-5-v0.1	llama	1B	2023-11-19	1178	-679.96	35.58
334	habanoz/TinyLlama-1.1B-intermediate-step-715k-1.5T-lr-5-1epch-airoboros3.1-1k-instruct-V1	llama	1B	2023-11-22	1279	-658.65	34.98
335	habanoz/TinyLlama-1.1B-intermediate-step-715k-1.5T-lr-5-2.2epochs-oasst1-top1-instruct-V1	llama	1B	2023-11-20	1283	-648.88	35.45
336	habanoz/TinyLlama-1.1B-intermediate-step-715k-1.5T-lr-5-3epochs-oasst1-top1-instruct-V1	llama	1B	2023-11-21	1277	-651.34	35.42
337	habanoz/TinyLlama-1.1B-intermediate-step-715k-1.5T-lr-5-4epochs-oasst1-top1-instruct-V1	llama	1B	2023-11-21	1288	-646.55	35.28
338	hakurei/instruct-12b	gpt_neox	12B	2023-04-09	1177	-602.56	38.63
339	hakurei/mommygpt-3B	llama	3B	2023-11-12	1177	-598.84	41.36
340	haoranxu/ALMA-13B (Xu et al., 2024b,a)	llama	13B	2023-09-17	2012	-563.34	50.16
341	haoranxu/ALMA-13B-Pretrain (Xu et al., 2024b,a)	llama	13B	2023-09-17	5510	-558.03	51.68
342	haoranxu/ALMA-7B (Xu et al., 2024b,a)	llama	7B	2023-09-17	1292	-599.26	45.32
343	harborwater/open-llama-3b-claude-30k	llama	3B	2023-11-21	1200	-603.17	40.93
344	health360/Healix-1.1B-V1-Chat-dDPO	llama	1B	2023-11-05	2790	-796.10	33.00
345	heegyu/LIMA2-13b-hf (Touvron et al., 2023b)	llama-2	13B	2023-08-07	3330	-596.33	52.98
346	heegyu/LIMA2-7b-hf (Touvron et al., 2023b)	llama-2	7B	2023-08-04	3463	-651.31	49.27
347	heegyu/LIMA-13b-hf	llama	13B	2023-08-01	3325	-544.21	52.61
348	heegyu/RedTulu-Uncensored-3B-0719	gpt_neox	3B	2023-07-23	1187	-703.96	39.19
349	heegyu/WizardVicuna2-13b-hf (Touvron et al., 2023b)	llama-2	13B	2023-08-07	6481	-612.15	51.05
350	heegyu/WizardVicuna-open-llama-3b-v2	llama	3B	2023-08-25	9364	-671.17	38.77
351	heegyu/WizardVicuna-Uncensored-3B-0719	llama	3B	2023-07-23	1168	-638.46	39.73
352	heegyu/WizardVicuna-3B-0719	llama	3B	2023-07-23	3356	-655.31	39.48
353	HenryJJ/Instruct_Mistral-7B-v0.1_Dolly15K	mistral	7B	2024-01-02	1166	-577.98	60.45
354	HenryJJ/Instruct_Yi-6B_Dolly15K	llama	6B	2024-01-06	1170	-524.87	56.85
355	HenryJJ/Instruct_Yi-6B_Dolly_CodeAlpaca	llama	6B	2024-01-07	1177	-527.06	56.11
356	hfl/chinese-alpaca-2-13b	llama	13B	2023-08-14	1308	-592.75	57.41
357	hfl/chinese-llama-2-13b	llama-2	13B	2023-08-11	1176	-665.69	52.00

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
358	hfl/chinese-llama-2-1.3b	llama-2	1B	2023-10-08	2269	-1132.44	28.59
359	HiTZ/GoLLIE-7B (Sainz et al., 2024)	llama-2	7B	2023-09-25	1394	-632.52	37.48
360	hiyouga/Baichuan2-7B-Base-LLaMAfied	llama-2	7B	2023-09-08	1218	-530.80	48.99
361	hiyouga/Baichuan2-7B-Chat-LLaMAfied	llama-2	7B	2023-09-09	1211	-609.32	51.42
362	hoskinson-center/proofGPT-v0.1-6.7B	gpt_neox	6B	2023-02-04	1187	-974.98	29.72
363	hpcai-tech/Colossal-LLaMA-2-7b-base (Li et al., 2023b; Touvron et al., 2023b; Dao, 2023)	llama-2	7B	2023-09-18	1192	-690.99	51.39
364	HuggingFaceFW/ablation-model-fineweb-v1 (Lacoste et al., 2019)	llama	1B	2024-04-20	2238	-771.42	36.76
365	HuggingFaceH4/mistral-7b-sft-beta	mistral	7B	2023-10-26	7408	-553.64	59.78
366	HuggingFaceH4/zephyr-7b-alpha (Ding et al., 2023; Tunstall et al., 2023; Rafailov et al., 2024; Cui et al., 2024)	mistral	7B	2023-10-09	12911	-559.73	59.50
367	HuggingFaceH4/zephyr-7b-beta (Ding et al., 2023; Tunstall et al., 2023; Rafailov et al., 2024; Cui et al., 2024)	mistral	7B	2023-10-26	294019	-570.40	59.08
368	huggyllama/llama-13b	llama-1	13B	2023-04-03	6345	-541.90	51.33
369	huggyllama/llama-7b	llama-1	6B	2023-04-03	192383	-562.04	46.37
370	HWERI/Llama2-7b-sharegpt4	llama-2	7B	2023-08-04	1200	-660.22	51.05
371	HWERI/pythia-1.4b-deduped-sharegpt	gpt_neox	1B	2023-08-10	1200	-557.93	35.11
372	hyunseoki/ko-en-llama2-13b	llama-2	13B	2023-10-02	3351	-550.60	51.27
373	hyunseoki/ko-ref-llama2-13b	llama-2	13B	2023-10-04	3364	-775.26	43.62
374	hyunseoki/ko-ref-llama2-7b	llama-2	7B	2023-10-04	3312	-851.38	40.75
375	hywu/Camelidae-8x13B (Houlsby et al., 2019; Dettmers et al., 2023; Komatsuzaki et al., 2023; Wu et al., 2024)	camelidae	13B	2024-01-10	1886	-535.46	59.40
376	hywu/Camelidae-8x7B (Houlsby et al., 2019; Dettmers et al., 2023; Komatsuzaki et al., 2023; Wu et al., 2024)	camelidae	7B	2024-01-10	1901	-565.23	54.47
377	h2oai/h2ogpt-gm-oasst1-en-1024-12b	gpt_neox	12B	2023-05-02	1185	-495.39	40.65
378	h2oai/h2ogpt-gm-oasst1-en-2048-open-llama-7b-preview-300bt	llama-1	7B	2023-05-04	1185	-1042.98	34.32
379	h2oai/h2ogpt-gm-oasst1-en-2048-open-llama-7b-preview-300bt-v2	llama-1	7B	2023-05-10	1190	-746.41	37.55
380	h2oai/h2ogpt-oasst1-512-12b	gpt_neox	12B	2023-04-17	1319	-484.68	40.48
381	h2oai/h2ogpt-oig-oasst1-256-6_9b	gpt_neox	9B	2023-04-17	1191	-492.94	38.62
382	h2oai/h2ogpt-oig-oasst1-512-6_9b	gpt_neox	9B	2023-04-18	1749	-521.42	38.52
383	ibivibiv/llama-3-nectar-dpo-8B (Lacoste et al., 2019)	llama-3	8B	2024-05-14	6085	-576.43	67.92
384	ibm/merlinite-7b (Sudalairaj et al., 2024)	mistral	7B	2024-03-02	11156	-541.12	64.00
385	ibndias/NeuralHermes-MoE-2x7B	mixtral	12B	2024-01-03	1172	-533.10	64.08
386	ibranze/araproje-llama2-7b-hf	llama-2	7B	2023-10-06	1161	-549.87	49.73
387	IDEA-CCNL/Ziya-LLaMA-13B-Pretrain-v1 (IDEA-CCNL, 2021; Yang et al., 2022; Zhang et al., 2022a)	llama-1	13B	2023-06-01	1171	-1397.30	29.96
388	IDEA-CCNL/Ziya-LLaMA-13B-v1 (IDEA-CCNL, 2021; Yang et al., 2022; Zhang et al., 2022a)	llama-1	13B	2023-05-16	1244	-1397.34	29.82
389	ignos/LeoScorpius-GreenNode-Alpaca-7B-v1	mistral	7B	2023-12-15	1206	-558.75	74.74
390	ignos/LeoScorpius-GreenNode-Platypus-7B-v1	mistral	7B	2023-12-15	1188	-540.87	68.96
391	ignos/Mistral-T5-7B-v1	mistral	7B	2023-12-18	1266	-557.24	72.47
392	ikala/bloom-zh-3b-chat	bloom	3B	2023-05-07	1352	-722.56	37.58
393	IkariDev/Athena-v1	llama	13B	2023-08-30	1225	-566.12	54.11
394	IkariDev/Athena-v4	llama	13B	2023-10-07	1214	-555.80	57.23
395	INSAIT-Institute/BgGPT-7B-Instruct-v0.2	mistral	7B	2024-03-03	3265	-594.56	63.08
396	Intel/neural-chat-7b-v3-1 (Mukherjee et al., 2023)	mistral	7B	2023-11-14	3976	-563.71	59.90
397	Intel/neural-chat-7b-v3-2 (Yu et al., 2024b)	mistral	7B	2023-11-21	2102	-554.92	68.29
398	Intel/neural-chat-7b-v3-3 (Yu et al., 2024b)	mistral	7B	2023-12-09	166226	-575.02	69.83
399	invalid-coder/Sakura-SOLAR-Instruct-CarbonVillain-en-10.7B-v2-slerp	llama	10B	2024-01-10	12810	-560.73	74.45
400	itsliupeng/llama2_7b_code	llama-2	7B	2023-09-28	1236	-538.58	49.05
401	itsliupeng/openllama-7b-base	llama	7B	2023-12-08	1200	-536.60	47.09
402	itsliupeng/openllama-7b-icl (Shi et al., 2024)	llama	7B	2023-12-08	1186	-535.25	47.93
403	jae24/openhermes_dpo_norobot_0201	mistral	7B	2024-01-02	1161	-562.15	63.78
404	jan-hq/trinity-v1	mistral	7B	2023-12-14	1192	-559.84	74.80
405	Jayant9928/orpo_med_v3	llama	8B	2024-05-01	2705	-557.53	62.21
406	jeonsworld/CarbonVillain-en-10.7B-v1	llama	10B	2023-12-28	1164	-561.74	74.28
407	jeonsworld/CarbonVillain-en-10.7B-v4	llama	10B	2023-12-30	12898	-561.44	74.52
408	jerryjalapeno/nart-100k-7b	llama-1	7B	2023-07-14	1194	-585.01	46.39
409	Jiayi-Pan/Tiny-Vicuna-1B	llama	1B	2023-11-22	3094	-647.57	34.76
410	jingyeom/freeze_KoSoLAR-10.7B-v0.2_1.4_dedup	llama	10B	2024-01-29	2287	-594.17	60.06
411	johnsnowlabs/BioLing-7B-Dare	mistral	7B	2024-04-08	2682	-591.07	60.32
412	johnsnowlabs/JSL-MedLlama-3-8B-v2.0	llama-3	8B	2024-04-30	11813	-569.39	61.93
413	jondurbin/airoboros-gpt-3.5-turbo-100k-7b	llama-1	7B	2023-05-12	1473	-614.18	47.05
414	jondurbin/airoboros-l2-13b-gpt4-m2.0	llama	13B	2023-07-28	1378	-614.41	52.66
415	jondurbin/airoboros-l2-13b-gpt4-2.0	llama	13B	2023-07-27	1395	-575.42	52.49
416	jondurbin/airoboros-l2-13b-2.1	llama-2	13B	2023-08-28	2322	-565.56	53.34
417	jondurbin/bagel-8b-v1.0	llama-3	8B	2024-04-24	7960	-512.36	67.84
418	Josephgflowers/TinyLlama-3T-Cinder-v1.2	llama	1B	2023-12-31	1413	-777.50	35.26
419	JosephusCheung/Qwen-LLaMAfied-7B-Chat	llama-2	7B	2023-08-04	1221	-646.77	51.99
420	JosephusCheung/Qwen-VL-LLaMAfied-7B-Chat	llama-2	7B	2023-08-30	1186	-788.30	45.00
421	jphme/em_german_leo_mistral	mistral	7B	2023-10-07	1900	-625.44	51.69
422	jphme/Llama-2-13b-chat-german (Touvron et al., 2023b)	llama-2	13B	2023-07-21	1265	-572.28	55.07
423	jphme/orca_mini_v2_ger_7b (Su et al., 2023b; Mathur, 2023a; Conover et al., 2023; Touvron et al., 2023a; Harries, 2023; Xu et al., 2023a; Taori et al., 2023)	llama-1	7B	2023-07-04	1181	-603.44	47.65
424	junelee/wizard-vicuna-13b	llama-1	13B	2023-05-03	2196	-610.98	52.73
425	Kabster/BioMistral-Zephyr-Beta-SLERP	mistral	7B	2024-03-09	2758	-597.30	56.35
426	Kabster/Bio-Mistralv2-Squared	mistral	7B	2024-03-09	2762	-603.34	57.73
427	kaist-ai/mistral-orpo-capybara-7k (Hong et al., 2024)	mistral	7B	2024-03-23	4821	-540.67	63.36
428	kekmodel/StopCarbon-10.7B-v5	llama	10B	2023-12-30	13910	-560.12	74.41
429	kevin009/llamaRAGdrama	mistral	7B	2024-02-04	4248	-610.94	74.65
430	kfkas/Llama-2-ko-7b-Chat (Touvron et al., 2023b)	llama-2	7B	2023-07-25	3436	-767.03	40.27
431	kingbri/airolima-chronos-grad-l2-13B	llama-2	13B	2023-08-04	1174	-587.96	55.50
432	kingbri/chronolima-airo-grad-l2-13B	llama-2	13B	2023-08-04	1182	-582.73	55.50
433	klyang/MentaLLaMA-chat-7B (Yang et al., 2024b)	llama	7B	2023-09-26	2634	-640.26	51.17
434	KnutJaegersberg/deacon-13b	llama	13B	2023-09-20	2018	-533.27	53.63
435	KnutJaegersberg/deacon-3b	llama	3B	2023-09-18	2016	-621.85	39.05
436	KnutJaegersberg/falcon-1b-t-sft	falcon	1B	2023-12-04	2029	-978.54	35.02

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
437	KnutJaegersberg/LLongMA-3b-LIMA	llama	3B	2023-09-03	2010	-617.15	38.51
438	KnutJaegersberg/MistralInstructLongish	mistral	7B	2023-11-15	1970	-543.74	53.62
439	KnutJaegersberg/Qwen-1_8B-Llamafied	llama	1B	2024-01-03	2839	-760.12	44.75
440	KnutJaegersberg/Walter-Falcon-1B	falcon	1B	2023-12-09	2014	-1010.10	34.07
441	KnutJaegersberg/webMistral-7B	mistral	7B	2023-11-17	1972	-554.62	53.97
442	KoboldAI/GPT-J-6B-Adventure	gptj	6B	2022-03-02	1297	-661.77	35.95
443	KoboldAI/GPT-J-6B-Janeway (Gao et al., 2020; Wang and Komatsuzaki, 2021)	gptj	6B	2022-03-02	4382	-496.37	39.54
444	KoboldAI/GPT-J-6B-Shinen (Gao et al., 2020; Wang and Komatsuzaki, 2021)	gptj	6B	2022-03-02	1498	-498.02	39.60
445	KoboldAI/GPT-J-6B-Skein (Lacoste et al., 2019; Wang, 2021)	gptj	6B	2022-03-02	1236	-509.27	40.02
446	KoboldAI/LLaMA2-13B-Holomax	llama-2	13B	2023-08-14	1255	-563.47	54.52
447	KoboldAI/LLaMA2-13B-Tiefighter	llama-2	13B	2023-10-18	2321	-599.88	54.51
448	KoboldAI/OPT-13B-Erebus (Zhang et al., 2022b)	opt	13B	2022-09-09	6998	-652.81	39.61
449	KoboldAI/OPT-13B-Nerybus-Mix (Zhang et al., 2022b)	opt	13B	2023-02-13	1657	-646.08	39.61
450	KoboldAI/OPT-13B-Nerys-v2 (Zhang et al., 2022b)	opt	13B	2022-09-19	4259	-652.01	39.53
451	KoboldAI/OPT-2.7B-Erebus (Zhang et al., 2022b)	opt	2B	2022-09-19	5559	-692.73	36.96
452	KoboldAI/OPT-2.7B-Nerybus-Mix	opt	2B	2023-02-09	1464	-686.59	36.88
453	KoboldAI/OPT-6B-nerys-v2 (Zhang et al., 2022b)	opt	6B	2022-06-26	4974	-656.62	38.72
454	KoboldAI/OPT-6.7B-Erebus (Zhang et al., 2022b)	opt	6B	2022-09-15	5587	-641.11	39.09
455	KoboldAI/OPT-6.7B-Nerybus-Mix	opt	6B	2023-02-13	1585	-642.68	38.83
456	kodonho/SolarM-SakuraSolar-SLERP	llama	10B	2024-01-12	3254	-562.65	74.29
457	kodonho/Solar-OrcaDPO-Solar-Instruct-SLERP	llama	10B	2024-01-12	3266	-560.31	74.35
458	Korabbit/Llama-2-7b-chat-hf-afr-100step-flan	llama-2	7B	2023-11-30	1208	-650.87	52.88
459	Korabbit/Llama-2-7b-chat-hf-afr-100step-flan-v2	llama-2	7B	2023-12-03	1208	-650.96	52.92
460	Korabbit/Llama-2-7b-chat-hf-afr-100step-v2	llama-2	7B	2023-11-22	1205	-655.71	50.89
461	Korabbit/Llama-2-7b-chat-hf-afr-200step-flan	llama-2	7B	2023-11-30	1201	-647.72	52.62
462	Korabbit/Llama-2-7b-chat-hf-afr-200step-flan-v2	llama-2	7B	2023-12-03	1211	-648.67	52.75
463	Korabbit/Llama-2-7b-chat-hf-afr-200step-merged	llama-2	7B	2023-11-21	1222	-653.05	52.26
464	Korabbit/Llama-2-7b-chat-hf-afr-200step-v2	llama-2	7B	2023-11-22	1203	-658.69	50.21
465	Korabbit/Llama-2-7b-chat-hf-afr-300step-flan-v2	llama-2	7B	2023-12-03	1208	-647.99	52.41
466	Korabbit/Llama-2-7b-chat-hf-afr-441step-flan-v2	llama-2	7B	2023-12-03	1209	-647.85	52.28
467	Kukedlc/NeuralExperiment-7b-MagicCoder-v7.5	mistral	7B	2024-03-07	4044	-569.72	74.28
468	Kukedlc/NeuralLLaMa-3-8b-DT-v0.1	llama-3	8B	2024-05-11	5707	-571.65	72.52
469	Kukedlc/NeuralLLaMa-3-8b-ORPO-v0.3	llama-3	8B	2024-05-14	7985	-562.41	72.66
470	Kukedlc/NeuralSynthesis-7B-v0.1	mistral	7B	2024-04-06	8910	-596.75	76.80
471	Kukedlc/NeuralSynthesis-7B-v0.3	mistral	7B	2024-04-07	3114	-597.52	76.70
472	Kukedlc/NeuralSynthesis-7b-v0.4-slerp	mistral	7B	2024-04-12	3067	-597.78	76.76
473	kyujinpy/Sakura-SOLAR-Instruct	llama	10B	2023-12-24	4271	-559.44	74.40
474	kyujinpy/Sakura-SOLAR-Instruct-DPO-v2	llama	10B	2023-12-24	3298	-560.08	74.14
475	kyujinpy/Sakura-SOLRCA-Math-Instruct-DPO-v1	llama	10B	2023-12-25	3284	-562.45	74.13
476	Lazycuber/L2-7b-Base-Guanaco-Uncensored	llama	7B	2023-09-19	1172	-555.04	50.45
477	lcw99/llama-3-10b-it-kor-extented-chang	llama-3	9B	2024-05-15	2230	-558.58	54.76
478	lcw99/llama-3-10b-it-kor-extented-chang-pro8	llama-3	9B	2024-05-21	2234	-561.88	63.76
479	lcw99/llama-3-10b-it-ko-2024-0527	llama-3	9B	2024-05-27	2236	-564.03	63.70
480	lcw99/llama-3-8b-it-kor-extented-chang	llama-3	8B	2024-05-02	2259	-527.86	66.27
481	LDCC/LDCC-SOLAR-10.7B (Kim et al., 2024b)	llama	10B	2024-01-03	3267	-549.97	71.40
482	lemon-mint/gemma-ko-1.1-2b-it	gemma	2B	2024-04-26	2337	-1058.64	30.92
483	lemon-mint/gemma-ko-7b-instruct-v0.62	gemma	8B	2024-04-03	7543	-589.67	69.25
484	lemon-mint/gemma-ko-7b-instruct-v0.71	gemma	8B	2024-04-09	2232	-711.63	59.23
485	lemon-mint/gemma-7b-openhermes-v0.80	gemma	8B	2024-04-09	4867	-691.34	56.91
486	LeoLM/leo-hessianai-7b	llama	7B	2023-08-22	4502	-592.31	47.72
487	LeoLM/leo-hessianai-7b-chat	llama	7B	2023-09-10	3924	-735.02	49.29
488	leveldevai/TurdusBeagle-7B	mistral	7B	2024-01-18	1926	-577.38	75.15
489	lex-hue/Delexa-7b	mistral	7B	2024-04-05	11797	-575.31	70.86
490	lgaalves/llama-2-13b-chat-platypus	llama-2	13B	2023-09-06	1193	-570.00	53.92
491	lgaalves/llama-2-7b-hf_open-platypus	llama-2	6B	2023-08-30	1204	-564.27	49.73
492	lgaalves/mistral-7b-platypus1k	mistral	7B	2023-10-10	1183	-539.54	58.19
493	lgaalves/mistral-7b_open_platypus	mistral	7B	2023-10-13	1259	-557.93	56.29
494	lgaalves/tinyllama-1.1b-chat-v0.3_platypus	llama	1B	2023-10-09	1193	-665.44	34.50
495	lightblue/suzume-llama-3-8B-multilingual (Devine, 2024)	llama-3	8B	2024-04-23	16442	-557.50	65.55
496	lighteternal/Llama3-merge-biomed-8b (Yadav et al., 2023a; Yu et al., 2024a)	llama-3	8B	2024-05-28	2731	-575.20	66.30
497	liminerity/M7-7b	mistral	7B	2024-03-07	4186	-599.03	76.82
498	LinkSoul/Chinese-Llama-2-7b	llama-2	7B	2023-07-20	40799	-584.42	52.59
499	llm-agents/tora-code-7b-v1.0 (Gou et al., 2024)	llama-2	7B	2023-10-08	1242	-682.23	40.21
500	llm-agents/tora-7b-v1.0 (Gou et al., 2024)	llama-2	7B	2023-10-08	1177	-628.41	48.50
501	lmsys/longchat-13b-16k	llama-1	13B	2023-06-28	12228	-621.21	49.64
502	lmsys/longchat-7b-v1.5-32k	llama	7B	2023-08-01	5114	-650.92	47.95
503	lmsys/vicuna-13b-delta-v1.1 (Zheng et al., 2023; Touvron et al., 2023a)	llama-1	13B	2023-04-12	1216	-1396.42	53.28
504	lmsys/vicuna-13b-v1.1 (Zheng et al., 2023; Touvron et al., 2023a)	llama-1	13B	2023-04-12	2662	-646.61	53.28
505	lmsys/vicuna-13b-v1.3 (Zheng et al., 2023; Touvron et al., 2023a)	llama-1	13B	2023-06-18	11951	-604.80	54.27
506	lmsys/vicuna-13b-v1.5 (Touvron et al., 2023b; Zheng et al., 2023)	llama-2	13B	2023-07-29	48653	-585.07	55.41
507	lmsys/vicuna-13b-v1.5-16k (Touvron et al., 2023b; Zheng et al., 2023)	llama-2	13B	2023-08-01	29580	-595.73	54.97
508	lmsys/vicuna-7b-delta-v1.1 (Zheng et al., 2023; Touvron et al., 2023a)	llama-1	7B	2023-04-12	1330	-1396.42	50.37
509	lmsys/vicuna-7b-v1.3 (Zheng et al., 2023; Touvron et al., 2023a)	llama-1	7B	2023-06-18	31436	-621.79	49.78
510	lmsys/vicuna-7b-v1.5 (Touvron et al., 2023b; Zheng et al., 2023)	llama-2	7B	2023-07-29	368410	-608.63	52.06
511	lmsys/vicuna-7b-v1.5-16k (Touvron et al., 2023b; Zheng et al., 2023)	llama-2	7B	2023-07-31	3362	-622.49	51.42
512	Locutusque/Hercules-2.5-Mistral-7B	mistral	7B	2024-02-10	1953	-533.73	63.59
513	Locutusque/Hercules-3.1-Mistral-7B	mistral	7B	2024-02-19	2780	-528.16	62.09
514	Locutusque/Orca-2-13b-SFT-v4	llama	13B	2023-11-25	2345	-617.05	59.75
515	Locutusque/Orca-2-13b-SFT-v6	llama	13B	2023-12-22	2319	-674.35	56.15
516	Locutusque/Orca-2-13b-SFT_v5	llama	13B	2023-12-13	2191	-602.83	56.77
517	LTC-AI-Labs/L2-7b-Base-WVG-Uncensored	llama	7B	2023-09-23	1241	-556.11	50.63
518	LTC-AI-Labs/L2-7b-Beluga-WVG-Test	llama	7B	2023-10-03	1222	-567.10	52.04
519	LTC-AI-Labs/L2-7b-Hermes-Synthia	llama-2	7B	2023-11-23	1232	-562.42	52.21
520	LTC-AI-Labs/L2-7b-Hermes-WVG-Test	llama	7B	2023-09-27	1225	-570.23	51.35
521	LTC-AI-Labs/L2-7b-Synthia-WVG-Test	llama	7B	2023-09-28	1233	-576.22	51.25

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
522	luffycodes/nash-vicuna-13b-v1dot5-ep2-w-rag-w-simple (Sonkar et al., 2023)	llama-2	13B	2023-08-21	1165	-603.30	55.40
523	luffycodes/vicuna-class-shishya-13b-ep3 (Sonkar et al., 2023)	llama-2	13B	2023-12-21	2092	-633.43	48.52
524	luffycodes/vicuna-class-shishya-7b-ep3 (Sonkar et al., 2023)	llama-2	7B	2023-12-14	3501	-649.39	46.14
525	luffycodes/vicuna-class-tutor-13b-ep3 (Sonkar et al., 2023)	llama-2	13B	2023-12-21	1514	-605.75	55.88
526	luffycodes/vicuna-class-tutor-7b-ep3 (Sonkar et al., 2023)	llama-2	7B	2023-12-15	3735	-643.38	51.45
527	lu-vae/llama2-13B-sharegpt4-orca-openplatypus-8w	llama-2	13B	2023-09-14	1190	-544.97	55.75
528	lu-vae/llama2-13b-sharegpt4-test	llama-2	13B	2023-09-07	1195	-557.27	55.69
529	lyogavin/Anima-7B-100K	llama-2	7B	2023-09-14	1207	-611.20	42.98
530	l3utterfly/llama2-7b-layla	llama-2	7B	2023-08-07	1182	-561.12	52.05
531	l3utterfly/minima-3b-layla-v1	llama-2	3B	2023-12-12	1178	-548.62	43.21
532	l3utterfly/minima-3b-layla-v2	llama-2	3B	2023-12-19	1176	-547.03	43.39
533	l3utterfly/mistral-7b-v0.1-layla-v1	mistral	7B	2023-10-31	1206	-602.61	57.56
534	l3utterfly/mistral-7b-v0.1-layla-v2	mistral	7B	2023-12-16	1199	-567.42	57.60
535	l3utterfly/open-llama-3b-v2-layla	llama	3B	2023-08-18	1172	-832.18	40.25
536	macadeliccc/laser-dolphin-mixtral-2x7b-dpo (Gao et al., 2021a; Sharma et al., 2023)	mixtral	12B	2024-01-08	1275	-594.20	67.16
537	macadeliccc/WestLake-7B-v2-laser-truthy-dpo	mistral	7B	2024-01-27	5064	-593.71	75.37
538	malhajar/Mistral-7B-v0.2-meditron-turkish	mistral	7B	2024-01-05	3865	-641.84	63.34
539	martyn/llama2-megamerge-dare-13b-v2	llama-2	13B	2023-12-17	1213	-594.59	57.94
540	martyn/mistral-megamerge-dare-7b	mistral	7B	2023-12-14	1203	-638.88	48.93
541	martyn/solar-megamerge-dare-10.7b-v1	llama	10B	2023-12-31	1190	-533.45	68.79
542	matsuo-lab/weblab-10b	gpt_neox	10B	2023-08-04	1596	-485.56	38.59
543	matsuo-lab/weblab-10b-instruction-sft	gpt_neox	10B	2023-08-04	1235	-515.77	39.13
544	MayaPH/FinOPT-Franklin	opt	1B	2023-05-26	1207	-1380.66	29.78
545	MayaPH/opt-flan-iml-6.7b (Iyer et al., 2023)	opt	6B	2023-08-15	1193	-729.78	35.84
546	maywell/PiVoT-0.1-early	mistral	7B	2023-11-24	3267	-681.01	64.58
547	maywell/PiVoT-10.7B-Mistral-v0.2	mistral	10B	2023-12-13	3228	-578.51	64.25
548	maywell/Synatra-RP-Orca-2-7b-v0.1	llama	6B	2023-11-21	3278	-632.02	59.65
549	maywell/Synatra-V0.1-7B-Instruct	mistral	7B	2023-10-09	3381	-623.00	55.86
550	maywell/Synatra-10.7B-v0.4	llama	10B	2023-12-27	3282	-532.85	65.48
551	maywell/Synatra-7B-v0.3-dpo	mistral	7B	2023-11-08	4302	-575.15	60.55
552	maywell/Synatra-7B-v0.3-RP	mistral	7B	2023-10-29	8360	-582.37	59.26
553	MaziyarPanahi/Llama-3-8B-Instruct-v0.4	llama-3	8B	2024-05-01	1407	-569.34	70.30
554	MaziyarPanahi/Llama-3-8B-Instruct-v0.8	llama-3	8B	2024-05-01	7281	-576.95	73.17
555	MaziyarPanahi/Llama-3-8B-Instruct-v0.9	llama-3	8B	2024-05-30	6241	-575.60	73.29
556	MaziyarPanahi/Mistral-7B-Instruct-v0.3	mistral	7B	2024-05-22	5230	-549.67	65.21
557	MaziyarPanahi/Mistral-7B-v0.3	mistral	7B	2024-05-22	5758	-531.44	60.40
558	medalpaca/medalpaca-7b (Li et al., 2023d)	llama-1	7B	2023-03-29	8136	-628.90	48.45
559	MediaTek-Research/Breeze-7B-Instruct-v1_0 (Hsu et al., 2024)	mistral	7B	2024-03-05	3065	-559.49	63.40
560	meta-llama/Llama-2-13b-chat-hf (Touvron et al., 2023b)	llama-2	13B	2023-07-13	266090	-616.08	54.91
561	meta-llama/Llama-2-13b-hf (Touvron et al., 2023b)	llama-2	13B	2023-07-13	120871	-530.82	55.69
562	meta-llama/Llama-2-7b-chat-hf (Touvron et al., 2023b)	llama-2	6B	2023-07-13	1402244	-656.64	50.74
563	meta-llama/Llama-2-7b-hf (Touvron et al., 2023b)	llama-2	6B	2023-07-13	1294737	-549.87	50.97
564	meta-llama/Meta-Llama-3-8B (AI@Meta, 2024)	llama-3	8B	2024-04-17	675850	-514.33	62.62
565	meta-llama/Meta-Llama-3-8B-Instruct (AI@Meta, 2024)	llama-3	8B	2024-04-17	2058011	-573.95	66.87
566	meta-math/MetaMath-Llemma-7B (Yu et al., 2024b; Azerbayev et al., 2024)	llama	7B	2023-11-19	1322	-612.21	53.19
567	meta-math/MetaMath-Mistral-7B (Jiang et al., 2023; Yu et al., 2024b)	mistral	7B	2023-10-22	3022	-571.33	65.78
568	microsoft/Orca-2-13b (Mitra et al., 2023)	llama	13B	2023-11-14	13616	-651.29	58.64
569	microsoft/Orca-2-7b (Mitra et al., 2023)	llama	7B	2023-11-14	11084	-684.33	54.55
570	microsoft/phi-1_5 (Li et al., 2023c)	phi	1B	2023-09-10	112476	-724.13	47.69
571	microsoft/phi-2	phi	2B	2023-12-13	237268	-621.44	61.33
572	microsoft/Phi-3-medium-128k-instruct	phi3	13B	2024-05-07	34054	-544.42	73.00
573	microsoft/Phi-3-medium-4k-instruct	phi3	13B	2024-05-07	35987	-552.43	73.45
574	migtissera/Llama-3-8B-Synthia-v3.5	llama-3	8B	2024-05-17	3310	-533.92	67.15
575	migtissera/SynthIA-7B-v1.3 (Mukherjee et al., 2023; Tissera, 2023)	mistral	7B	2023-09-28	3336	-541.46	59.34
576	migtissera/SynthIA-7B-v1.5	mistral	7B	2023-10-07	1228	-537.68	59.59
577	migtissera/Synthia-7B-v3.0	mistral	7B	2023-12-08	1204	-531.17	61.99
578	migtissera/Tess-XS-v1.1	mistral	7B	2023-11-22	1213	-546.97	59.39
579	migtissera/Tess-2.0-Llama-3-8B	llama-3	8B	2024-05-05	3314	-523.68	64.81
580	migtissera/Tess-7B-v1.4	mistral	7B	2023-12-04	1234	-599.60	62.19
581	Mihaiii/Metis-0.3	mistral	7B	2023-12-16	1279	-567.64	65.44
582	mindy-labs/mindy-7b-v2	mistral	7B	2023-12-14	1238	-550.38	72.11
583	Minirecord/Mini_DPO_test02	mistral	7B	2023-11-30	3470	-547.90	61.23
584	mistralai/Mistral-7B-Instruct-v0.1 (Jiang et al., 2023)	mistral	7B	2023-09-27	1370245	-593.06	54.96
585	mistralai/Mistral-7B-Instruct-v0.2 (Jiang et al., 2023)	mistral	7B	2023-12-11	3565248	-574.91	65.71
586	mistralai/Mistral-7B-v0.1 (Jiang et al., 2023)	mistral	7B	2023-09-20	1750089	-527.60	60.97
587	mistralai/Mistral-7B-v0.3 (Jiang et al., 2023)	mistral	7B	2024-05-22	1443065	-531.44	60.28
588	mistral-community/Mistral-7B-v0.2	mistral	7B	2024-03-23	30074	-532.63	60.41
589	mlabonne/AlphaMonarch-7B	mistral	7B	2024-02-14	12804	-593.46	75.99
590	mlabonne/Beagle14-7B	mistral	7B	2024-01-15	1986	-563.97	74.76
591	mlabonne/ChimeraLlama-3-8B-v2	llama-3	8B	2024-04-22	2806	-565.81	69.69
592	mlabonne/ChimeraLlama-3-8B-v3	llama-3	8B	2024-05-01	5718	-568.78	70.06
593	mlabonne/Daredevil-8B-abliterated	llama	8B	2024-05-26	9010	-565.64	71.82
594	mlabonne/GML-Mistral-merged-v1	mistral	8B	2023-12-27	1166	-1384.76	48.54
595	mlabonne/Marcoro14-7B-slerp	mistral	7B	2023-12-29	3792	-554.34	73.01
596	mlabonne/NeuralMarcoro14-7B	mistral	7B	2024-01-06	2015	-566.69	73.57
597	mlabonne/NeuralMonarch-7B	mistral	7B	2024-02-14	13486	-591.01	76.15
598	mncai/agiin-13.6B-v0.1	mistral	13B	2023-12-15	3317	-595.75	68.40
599	MoaData/Myrrh_solar_10.7b_3.0	llama	10B	2024-04-26	9855	-673.46	67.61
600	monology/openinstruct-mistral-7b	mistral	7B	2023-11-20	1191	-537.83	63.64
601	mosaicml/mpt-7b (Henry et al., 2020; Shoeybi et al., 2020; Dao et al., 2022; Press et al., 2022; Touvron et al., 2023a; Chen et al., 2023a; MosaicML NLP Team, 2023b)	mpt	7B	2023-05-05	31480	-548.53	44.28
602	mosaicml/mpt-7b-chat (Henry et al., 2020; Press et al., 2022; Dao et al., 2022; MosaicML NLP Team, 2023b)	mpt	7B	2023-05-04	88069	-647.13	45.39
603	mosaicml/mpt-7b-instruct (Henry et al., 2020; Press et al., 2022; Dao et al., 2022; MosaicML NLP Team, 2023b)	mpt	7B	2023-05-05	8599	-569.62	44.83

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
604	mosaicml/mpt-7b-storywriter (Press et al., 2022; Dao et al., 2022; Chen et al., 2023a; MosaicML NLP Team, 2023b)	mpt	7B	2023-05-04	1920	-650.05	39.31
605	mosaicml/mpt-7b-8k (Henry et al., 2020; Shoeybi et al., 2020; Dao et al., 2022; Press et al., 2022; Touvron et al., 2023a; Chen et al., 2023a; MosaicML NLP Team, 2023b)	mpt	7B	2023-06-30	2106	-546.57	47.24
606	mosaicml/mpt-7b-8k-chat (Henry et al., 2020; Press et al., 2022; Dao et al., 2022; MosaicML NLP Team, 2023a)	mpt	7B	2023-06-22	1304	-586.09	47.78
607	mrm8488/llama-2-coder-7b (Manuel Romero, 2023)	llama-2	7B	2023-07-26	1313	-572.92	49.95
608	MTSAIR/multi_verse_model	mistral	7B	2024-03-07	6357	-598.96	76.74
609	mwitiderrick/open_llama_3b_code_instruct_0.1	llama	3B	2023-12-11	1252	-598.99	39.72
610	NekoPunchBBB/Llama-2-13b-hf_Open-Platypus-QLoRA-multigpu	llama-2	13B	2023-09-15	1208	-536.60	54.40
611	Neko-Institute-of-Science/metharme-7b	llama-1	6B	2023-04-30	1161	-562.36	47.48
612	Neko-Institute-of-Science/pygmalion-7b	llama-1	6B	2023-04-30	1188	-562.13	46.04
613	NeverSleep/Llama-3-Lumimaid-8B-v0.1	llama-3	8B	2024-04-30	1499	-545.48	66.55
614	NeverSleep/Noromaid-7b-v0.2	mistral	7B	2023-12-21	1179	-552.34	61.78
615	NewstaR/Koss-7B-chat	llama	7B	2023-10-01	1180	-660.75	50.37
616	NewstaR/Starlight-7B (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2021a; Gao et al., 2021a; Lin et al., 2022a; Beeching et al., 2023)	llama-2	7B	2023-09-11	1180	-549.87	49.73
617	NExtNewChattingAI/shark_tank_ai_7b_v2	mistral	7B	2023-12-23	1208	-618.14	66.54
618	NExtNewChattingAI/shark_tank_ai_7_b	mistral	7B	2023-12-17	1224	-546.93	71.10
619	Nexusflow/NexusRaven-V2-13B (Nexusflow.ai team, 2023; Rozière et al., 2024)	llama	13B	2023-12-04	3966	-593.55	48.21
620	Nexusflow/Starling-LM-7B-beta (Ziegler et al., 2020; Zhu et al., 2023a)	mistral	7B	2024-03-19	4847	-559.22	69.88
621	NickyNicky/Mistral-7B-OpenOrca-oasst_top1_2023-08-25-v2 (Xiao et al., 2024)	mistral	7B	2023-10-11	1236	-560.11	61.65
622	NickyNicky/Mistral-7B-OpenOrca-oasst_top1_2023-08-25-v3 (Dao et al., 2022; Xiao et al., 2024)	mistral	7B	2023-10-13	1227	-568.73	61.26
623	nlpguy/ColorShadow-7B	mistral	7B	2023-12-30	1214	-552.87	68.34
624	nlpguy/ColorShadow-7B-v2	mistral	7B	2023-12-30	1214	-569.63	66.88
625	nlpguy/ColorShadow-7B-v3	mistral	7B	2023-12-30	1222	-561.80	67.29
626	Norquinal/llama-2-7b-claude-chat	llama-2	7B	2023-08-11	1221	-557.34	50.98
627	Norquinal/llama-2-7b-claude-chat-rp	llama-2	7B	2023-08-14	1218	-557.09	51.25
628	Norquinal/Mistral-7B-claude-instruct	mistral	7B	2023-09-28	1248	-534.14	59.27
629	NousResearch/CodeLlama-13b-hf	llama	13B	2023-08-24	6299	-582.02	43.35
630	NousResearch/CodeLlama-7b-hf	llama	7B	2023-08-24	9428	-597.05	39.81
631	NousResearch/Hermes-2-Pro-Llama-3-8B ("Teknium" et al., 2024b)	llama-3	8B	2024-04-30	26797	-576.90	68.73
632	NousResearch/Hermes-2-Pro-Mistral-7B ("interstellarninja" et al., 2024)	mistral	7B	2024-03-11	14586	-620.93	67.35
633	NousResearch/Hermes-2-Theta-Llama-3-8B ("Teknium" et al., 2024a)	llama-3	8B	2024-05-05	12392	-565.47	69.21
634	NousResearch/Meta-Llama-3-8B-Instruct (AI@Meta, 2024)	llama-3	8B	2024-04-18	104388	-573.95	67.10
635	NousResearch/Nous-Hermes-Llama2-13b	llama-2	13B	2023-07-20	39412	-586.34	55.75
636	NousResearch/Nous-Hermes-llama-2-7b	llama-2	6B	2023-07-25	12019	-569.93	51.87
637	NousResearch/Nous-Hermes-13b	llama-1	13B	2023-06-03	1536	-605.76	54.04
638	NousResearch/Nous-Hermes-2-Mistral-7B-DPO ("Teknium" et al., 2024c)	mistral	7B	2024-02-18	8725	-565.26	68.10
639	NousResearch/Nous-Hermes-2-SOLAR-10.7B	llama	10B	2024-01-01	9918	-542.37	71.00
640	NousResearch/Yarn-Mistral-7b-128k (Peng et al., 2023b)	mistral	7B	2023-10-31	20727	-532.22	59.42
641	NousResearch/Yarn-Mistral-7b-64k (Peng et al., 2023b)	mistral	7B	2023-10-31	10368	-532.16	59.63
642	NTQAI/Nxcode-CQ-7B-orpo (Hong et al., 2024)	qwen2	7B	2024-04-24	12700	-685.73	42.98
643	nvidia/Llama3-ChatQA-1.5-8B (Liu et al., 2024)	llama-3	8B	2024-04-28	10608	-515.82	56.71
644	occultml/CatMarcoro14-7B-slerp	mistral	7B	2024-01-06	1234	-551.33	73.25
645	occultml/Helios-10.7B	llama	7B	2023-12-31	1215	-1381.38	42.19
646	occultml/Helios-10.7B-v2	llama	7B	2023-12-31	1218	-1381.35	42.25
647	OEvortex/EMO-2B	gemma	2B	2024-04-28	4137	-976.17	44.26
648	oh-yeontaek/llama-2-13B-LoRA-assemble	llama-2	13B	2023-09-13	3303	-568.38	57.91
649	oh-yeontaek/llama-2-7B-LoRA-assemble	llama-2	7B	2023-09-13	1337	-614.33	52.26
650	openaccess-ai-collective/DPOpenHermes-11B	mistral	10B	2023-12-03	1185	-591.33	66.83
651	openaccess-ai-collective/jackalope-7b (Mukherjee et al., 2023; Longpre et al., 2023; Lian et al., 2023b)	mistral	7B	2023-10-07	1198	-550.33	61.16
652	openaccess-ai-collective/manticore-13b	llama-1	13B	2023-05-17	1224	-553.22	54.86
653	openaccess-ai-collective/manticore-13b-chat-pyg	llama-1	13B	2023-05-22	2148	-553.85	54.13
654	openaccess-ai-collective/minotaur-13b	llama-1	13B	2023-06-06	1202	-577.38	53.97
655	openaccess-ai-collective/minotaur-13b-fixed	llama-1	13B	2023-06-12	1194	-568.52	55.19
656	openaccess-ai-collective/mistral-7b-slimorcaboros	mistral	7B	2023-10-13	1186	-569.20	61.18
657	openaccess-ai-collective/wizard-mega-13b	llama-1	13B	2023-05-14	2168	-565.46	54.27
658	OpenAssistant/codellama-13b-oasst-sft-v10	llama-2	13B	2023-08-26	2122	-591.56	44.85
659	OpenAssistant/llama2-13b-orca-8k-3319 (Mukherjee et al., 2023)	llama-2	13B	2023-07-24	1270	-531.87	55.09
660	OpenAssistant/oasst-sft-1-pythia-12b	gpt_neox	12B	2023-03-09	22359	-615.25	40.77
661	OpenAssistant/oasst-sft-4-pythia-12b-epoch-3.5	gpt_neox	12B	2023-04-03	458718	-572.27	41.31
662	OpenAssistant/pythia-12b-sft-v8-7k-steps	gpt_neox	12B	2023-05-07	1323	-524.99	42.21
663	OpenAssistant/stablelm-7b-sft-v7-epoch-3	gpt_neox	7B	2023-04-20	1215	-764.93	34.85
664	openbmb/UltraLM-13b-v2.0	llama	13B	2023-09-22	1174	-554.24	58.72
665	OpenBuddy/openbuddy-atom-13b-v9-bf16	llama	13B	2023-08-05	1206	-625.44	52.31
666	OpenBuddy/openbuddy-llama2-13b-v11-bf16	llama-2	13B	2023-08-23	1199	-606.17	52.93
667	OpenBuddy/openbuddy-llama2-13b-v11.1-bf16	llama-2	13B	2023-08-24	1210	-595.93	55.28
668	OpenBuddy/openbuddy-llama2-13b-v8.1-fp16	llama-2	13B	2023-07-25	7339	-587.40	57.76
669	OpenBuddy/openbuddy-llama3-8b-v21.1-8k	llama-3	8B	2024-04-20	2293	-578.14	65.31
670	OpenBuddy/openbuddy-mistral2-7b-v20.3-32k	mistral	7B	2024-03-27	2302	-642.78	62.73
671	OpenBuddy/openbuddy-mistral-7b-v13	mistral	7B	2023-10-10	1335	-643.47	53.50
672	OpenBuddy/openbuddy-mistral-7b-v13-base	mistral	7B	2023-10-11	1197	-677.88	51.99
673	OpenBuddy/openbuddy-mistral-7b-v13.1	mistral	7B	2023-10-11	1217	-638.68	52.62
674	OpenBuddy/openbuddy-openllama-13b-v7-fp16	llama-1	13B	2023-07-03	1209	-654.65	49.31
675	OpenBuddy/openbuddy-openllama-3b-v10-bf16	llama	3B	2023-08-10	1206	-794.41	36.87
676	OpenBuddy/openbuddy-openllama-7b-v12-bf16	llama	7B	2023-09-19	3153	-854.52	45.28
677	OpenBuddy/openbuddy-zephyr-7b-v14.1	mistral	7B	2023-11-06	3392	-657.03	51.86
678	openchat/openchat-3.5-0106 (Wang et al., 2024a; OpenAI et al., 2024)	mistral	7B	2024-01-07	26858	-554.42	69.30
679	openchat/openchat-3.5-0106-gemma (Wang et al., 2024a)	gemma	8B	2024-03-09	7817	-937.52	69.42
680	openchat/openchat-3.5-1210 (Wang et al., 2024a; OpenAI et al., 2024)	mistral	7B	2023-12-12	2123	-583.16	68.89
681	openchat/openchat-3.6-8b-20240522 (Wang et al., 2024a)	llama-3	8B	2024-05-07	10464	-561.80	68.14

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
682	openlm-research/open_llama_13b (Geng and Liu, 2023; Together Computer, 2023; Touvron et al., 2023a)	llama-1	13B	2023-06-15	2373	-582.20	47.26
683	openlm-research/open_llama_3b (Geng and Liu, 2023; Together Computer, 2023; Touvron et al., 2023a)	llama-1	3B	2023-06-07	157405	-612.46	38.26
684	openlm-research/open_llama_3b_v2 (Geng and Liu, 2023; Together Computer, 2023; Touvron et al., 2023a)	llama-1	3B	2023-07-16	21275	-578.08	40.28
685	openlm-research/open_llama_7b (Geng and Liu, 2023; Together Computer, 2023; Touvron et al., 2023a)	llama-1	7B	2023-06-07	44968	-610.20	42.31
686	openlm-research/open_llama_7b_v2 (Geng and Liu, 2023; Together Computer, 2023; Touvron et al., 2023a)	llama-1	7B	2023-07-06	2792	-556.22	44.26
687	OpenModels4all/gemma-1.1-7b-it (Joshi et al., 2017; Zhao et al., 2018; Mihaylov et al., 2018; Zellers et al., 2019; Clark et al., 2019; Chollet, 2019; Sakaguchi et al., 2019; Talmor et al., 2019; Bisk et al., 2019; Sap et al., 2019; Hendrycks et al., 2021a; Cobbe et al., 2021; Chen et al., 2021; Austin et al., 2021; Parrish et al., 2022; Zhong et al., 2023; Srivastava et al., 2023; Gemini Team et al., 2024)	gemma	8B	2024-04-06	5128	-1355.60	59.78
688	OpenPipe/mistral-ft-optimized-1218	mistral	7B	2023-12-17	1405	-546.73	71.94
689	OpenPipe/mistral-ft-optimized-1227	mistral	7B	2023-12-27	6273	-564.28	70.54
690	openthaigpt/openthaigpt-1.0.0-beta-13b-chat-hf	llama	13B	2023-12-18	1225	-712.61	50.45
691	openthaigpt/openthaigpt-1.0.0-beta-7b-chat-ckpt-hf	llama	7B	2023-08-14	1291	-772.50	45.35
692	Open-Orca/LlongOrca-13B-16k (Wang et al., 2023a; Mukherjee et al., 2023; Dale et al., 2023; Touvron et al., 2023b; Longpre et al., 2023)	llama-2	13B	2023-08-16	1217	-551.07	56.59
693	Open-Orca/LlongOrca-7B-16k (Wang et al., 2023a; Lian et al., 2023c; Mukherjee et al., 2023; Touvron et al., 2023b; Longpre et al., 2023)	llama-2	7B	2023-08-05	1223	-583.06	53.02
694	Open-Orca/Mistral-7B-OpenOrca (Mukherjee et al., 2023; Longpre et al., 2023; Lian et al., 2023d)	mistral	7B	2023-09-29	18070	-575.75	60.17
695	Open-Orca/Mistral-7B-SlimOrca (Mukherjee et al., 2023; Longpre et al., 2023; Lian et al., 2023f,e)	mistral	7B	2023-10-08	3696	-569.42	60.37
696	Open-Orca/OpenOrcaxOpenChat-Preview2-13B (Wang et al., 2023a; Mukherjee et al., 2023; Touvron et al., 2023b; Longpre et al., 2023; Wang et al., 2023b)	llama-2	13B	2023-07-31	2172	-547.85	56.70
697	Open-Orca/OpenOrca-Platypus2-13B (Hu et al., 2022; Wang et al., 2023a; Lee et al., 2023; Mukherjee et al., 2023; Touvron et al., 2023b; Longpre et al., 2023; Wang et al., 2023b; Lee et al., 2024)	llama	13B	2023-08-11	6853	-557.72	57.28
698	Open-Orca/OpenOrca-Preview1-13B (Mukherjee et al., 2023; Longpre et al., 2023; Lian et al., 2023a; Touvron et al., 2023a)	llama-1	13B	2023-07-12	1240	-679.87	51.38
699	Orenguteng/Llama-3-8B-Lexi-Uncensored	llama-3	8B	2024-04-23	9653	-564.93	66.18
700	pankajmathur/orca_mini_v3_13b (Mukherjee et al., 2023; Touvron et al., 2023b; Mathur, 2023b)	llama	13B	2023-08-09	4717	-546.48	57.24
701	pankajmathur/orca_mini_v3_7b (Mukherjee et al., 2023; Touvron et al., 2023b; Mathur, 2023c; Touvron et al., 2023a)	llama	7B	2023-08-07	2338	-572.18	53.47
702	PathFinderKR/Waktaverse-Llama-3-KO-8B-Instruct (AI@Meta, 2024; AI@Waktaverse, 2024)	llama-3	8B	2024-04-19	2305	-552.92	66.77
703	pe-nlp/llama-2-13b-vicuna-wizard	llama-2	13B	2023-08-11	1176	-549.39	51.94
704	pillowtalks-ai/delta13b	llama-1	13B	2023-04-14	1168	-646.61	53.29
705	Pirr/pythia-13b-deduped-green_devil	gpt_neox	13B	2023-02-09	1351	-505.96	40.31
706	PistachioAlt/Synatra-MCS-7B-v0.3-RP-Slerp	mistral	7B	2023-12-11	1162	-549.18	69.18
707	PracticeLLM/SOLAR-tail-10.7B-Merge-v1.0	llama	10B	2023-12-26	2264	-530.70	71.68
708	princeton-nlp/Sheared-LLaMA-1.3B (Xia et al., 2024)	llama	1B	2023-10-10	28905	-646.81	35.95
709	princeton-nlp/Sheared-LLaMA-1.3B-ShareGPT (Xia et al., 2024)	llama	1B	2023-11-22	1737	-721.70	37.14
710	princeton-nlp/Sheared-LLaMA-2.7B (Xia et al., 2024)	llama-2	2B	2023-10-10	2590	-610.74	40.84
711	princeton-nlp/Sheared-LLaMA-2.7B-ShareGPT (Xia et al., 2024)	llama-2	2B	2023-11-22	1866	-681.47	42.11
712	project-baize/baize-v2-13b (Xu et al., 2023b)	llama-1	13B	2023-05-23	2149	-578.69	52.94
713	project-baize/baize-v2-7b (Xu et al., 2023b)	llama-1	7B	2023-05-23	1212	-638.05	46.72
714	PygmalionAI/metharme-1.3b	gpt_neox	1B	2023-06-02	1235	-541.84	35.04
715	PygmalionAI/mythalion-13b	llama-2	13B	2023-09-05	2383	-552.46	56.48
716	PygmalionAI/pygmalion-1.3b	gpt_neox	1B	2022-12-25	1519	-963.99	31.14
717	PygmalionAI/pygmalion-2-13b	llama-2	13B	2023-09-04	2083	-540.27	55.12
718	PygmalionAI/pygmalion-2-7b	llama-2	6B	2023-09-04	2063	-559.12	51.11
719	PygmalionAI/pygmalion-2.7b	gpt_neo	2B	2023-01-05	1986	-659.52	33.98
720	PygmalionAI/pygmalion-6b	gptj	6B	2023-01-07	4312	-555.36	38.47
721	qnguyen3/Master-Yi-9B	llama	8B	2024-05-18	8810	-522.22	67.44
722	quantumaikr/llama-2-7b-hf-guanaco-1k	llama-2	7B	2023-08-06	1186	-605.93	50.13
723	quantumaikr/QuantumLM-7B	llama	7B	2023-07-22	1192	-625.18	49.51
724	quantumaikr/quantum-v0.01	mistral	7B	2023-12-17	1192	-559.84	74.68
725	Qwen/CodeQwen1.5-7B-Chat (Bai et al., 2023)	qwen2	7B	2024-04-15	77169	-691.05	43.26
726	Qwen/Qwen1.5-1.8B (Bai et al., 2023)	qwen2	1B	2024-01-22	137256	-610.58	46.55
727	Qwen/Qwen1.5-1.8B-Chat (Bai et al., 2023)	qwen2	1B	2024-01-30	10856	-673.92	43.99
728	Qwen/Qwen1.5-4B (Bai et al., 2023)	qwen2	3B	2024-01-22	6431	-553.86	57.05
729	Qwen/Qwen1.5-4B-Chat (Bai et al., 2023)	qwen2	3B	2024-01-30	5525	-601.85	46.79
730	Qwen/Qwen1.5-7B (Bai et al., 2023)	qwen2	7B	2024-01-22	122768	-533.48	61.76
731	Qwen/Qwen1.5-7B-Chat (Bai et al., 2023)	qwen2	7B	2024-01-30	26143	-608.59	55.15
732	Qwen/Qwen2-1.5B (Yang et al., 2024a)	qwen2	1B	2024-05-31	46510	-581.07	55.80
733	Qwen/Qwen2-7B (Yang et al., 2024a)	qwen2	7B	2024-06-04	37173	-507.30	68.40
734	Q-bert/Bumblebee-7B	mistral	7B	2023-12-03	1226	-549.82	67.73
735	Q-bert/Optimus-7B	mistral	7B	2023-12-03	1237	-550.18	69.09
736	Q-bert/Terminis-7B	mistral	7B	2023-12-12	1233	-564.26	70.73
737	refuelai/Llama-3-Refueled	llama-3	8B	2024-05-03	1246	-582.52	63.62
738	revolutionarybukhari/Llama-2-7b-chat-finetune-AUTOMATE	llama-2	7B	2023-10-14	1166	-613.72	50.68
739	rinna/bilingual-gpt-neox-4b (Zhao et al., 2023b; Sawada et al., 2024)	gpt_neox	3B	2023-07-31	3267	-594.93	32.14
740	rinna/japanese-gpt-neox-3.6b (Zhao and Sawada, 2023; Sawada et al., 2024)	gpt_neox	3B	2023-05-17	6042	-1117.39	29.28
741	rinna/llama-3-youko-8b (Andonian et al., 2021; AI@Meta, 2024; Sawada et al., 2024; Mitsuda et al., 2024)	llama-3	8B	2024-05-01	1468	-502.57	57.55
742	rinna/youri-7b (Andonian et al., 2021; Zhao et al., 2023a; Touvron et al., 2023b; Sawada et al., 2024)	llama-2	7B	2023-10-30	2512	-545.03	47.11
743	rishiraj/CatPPT-base (Acharya, 2023)	mistral	7B	2023-12-17	4392	-558.36	72.25
744	RoversX/llama-2-7b-hf-small-shards-Samantha-V1-SFT	llama-2	7B	2023-08-11	1164	-558.46	49.96
745	RubielLabarta/LogoS-7Bx2-MoE-13B-v0.2	mixtral	12B	2024-01-21	3055	-589.76	77.15

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
746	ruslanmv/ai-medical-model-32bit	llama	8B	2024-05-13	2810	-557.43	67.67
747	ruslanmv/Medical-Llama3-8B	llama-3	8B	2024-04-21	5457	-514.00	60.61
748	rwitz2/go-bruins-v2.1 (Murias, 2023)	mistral	7B	2023-12-14	1193	-561.68	74.50
749	rwitz2/go-bruins-v2.1.1 (Murias, 2023)	mistral	7B	2023-12-14	1205	-562.22	74.95
750	RWKV/rwkv-raven-1b5	rwkv	1B	2023-05-04	1570	-526.30	33.56
751	saberai/Zro1.5_3B	gpt_neox	2B	2023-12-25	1200	-695.87	38.02
752	Salesforce/codegen-6B-multi (Nijkamp et al., 2023)	codegen	6B	2022-04-13	1651	-733.14	32.43
753	Salesforce/codegen-6B-nl (Nijkamp et al., 2023)	codegen	6B	2022-04-13	1176	-478.78	40.00
754	samir-fama/FernandoGPT-v1	mistral	7B	2023-12-30	1209	-551.17	72.87
755	samir-fama/SamirGPT-v1	mistral	7B	2023-12-28	1215	-552.06	73.11
756	SanjiWatsuki/Kunoichi-DPO-v2-7B	mistral	7B	2024-01-13	1216	-586.90	72.40
757	SanjiWatsuki/Kunoichi-7B	mistral	7B	2024-01-04	1230	-579.35	72.13
758	SanjiWatsuki/Loyal-Macaroni-Maid-7B	mistral	7B	2023-12-24	1297	-575.31	71.68
759	SanjiWatsuki/Silicon-Maid-7B	mistral	7B	2023-12-27	1578	-580.90	70.31
760	SanjiWatsuki/Sonya-7B	mistral	7B	2023-12-31	5399	-594.55	68.48
761	sarvamai/OpenHathi-7B-Hi-v0.1-Base	llama-2	6B	2023-12-13	1766	-673.13	46.64
762	scaledown/ScaleDown-7B-slerp-v0.1	mistral	7B	2024-01-01	1206	-538.58	71.57
763	scb10x/llama-3-typhoon-v1.5-8b-instruct (Pipatanakul et al., 2023)	llama-3	8B	2024-05-06	6088	-590.46	65.62
764	scb10x/typhoon-7b (Pipatanakul et al., 2023)	mistral	7B	2023-12-20	1908	-617.60	58.05
765	SciPhi/SciPhi-Self-RAG-Mistral-7B-32k (Mukherjee et al., 2023; Longpre et al., 2023; Asai et al., 2023)	mistral	7B	2023-10-27	1214	-664.82	56.46
766	SeaLLMs/SeaLLM-7B-v2 (Zhang et al., 2023; Zheng et al., 2023; Kojima et al., 2023; Nguyen et al., 2024)	mistral	7B	2024-01-29	6294	-548.56	67.57
767	SeaLLMs/SeaLLM-7B-v2.5 (Zhang et al., 2023; Nguyen et al., 2024)	gemma	8B	2024-04-03	13928	-565.35	69.07
768	selfrag/selfrag_llama2_7b (Asai et al., 2023)	llama-2	7B	2023-10-18	4388	-605.61	51.30
769	senseable/WestLake-7B-v2	mistral	7B	2024-01-22	1189	-594.85	74.68
770	sethuiyer/Medichat-Llama3-8B	llama-3	8B	2024-04-22	4542	-535.77	66.03
771	Severian/ANIMA-Phi-Neptune-Mistral-7B	mistral	7B	2023-10-11	1191	-631.73	55.54
772	shadowml/BeagSake-7B	mistral	7B	2024-01-31	11842	-562.99	75.38
773	shanchen/llama3-8B-slerp-biomed-chat-chinese	llama-3	8B	2024-04-30	2708	-591.11	63.00
774	shanchen/llama3-8B-slerp-med-chinese	llama-3	8B	2024-04-30	8028	-593.26	58.99
775	shanchen/llama3-8B-slerp-med-262k	llama-3	8B	2024-04-30	2697	-599.35	53.65
776	shenzhi-wang/Llama3-8B-Chinese-Chat (Wang et al., 2024b)	llama-3	8B	2024-04-21	55718	-550.92	67.10
777	shibing624/chinese-alpaca-plus-7b-hf (Ming, 2023)	llama-1	7B	2023-05-01	1527	-741.70	44.77
778	shitshow123/tinylamma-20000	llama	1B	2024-01-09	1195	-1213.36	27.95
779	SJ-Donald/llama3-passthrough-chat	llama-3	11B	2024-05-17	2241	-607.16	60.15
780	SJ-Donald/SJ-SOLAR-10.7b-DPO	llama	10B	2024-01-25	2306	-533.83	72.67
781	SJ-Donald/SOLAR-10.7B-slerp	llama	10B	2024-01-12	2296	-533.67	72.58
782	skfrost19/BioMistralMerged	mistral	7B	2024-04-21	4013	-600.14	57.44
783	speakleash/Bielik-7B-Instruct-v0.1 (Levine et al., 2020; Granziol et al., 2021; Ociepa et al., 2024c; Wang et al., 2024a; Ociepa et al., 2024b)	mistral	7B	2024-03-30	3573	-871.06	51.24
784	speakleash/Bielik-7B-v0.1 (Ociepa et al., 2024a,c)	mistral	7B	2024-03-30	2822	-721.18	50.01
785	stabilityai/japanese-stablelm-base-gamma-7b (Jiang et al., 2023)	mistral	7B	2023-10-16	2072	-581.58	52.59
786	stabilityai/japanese-stablelm-instruct-gamma-7b (Jiang et al., 2023)	mistral	7B	2023-10-16	1412	-584.50	52.82
787	stabilityai/StableBeluga-13B (Mukherjee et al., 2023; Touvron et al., 2023b)	llama	13B	2023-07-27	6154	-540.85	57.05
788	stabilityai/StableBeluga-7B (Mukherjee et al., 2023; Touvron et al., 2023b)	llama	6B	2023-07-27	6691	-565.19	53.56
789	stabilityai/stablelm-base-alpha-3b (Andonian et al., 2021)	gpt_neox	3B	2023-04-17	1728	-677.47	31.50
790	stabilityai/stablelm-base-alpha-7b (Andonian et al., 2021)	gpt_neox	7B	2023-04-11	1750	-623.09	34.37
791	stabilityai/stablelm-base-alpha-7b-v2 (Gao et al., 2020; Shazeer, 2020; Rajbhandari et al., 2020; Su et al., 2023a; Li et al., 2023a; Tow, 2023)	stablelm_alpha	6B	2023-08-04	2209	-505.45	46.18
792	stabilityai/stablelm-tuned-alpha-3b (Taori et al., 2023; Chiang et al., 2023; Anand et al., 2023)	gpt_neox	3B	2023-04-19	2145	-736.44	32.14
793	stabilityai/stablelm-tuned-alpha-7b (Taori et al., 2023; Chiang et al., 2023; Anand et al., 2023)	gpt_neox	7B	2023-04-19	3765	-694.37	34.04
794	stabilityai/stablelm-zephyr-3b (Zheng et al., 2023; Rafailov et al., 2024)	stablelm	2B	2023-11-21	8261	-777.98	53.43
795	stabilityai/stablelm-2-zephyr-1_6b (Stability AI Language Team, 2024; Rafailov et al., 2024)	stablelm	1B	2024-01-19	18239	-660.56	49.99
796	stabilityai/stablelm-2-12b-chat (Stability AI Language Team, 2024; Rafailov et al., 2024)	stablelm	12B	2024-04-04	4843	-588.55	68.38
797	stabilityai/stablelm-2-1_6b-chat (Stability AI Language Team, 2024; Rafailov et al., 2024)	stablelm	1B	2024-04-08	4156	-683.91	50.71
798	stabilityai/stablelm-3b-4e1t (Ba et al., 2016; Zhang and Sennrich, 2019; Gao et al., 2020; Rajbhandari et al., 2020; Black et al., 2022; Su et al., 2023a; Tow et al., 2023; Li et al., 2023a; Touvron et al., 2023b)	stablelm	2B	2023-09-29	11112	-510.56	46.58
799	stabilityai/stable-code-3b (Rajbhandari et al., 2020; Black et al., 2022; Su et al., 2023a; Li et al., 2023a; Touvron et al., 2023b; Yu et al., 2024b; Azerbayev et al., 2024; Pinnaparaju et al., 2024)	stablelm	2B	2024-01-09	5670	-616.25	41.53
800	starmpcc/Asclepius-Llama2-13B (Kweon et al., 2024)	llama-2	13B	2023-09-19	1175	-645.45	50.25
801	starmpcc/Asclepius-Llama2-7B (Kweon et al., 2024)	llama-2	7B	2023-09-19	1231	-688.58	47.15
802	statking/zephyr-7b-sft-full-orpo	mistral	7B	2024-05-18	2278	-548.89	53.16
803	StudentLLM/Alpagasus-2-13b-QLoRA-merged (Chen et al., 2024a)	llama	13B	2023-09-02	1303	-533.28	54.31
804	SuperAGI/SAM	mistral	7B	2023-12-22	1235	-550.64	59.30
805	swap-uniba/LLaMAntino-3-ANITA-8B-Inst-DPO-ITA (Basile et al., 2023; AI@Meta, 2024; Polignano et al., 2024)	llama-3	8B	2024-04-29	5995	-635.92	75.12
806	S4sch/zephyr-neural-chat-frankenmerge11b	mistral	11B	2023-11-28	1181	-618.27	58.57
807	TaylorAI/Flash-Llama-13B	llama	13B	2023-08-19	1181	-530.82	53.67
808	TaylorAI/Flash-Llama-3B	llama	3B	2023-08-13	1177	-578.13	40.13
809	TaylorAI/Flash-Llama-7B	llama	7B	2023-08-19	1184	-549.87	49.73
810	TeeZee/Bielik-SOLAR-LIKE-10.7B-Instruct-v0.1	mistral	10B	2024-04-10	1566	-854.79	53.50
811	TehVenom/Dolly_Malion-6b	gptj	6B	2023-03-27	1201	-486.77	39.77
812	TehVenom/Dolly_Shygmalion-6b	gptj	6B	2023-03-29	1195	-491.76	39.89
813	TehVenom/Dolly_Shygmalion-6b-Dev_V8P2 (Wang and Komatsuzaki, 2021)	gptj	6B	2023-05-23	1197	-487.65	40.11
814	TehVenom/GPT-J-Pyg_PPO-6B	gptj	6B	2023-03-05	1203	-495.88	39.60
815	TehVenom/GPT-J-Pyg_PPO-6B-Dev-V8p4	gptj	6B	2023-03-26	1191	-493.17	39.61
816	TehVenom/Metharme-13b-Merged	llama-1	13B	2023-05-18	1199	-561.32	54.15
817	TehVenom/Moderator-Chan_GPT-JT-6b	gptj	6B	2023-03-19	1185	-502.43	42.17
818	TehVenom/PPO_Pygway-V8p4_Dev-6b (Wang and Komatsuzaki, 2021)	gptj	6B	2023-03-17	1191	-489.49	39.85

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
819	TehVenom/PPO_Shygmalion-V8p4_Dev-6b	gptj	6B	2023-03-23	1188	-490.13	39.85
820	TehVenom/PPO_Shygmalion-6b	gptj	6B	2023-03-23	1199	-489.16	39.35
821	TehVenom/Pygmalion-Vicuna-1.1-7b	llama-1	6B	2023-05-02	1246	-572.52	49.25
822	TehVenom/Pygmalion-13b-Merged	llama-1	13B	2023-05-18	1207	-588.35	48.49
823	TehVenom/Pygmalion_AlpacaLora-7b	llama-1	7B	2023-04-30	1195	-605.84	46.49
824	teilomillet/MiniMerlin-3B	llama	3B	2023-12-15	1168	-681.62	47.63
825	teknium/OpenHermes-13B	llama-2	13B	2023-09-06	1594	-543.02	55.24
826	teknium/OpenHermes-2.5-Mistral-7B	mistral	7B	2023-10-29	100997	-560.57	61.45
827	Telugu-LLM-Labs/Indic-gemma-7b-finetuned-sft-Navarasa-2.0	gemma	8B	2024-03-17	2025	-650.29	55.74
828	TencentARC/LLaMA-Pro-8B	llama-2	8B	2024-01-05	1319	-557.58	51.67
829	TencentARC/LLaMA-Pro-8B-Instruct	llama-2	8B	2024-01-06	1423	-634.71	58.06
830	TheBloke/airoboros-13B-HF	llama-1	13B	2023-05-23	1214	-581.17	54.05
831	TheBloke/airoboros-7b-gpt4-fp16	llama-1	7B	2023-06-04	1208	-603.39	47.70
832	TheBloke/CodeLlama-13B-Instruct-fp16	llama-2	13B	2023-08-24	2193	-579.86	45.82
833	TheBloke/CodeLlama-13B-Python-fp16	llama-2	13B	2023-08-24	2114	-587.68	37.52
834	TheBloke/gpt4-alpaca-lora-13B-HF	llama-1	13B	2023-04-17	1181	-547.42	53.98
835	TheBloke/guanaco-13B-HF	llama-1	13B	2023-05-25	1222	-588.92	53.54
836	TheBloke/guanaco-7B-HF	llama-1	7B	2023-05-25	1267	-586.15	47.34
837	TheBloke/koala-13B-HF	llama-1	13B	2023-04-07	2501	-594.23	51.16
838	TheBloke/koala-7B-HF	llama-1	7B	2023-04-07	1238	-617.95	44.29
839	TheBloke/Llama-2-13B-fp16	llama-2	13B	2023-07-18	6470	-530.82	53.67
840	TheBloke/Nous-Hermes-13B-SuperHOT-8K-fp16	llama-1	13B	2023-06-26	1225	-633.76	52.18
841	TheBloke/Planner-7B-fp16	llama-1	7B	2023-06-05	1219	-562.04	45.65
842	TheBloke/stable-vicuna-13B-HF (von Werra et al., 2023; Touvron et al., 2023a; Anand et al., 2023; Taori et al., 2023; Chiang et al., 2023)	llama-1	13B	2023-04-28	1337	-590.21	51.64
843	TheBloke/tulu-13B-fp16 (Chaudhary, 2023; Conover et al., 2023; Touvron et al., 2023a; Wang et al., 2023c; Köpf et al., 2023; Longpre et al., 2023; Peng et al., 2023a)	llama-1	13B	2023-06-10	1231	-583.06	53.58
844	TheBloke/tulu-7B-fp16 (Chaudhary, 2023; Conover et al., 2023; Touvron et al., 2023a; Wang et al., 2023c; Köpf et al., 2023; Longpre et al., 2023; Peng et al., 2023a)	llama-1	7B	2023-06-10	4079	-616.51	50.24
845	TheBloke/UltraLM-13B-fp16 (Ding et al., 2023)	llama-1	13B	2023-06-29	1211	-567.38	54.62
846	TheBloke/Vicuna-13B-CoT-fp16 (Lacoste et al., 2019)	llama-1	13B	2023-06-08	1219	-646.61	53.28
847	TheBloke/WizardLM-13B-V1-1-SuperHOT-8K-fp16 (Xu et al., 2023a)	llama-1	13B	2023-07-07	1225	-626.40	53.16
848	TheBloke/wizard-vicuna-13B-HF	llama-1	13B	2023-05-04	1219	-610.98	52.75
849	TheBloke/Wizard-Vicuna-13B-Uncensored-HF	llama-1	13B	2023-05-13	1648	-579.62	54.14
850	TheBloke/Wizard-Vicuna-7B-Uncensored-HF	llama-1	7B	2023-05-18	1976	-607.14	48.27
851	TheTravellingEngineer/llama2-7b-chat-hf-dpo	llama-2	7B	2023-08-14	1201	-659.39	50.38
852	TheTravellingEngineer/llama2-7b-chat-hf-guanaco	llama-2	6B	2023-08-02	1209	-597.96	50.02
853	TheTravellingEngineer/llama2-7b-chat-hf-v2	llama-2	6B	2023-08-08	1208	-549.87	49.73
854	TheTravellingEngineer/llama2-7b-chat-hf-v3	llama-2	6B	2023-08-10	1203	-557.97	48.81
855	TheTravellingEngineer/llama2-7b-chat-hf-v4	llama-2	6B	2023-08-10	1213	-549.87	49.78
856	TheTravellingEngineer/llama2-7b-hf-guanaco	llama-2	6B	2023-07-25	1206	-554.41	50.12
857	TigerResearch/tigerbot-7b-base	llama	7B	2023-08-19	1228	-587.17	47.93
858	TIGER-Lab/MAmmoTH2-7B-Plus (Yue et al., 2024b)	mistral	7B	2024-05-06	10155	-673.56	67.75
859	TIGER-Lab/MAmmoTH2-8B-Plus (Yue et al., 2024b)	llama	8B	2024-05-06	12819	-595.71	67.49
860	TIGER-Lab/TIGERScore-13B (Jiang et al., 2024)	llama	13B	2023-11-26	1429	-548.56	56.79
861	tiiuae/falcon-rw-1b (Brown et al., 2020; Dao et al., 2022; Press et al., 2022; Penedo et al., 2023)	falcon	1B	2023-04-26	22921	-688.36	37.07
862	tiiuae/falcon-7b (Shazeer, 2019; Brown et al., 2020; Gao et al., 2020; Dao et al., 2022; Su et al., 2023a; Almazrouei et al., 2023; Penedo et al., 2023)	falcon	7B	2023-04-24	104010	-549.44	44.17
863	tiiuae/falcon-7b-instruct (Shazeer, 2019; Brown et al., 2020; Dao et al., 2022; Su et al., 2023a; Almazrouei et al., 2023; Penedo et al., 2023)	falcon	7B	2023-04-25	179952	-621.07	43.16
864	timpal0l/Mistral-7B-v0.1-flashback-v2	mistral	7B	2023-12-04	1275	-578.05	57.53
865	TinyLlama/TinyLlama-1.1B-Chat-v0.6	llama	1B	2023-11-20	15745	-642.54	34.94
866	TinyLlama/TinyLlama-1.1B-Chat-v1.0	llama	1B	2023-12-30	1078500	-619.71	37.17
867	TinyLlama/TinyLlama-1.1B-intermediate-step-1195k-token-2.5T	llama	1B	2023-12-11	1365	-613.87	36.26
868	TinyLlama/TinyLlama-1.1B-intermediate-step-1431k-3T	llama	1B	2023-12-28	798206	-608.46	36.42
869	TinyLlama/TinyLlama-1.1B-intermediate-step-955k-token-2T	llama	1B	2023-11-19	8119	-646.34	34.56
870	togethercomputer/GPT-JT-6B-v0	gptj	6B	2022-11-22	1422	-484.70	44.05
871	togethercomputer/GPT-JT-6B-v1 (Tay et al., 2022, 2023)	gptj	6B	2022-11-24	5765	-503.02	43.13
872	togethercomputer/LLaMA-2-7B-32K	llama-2	7B	2023-07-26	8884	-530.20	47.07
873	togethercomputer/Llama-2-7B-32K-Instruct (Liu et al., 2023b)	llama-2	7B	2023-08-08	5596	-564.83	50.02
874	togethercomputer/Pythia-Chat-Base-7B	gpt_neox	7B	2023-03-22	7040	-522.45	39.81
875	togethercomputer/RedPajama-INCITE-Base-3B-v1	gpt_neox	3B	2023-05-04	2635	-590.97	38.54
876	togethercomputer/RedPajama-INCITE-Chat-3B-v1	gpt_neox	3B	2023-05-05	1647	-610.76	39.53
877	togethercomputer/RedPajama-INCITE-Instruct-3B-v1	gpt_neox	3B	2023-05-05	2014	-552.27	39.06
878	togethercomputer/RedPajama-INCITE-7B-Base	gpt_neox	7B	2023-05-04	1487	-560.75	41.49
879	togethercomputer/RedPajama-INCITE-7B-Chat	gpt_neox	7B	2023-05-04	1583	-931.82	39.37
880	togethercomputer/RedPajama-INCITE-7B-Instruct	gpt_neox	7B	2023-05-05	1274	-525.00	42.38
881	TomGrc/FusionNet	llama	10B	2023-12-31	1166	-561.26	74.38
882	TomGrc/FusionNet_linear	llama	10B	2023-12-31	1167	-561.26	74.43
883	totally-not-an-llm/EverythingLM-13b-16k	llama-2	13B	2023-08-12	2132	-557.79	52.33
884	totally-not-an-llm/PuddleJumper-13b-V2	llama	13B	2023-09-21	1162	-633.45	54.19
885	Toten5/Marcoroni-neural-chat-7B-v2	mistral	7B	2023-12-12	1201	-557.34	72.50
886	TsinghuaC3I/Llama-3-8B-UltraMedical (Zhang et al., 2024c)	llama-3	8B	2024-04-27	3949	-558.79	63.73
887	Unbabel/TowerBase-7B-v0.1 (Alves et al., 2024)	llama	6B	2024-01-03	1856	-582.06	49.11
888	Unbabel/TowerInstruct-7B-v0.1 (Alves et al., 2024)	llama	6B	2024-01-04	2230	-590.30	52.39
889	Undi95/Meta-Llama-3-8B-hf (AI@Meta, 2024)	llama-3	8B	2024-04-18	11376	-514.33	62.35
890	Undi95/Mistral-11B-v0.1	mistral	10B	2023-10-09	1196	-556.96	58.05
891	Undi95/MLewdBoros-L2-13B	llama	13B	2023-09-09	1224	-542.84	56.51
892	Undi95/MLewd-Chat-v2-13B	llama	13B	2023-09-26	1187	-569.88	57.23
893	Undi95/MLewd-L2-Chat-13B	llama	13B	2023-09-16	1177	-551.09	57.75
894	Undi95/MLewd-L2-13B	llama	13B	2023-09-04	1172	-666.07	53.12
895	Undi95/MLewd-v2.4-13B	llama	13B	2023-09-26	1280	-571.03	56.37
896	Undi95/Nous-Hermes-13B-Code	llama	13B	2023-09-02	1207	-602.47	55.93

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
897	Undi95/OpenRP-13B	llama	13B	2023-09-11	1242	-536.49	56.57
898	Undi95/ReMM-SLERP-L2-13B	llama	13B	2023-09-04	1511	-585.33	56.03
899	Undi95/ReMM-v2-L2-13B	llama	13B	2023-09-09	1207	-565.06	56.99
900	Undi95/ReMM-v2.1-L2-13B	llama	13B	2023-09-12	1217	-564.83	56.71
901	Undi95/ReMM-v2.2-L2-13B	llama	13B	2023-09-21	1384	-567.22	57.10
902	Undi95/UndiMix-v1-13b	llama	13B	2023-08-31	1220	-649.19	55.50
903	Undi95/UndiMix-v4-13B	llama	13B	2023-09-12	1212	-569.06	56.93
904	Undi95/Unholy-v1-12L-13B	llama	13B	2023-09-10	1210	-541.04	57.47
905	Undi95/X-MythoChronos-13B	llama	13B	2023-11-18	1203	-604.69	58.43
906	unsloth/mistral-7b-v0.2	mistral	7B	2024-03-24	4119	-532.63	60.34
907	unsloth/Phi-3-medium-4k-instruct	mistral	13B	2024-05-23	3994	-552.43	73.57
908	unsloth/Phi-3-mini-4k-instruct	mistral	3B	2024-04-29	12224	-575.74	69.86
909	unsloth/tinyllama-chat	llama	1B	2024-02-14	5959	-619.71	37.24
910	upstage/SOLAR-10.7B-Instruct-v1.0 (Kim et al., 2024b,a)	llama	10B	2023-12-12	67725	-557.67	74.20
911	upstage/SOLAR-10.7B-v1.0 (Kim et al., 2024b)	llama	10B	2023-12-12	24478	-525.51	66.04
912	uukuguy/speechless-code-mistral-7b-v1.0	mistral	7B	2023-10-10	4393	-540.50	58.85
913	uukuguy/speechless-llama2-hermes-orca-platypus-wizardlm-13b (Touvron et al., 2023b)	llama-2	13B	2023-09-01	2101	-609.55	57.52
914	uukuguy/speechless-llama2-hermes-orca-platypus-13b (Touvron et al., 2023b)	llama-2	13B	2023-09-01	1368	-587.97	57.17
915	uukuguy/speechless-zephyr-code-functionary-7b	mistral	7B	2024-01-23	4098	-529.18	62.93
916	uukuguy/zephyr-7b-alpha-dare-0.85	mistral	7B	2023-11-23	6141	-529.44	62.35
917	uygarkurt/llama-3-merged-linear	llama-3	8B	2024-05-09	12676	-570.63	73.93
918	VAGOsolutions/Llama-3-SauerkrautLM-8b-Instruct	llama-3	8B	2024-04-19	55400	-579.40	73.74
919	VAGOsolutions/SauerkrautLM-Gemma-7b	gemma	8B	2024-02-27	5691	-561.27	67.83
920	VAGOsolutions/SauerkrautLM-SOLAR-Instruct	llama	10B	2023-12-20	1175	-560.76	74.21
921	VAGOsolutions/SauerkrautLM-7b-HerO	mistral	7B	2023-11-24	1226	-553.88	64.49
922	varox34/Bio-Saul-Dolphin-Beagle-Breadcrumbs	mistral	7B	2024-05-01	2679	-610.01	48.72
923	vibhorag101/llama-2-13b-chat-hf-phr_mental_therapy	llama-2	13B	2023-09-17	1269	-630.34	42.50
924	vicgalle/CarbonBeagle-11B (Wortsman et al., 2022)	mistral	10B	2024-01-21	6696	-549.83	74.64
925	vicgalle/CarbonBeagle-11B-truthy	mistral	10B	2024-02-10	13887	-554.84	76.10
926	vicgalle/ConfigurableBeagle-11B (Gallego, 2024)	mistral	10B	2024-02-17	6045	-548.90	75.40
927	vicgalle/ConfigurableHermes-7B (Gallego, 2024)	mistral	7B	2024-02-17	6085	-572.68	68.89
928	vicgalle/ConfigurableSOLAR-10.7B (Gallego, 2024)	llama	10B	2024-03-10	5135	-559.06	73.94
929	vicgalle/Configurable-Hermes-2-Pro-Llama-3-8B (Gallego, 2024)	llama-3	8B	2024-05-02	10273	-580.31	70.10
930	vicgalle/Configurable-Llama-3-8B-v0.3 (Gallego, 2024)	llama-3	8B	2024-04-20	6030	-555.07	68.79
931	vicgalle/Configurable-Yi-1.5-9B-Chat (Gallego, 2024)	llama	8B	2024-05-12	6537	-633.71	70.50
932	viethq188/LeoScorpius-7B	mistral	7B	2023-12-12	1194	-552.34	72.21
933	viethq188/Rabbit-7B-v2-DPO-Chat	mistral	7B	2023-12-12	1181	-579.23	69.36
934	vihangd/dopeyplats-1.1b-2T-v1	llama	1B	2023-11-26	1191	-682.92	35.28
935	vihangd/dopeyshearedplats-1.3b-v1	llama-2	1B	2023-12-12	1168	-766.02	36.74
936	vihangd/dopeyshearedplats-2.7b-v1	llama-2	2B	2023-12-16	1173	-709.66	42.90
937	vihangd/neuralfalcon-1b-v1	falcon	1B	2023-12-17	1185	-895.18	29.72
938	vihangd/shearedplats-1.3b-v1	llama-2	1B	2023-11-16	1184	-706.16	35.97
939	vihangd/shearedplats-2.7b-v2	llama-2	2B	2023-11-18	2081	-647.66	41.61
940	vihangd/smartyplats-3b-v1	llama	3B	2023-09-11	1180	-600.04	40.00
941	vihangd/smartyplats-3b-v2	llama	3B	2023-09-14	1179	-597.07	40.29
942	vikash06/llama-2-7b-small-model-new	llama-2	6B	2023-12-22	1174	-925.37	46.62
943	vmajor/Orca2-13B-selfmerge-26B	llama	13B	2023-12-01	2041	-653.10	62.24
944	vmajor/Orca2-13B-selfmerge-39B	llama	13B	2023-12-01	1208	-653.10	62.24
945	VMware/open-llama-0.7T-7B-open-instruct-v1.1	llama-1	7B	2023-05-31	1161	-652.26	41.11
946	VMware/open-llama-7b-open-instruct	llama-1	7B	2023-06-08	6470	-642.10	42.59
947	Voicelab/trurl-2-13b-academic	llama-2	13B	2023-09-18	2774	-568.46	53.94
948	Voicelab/trurl-2-7b	llama-2	7B	2023-08-16	2861	-602.32	50.58
949	vonjack/Qwen-LLaMAfied-HFTok-7B-Chat	llama-2	7B	2023-08-09	1189	-839.96	50.64
950	v1olet/v1olet_marcoroni-go-bruins-merge-7B	mistral	7B	2023-12-11	1225	-559.92	72.81
951	v1olet/v1olet_merged_dpo_7B	mistral	7B	2023-12-12	1210	-592.71	70.26
952	wang7776/Llama-2-7b-chat-hf-10-sparsity (Sun et al., 2024)	llama-2	6B	2023-12-11	1178	-660.87	52.48
953	wang7776/Llama-2-7b-chat-hf-20-sparsity (Sun et al., 2024)	llama-2	7B	2023-12-13	1175	-668.34	52.01
954	wang7776/Llama-2-7b-chat-hf-30-sparsity (Sun et al., 2024)	llama-2	6B	2023-12-11	1179	-676.50	51.02
955	webbigdata/ALMA-7B-Ja-V2 (Xu et al., 2024a)	llama-2	7B	2023-10-21	1177	-599.69	47.85
956	wei123602/Llama-2-13b-FINETUNE4_TEST	llama-2	13B	2023-09-18	1174	-538.02	53.62
957	wenbopan/Faro-Yi-9B (OpenAI et al., 2024)	llama	8B	2024-03-27	6127	-594.07	66.37
958	wenbopan/Faro-Yi-9B-DPO (OpenAI et al., 2024)	llama	8B	2024-04-07	6121	-597.40	68.77
959	wenge-research/yayi-7b	bloom	7B	2023-06-02	1192	-653.40	41.88
960	wenge-research/yayi-7b-llama2	llama-2	7B	2023-07-21	1195	-562.60	49.88
961	Weyaxi/ChatAYT-Lora-Assamble-Marcoroni	llama	13B	2023-09-14	1203	-569.45	57.76
962	Weyaxi/Dolphin2.1-OpenOrca-7B	mistral	7B	2023-10-11	1221	-557.64	60.47
963	Weyaxi/Instruct-v0.2-Seraph-7B	mistral	7B	2023-12-12	1203	-574.63	68.48
964	Weyaxi/Luban-Marcoroni-13B-v2	llama	13B	2023-09-13	1212	-566.43	57.92
965	Weyaxi/Luban-Marcoroni-13B-v3	llama	13B	2023-09-13	1214	-566.44	57.94
966	Weyaxi/MetaMath-Chupacabra-7B-v2.01-Slerp	mistral	7B	2023-12-08	1210	-546.54	70.26
967	Weyaxi/MetaMath-NeuralHermes-2.5-Mistral-7B-Linear	mistral	7B	2023-12-05	1205	-560.95	67.60
968	Weyaxi/MetaMath-NeuralHermes-2.5-Mistral-7B-Ties	mistral	7B	2023-12-05	1212	-604.36	67.03
969	Weyaxi/MetaMath-neural-chat-7b-v3-2-Slerp	mistral	7B	2023-12-08	1208	-544.26	69.79
970	Weyaxi/MetaMath-neural-chat-7b-v3-2-Ties	mistral	7B	2023-12-05	1212	-590.97	67.54
971	Weyaxi/MetaMath-OpenHermes-2.5-neural-chat-v3-3-Slerp	mistral	7B	2023-12-10	1224	-548.15	69.92
972	Weyaxi/MetaMath-Tulpar-7b-v2-Slerp	mistral	7B	2023-12-08	1217	-555.63	70.07
973	Weyaxi/MetaMath-una-cybertron-v2-bf16-Ties	mistral	7B	2023-12-06	1218	-602.74	68.88
974	Weyaxi/neural-chat-7b-v3-1-OpenHermes-2.5-7B	mistral	7B	2023-12-01	1205	-570.29	67.19
975	Weyaxi/openchat-3.5-1210-Seraph-Slerp	mistral	7B	2023-12-27	1210	-553.81	71.82
976	Weyaxi/OpenHermes-2.5-neural-chat-7b-v3-1-7B	mistral	7B	2023-11-24	1230	-567.74	67.84
977	Weyaxi/OpenHermes-2.5-neural-chat-7b-v3-2-7B	mistral	7B	2023-12-03	1221	-576.02	68.71
978	Weyaxi/OpenOrca-Zephyr-7B	mistral	7B	2023-10-11	1208	-564.12	64.97
979	Weyaxi/Samantha-Nebula-7B	mistral	7B	2023-10-05	1166	-584.51	54.58
980	Weyaxi/SauerkrautLM-UNA-SOLAR-Instruct	llama	10B	2023-12-21	1225	-561.98	74.26

ID	Model Name	Model Type	Size	Date	DLs	$\bar{\ell}_i$	Task
981	Weyaxi/Seraph-openchat-3.5-1210-Slerp	mistral	7B	2023-12-27	1210	-567.06	70.89
982	Weyaxi/Seraph-7B	mistral	7B	2023-12-11	1214	-546.73	71.86
983	Weyaxi/SlimOpenOrca-Mistral-7B	mistral	7B	2023-10-11	1220	-587.26	60.84
984	WhiteRabbitNeo/WhiteRabbitNeo-13B-v1	llama-2	13B	2023-12-17	2090	-615.73	49.11
985	winglian/Llama-2-3b-hf	llama-2	3B	2023-09-19	1565	-1376.16	29.53
986	winglian/llama-2-4b	llama-2	4B	2023-09-19	1193	-676.31	34.23
987	w601sxs/b1ade-1b	gpt_neox	1B	2023-07-17	1204	-929.29	32.59
988	xDAN-AI/xDAN-L1-Chat-RL-v1	mistral	7B	2023-12-20	1178	-574.92	68.38
989	Xwin-LM/Xwin-LM-13B-V0.1 (Xwin-LM Team, 2023)	llama-2	13B	2023-09-15	1401	-559.13	55.29
990	Xwin-LM/Xwin-LM-7B-V0.1 (Xwin-LM Team, 2023)	llama-2	7B	2023-09-15	1474	-592.93	52.08
991	yam-peleg/Hebrew-Mistral-7B	mistral	7B	2024-04-26	5993	-625.28	58.76
992	yanolja/Bookworm-10.7B-v0.4-DPO (Mukherjee et al., 2023; Lian et al., 2023g; Cui et al., 2024)	llama	10B	2024-01-18	2239	-586.77	66.59
993	yanolja/EEVE-Korean-Instruct-10.8B-v1.0 (Mukherjee et al., 2023; Lian et al., 2023g; Kim et al., 2024c; Cui et al., 2024)	llama	10B	2024-02-22	13241	-555.03	66.48
994	yeen214/llama2_7b_small_tuning_v1	llama-2	7B	2023-10-02	3289	-1402.04	28.56
995	yeen214/test_llama2_ko_7b	llama-2	7B	2023-10-02	3285	-1405.43	29.99
996	yeen214/test_llama2_7b	llama-2	7B	2023-09-30	3289	-549.87	49.73
997	yhyhy3/open_llama_7b_v2_med_instruct	llama-1	7B	2023-07-09	1197	-561.62	46.24
998	Yhyu13/chimera-inst-chat-13b-hf	llama-1	13B	2023-05-11	1295	-582.39	52.86
999	Yhyu13/LMCocktail-10.7B-v1 (Xiao et al., 2023)	llama-2	10B	2023-12-20	3332	-556.38	74.06
1000	Yukang/Llama-2-7b-longlora-32k-ft (Chen et al., 2024b)	llama-2	7B	2023-09-12	2423	-1387.05	29.20
1001	Yukang/LongAlpaca-13B (Chen et al., 2023b, 2024b)	llama	13B	2023-10-08	1896	-832.57	41.74
1002	Yukang/LongAlpaca-7B (Chen et al., 2023b, 2024b)	llama	6B	2023-10-07	2773	-794.46	39.36
1003	yulan-team/YuLan-Chat-2-13b-fp16 (YuLan-Team, 2023)	llama	13B	2023-08-04	1165	-645.56	57.01
1004	yunconglong/DARE_TIES_13B (Yadav et al., 2023a; Yu et al., 2024a)	mixtral	12B	2024-01-30	7029	-611.23	77.10
1005	yunconglong/MoE_13B_DPO	mixtral	12B	2024-01-28	3939	-603.88	77.05
1006	yunconglong/Truthful_DPO_TomGrc_FusionNet_7Bx2_MoE_13B	mixtral	12B	2024-01-21	7965	-598.84	77.44
1007	01-ai/Yi-1.5-6B (01. AI et al., 2025)	llama	6B	2024-05-11	5102	-525.57	61.57
1008	01-ai/Yi-1.5-6B-Chat (01. AI et al., 2025)	llama	6B	2024-05-11	19443	-645.33	66.17
1009	01-ai/Yi-1.5-9B (01. AI et al., 2025)	llama	8B	2024-05-11	20252	-513.39	66.73
1010	01-ai/Yi-1.5-9B-Chat (01. AI et al., 2025)	llama	8B	2024-05-10	20108	-626.90	69.56
1011	01-ai/Yi-1.5-9B-Chat-16K (01. AI et al., 2025)	llama	8B	2024-05-15	14262	-601.39	66.98
1012	01-ai/Yi-1.5-9B-32K (01. AI et al., 2025)	llama	8B	2024-05-15	9287	-510.38	55.22
1013	01-ai/Yi-6B (Zhang et al., 2024b; Yue et al., 2024a; 01. AI et al., 2025)	llama	6B	2023-11-01	6997	-515.62	54.02
1014	01-ai/Yi-6B-200K (Zhang et al., 2024b; Yue et al., 2024a; 01. AI et al., 2025)	llama	6B	2023-11-06	8370	-529.12	56.76
1015	01-ai/Yi-9B-200K (Zhang et al., 2024b; Yue et al., 2024a; 01. AI et al., 2025)	llama	8B	2024-03-15	8915	-517.82	61.94
1016	42dot/42dot_LLM-PLM-1.3B (42dot Inc., 2023)	llama	1B	2023-09-04	1268	-600.03	35.70
1017	42dot/42dot_LLM-SFT-1.3B (42dot Inc., 2023)	llama	1B	2023-09-04	1453	-605.30	36.61
1018	922-CA/monika-ddlc-7b-v1	llama-2	7B	2023-10-13	1191	-635.19	50.49