Self-calibration for Language Model Quantization and Pruning

Miles Williams[◇] **George Chrysostomou** **Nikolaos Aletras**[◇]
[◇]University of Sheffield
United Kingdom
{mwilliams15, n.aletras}@sheffield.ac.uk

Abstract

Quantization and pruning are fundamental approaches for model compression, enabling efficient inference for language models. In a post-training setting, state-of-the-art quantization and pruning methods require calibration data, a small set of unlabeled examples. Conventionally, randomly sampled web text is used, aiming to reflect the model training data. However, this poses two key problems: (1) unrepresentative calibration examples can harm model performance, and (2) organizations increasingly avoid releasing model training data. In this paper, we propose self-calibration as a solution. Our approach requires no external data, instead leveraging the model itself to generate synthetic calibration data as a better approximation of the pre-training data distribution. We extensively compare the performance of self-calibration with several baselines, across a variety of models, compression methods, and tasks. Our approach proves consistently competitive in maximizing downstream task performance, frequently outperforming even using real data.

1 Introduction

Large language models (LLMs) trained using vast corpora have delivered remarkable advances across a variety of domains and tasks (Touvron et al., 2023a; Jiang et al., 2023; Gemma Team et al., 2024). However, they demand extensive computational resources for inference (Wu et al., 2022; Luccioni et al., 2023), presenting a limiting factor in their practical use. Consequently, this has prompted the development of an extensive collection of methods to improve inference efficiency (Treviso et al., 2023). In particular, model compression aims to reduce the size of a model yet retain downstream task performance (Wan et al., 2024).

Quantization and pruning have emerged as prominent model compression approaches for LLMs (Gholami et al., 2021; Wan et al., 2024). Pruning removes less important weights from the model, while quantization represents the weights (and possibly activations) using fewer bits. Both quantization and pruning can be effectively applied in a post-training setting, retaining comparable performance across a range of downstream tasks (Frantar et al., 2023; Frantar and Alistarh, 2023; Sun et al., 2024; Lin et al., 2024).

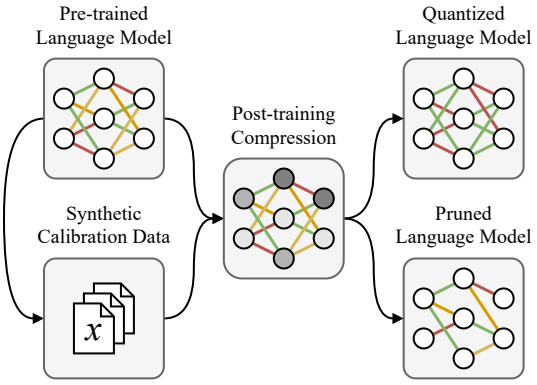

Figure 1: Self-calibration for the post-training quantization and pruning of language models.

Post-training quantization and pruning typically depend upon *calibration data*, a small set of unlabeled examples (Nagel et al., 2020; Hubara et al., 2021) used to generate layer activations throughout the model. Conventionally, LLM calibration data consists of randomly sampled web text (Frantar et al., 2023; Frantar and Alistarh, 2023; Sun et al., 2024; Lin et al., 2024), aiming to reflect the model training data distribution.

However, recent work has raised concerns surrounding the impact of calibration data upon LLM compression. Jaiswal et al. (2024) suggest that the careful selection of calibration examples may play an essential role in high-sparsity pruning. Concurrently, Williams and Aletras (2024) demonstrate the influence of calibration data in post-training quantization and pruning. Finally, Zeng et al. (2024) illustrate the importance of linguistically diverse calibration data for multilingual performance.

To further complicate matters, organizations are increasingly reluctant to release model training data or disclose necessary replication details. Table 1 illustrates that although the weights of some state-of-the-art LLMs are openly available, their training data is largely unavailable. This may be due to (1) legal liability concerns arising from data licensing (Eckart de Castilho et al., 2018), and (2) privacy concerns when using proprietary or personal data (Carlini et al., 2021). Moreover, publicly released training data can later become unavailable. For example, The Pile (Gao et al., 2020), is no longer distributed due to copyright violations. The absence of training data raises the question of how representative calibration data can be selected, when the training distribution itself is unknown. This issue is especially relevant for models trained primarily with private datasets, such as Microsoft's Phi series of models (Gunasekar et al., 2023; Li et al., 2023b).

In this paper, we propose self-calibration as a solution to concerns surrounding the availability and quality of calibration data. Our approach removes the need for external calibration data sources, instead leveraging the model itself to automatically generate synthetic calibration data. We compare our approach to various real and synthetic datasets, including data sampled from a large mixture-of-experts model. Our approach is consistently competitive in maximizing the performance of compressed models, across a variety of models and compression methods. In many cases, we find that self-calibration can outperform even real data.

2 Related Work

2.1 Model Compression

Model compression aims to reduce the size of a model without compromising downstream task performance, therefore reducing the computational resources required for inference (Treviso et al., 2023). Quantization and pruning are two prominent model compression approaches that have been widely applied to LLMs (Wan et al., 2024).

Pruning. The goal of pruning is to remove redundant model weights (LeCun et al., 1989). Pruning often relies upon a fine-tuning step (Han et al., 2015; Sanh et al., 2020), however this is challenging at the scale of LLMs. Alternatively, there have been various efforts towards adapting the Optimal Brain Surgeon (OBS) framework (LeCun et al., 1989; Hassibi et al., 1993) for language model pruning (Frantar et al., 2021; Kurtic et al., 2022; Frantar and Alistarh, 2022). However, the extensive size of LLMs makes it impractical to apply such methods. SparseGPT (Frantar and Alistarh, 2023) presents an approximate weight reconstruction approach, enabling efficient LLM pruning without compromising performance. Separately, Wanda (Sun et al., 2024) relies on a pruning criterion that does not require second-order information, allowing pruning with a single forward-pass.

	Open Source	
Model Family	Weights	Data
OPT (Zhang et al., 2022)	✓	✓
GPT-4 (OpenAI et al., 2023)	✗	✗
Mistral (Jiang et al., 2023)	✓	✗
Llama 2 (Touvron et al., 2023b)	✓	✗
Falcon (Almazrouei et al., 2023)	✓	✓
Phi-2 (Javaheripi et al., 2023)	✓	✗
Gemini (Gemini Team et al., 2024)	✗	✗
OLMo (Groeneveld et al., 2024)	✓	✓
Claude 3 (Anthropic, 2024)	✗	✗
Gemma (Gemma Team et al., 2024)	✓	✗

Table 1: The training data for state-of-the-art LLMs is rarely available. Models selected according to benchmark performance and ordered by publication date.

Quantization. The aim of quantization is to represent model weights (and potentially activations) using fewer bits. Large magnitude outlier features pose a significant problem for the quantization of LLMs, which can be addressed through holding these in higher precision (Dettmers et al., 2022). However, this approach is less hardware-friendly. Instead, SmoothQuant (Xiao et al., 2023) migrates the difficulty of activation quantization to the weights, which are easier to quantize. AWQ (Lin et al., 2024) presents a hardware-friendly approach for holding a small fraction of the weights in higher precision. In a separate line of work, Frantar and Alistarh (2022) adapt the OBS framework to quantization. GPTQ (Frantar et al., 2023) builds upon this work to enable second-order low-bit quantization for LLMs.

2.2 Calibration Data

In a post-training setting, model compression methods rely upon calibration data (Wan et al., 2024). This consists of a small set of unlabeled examples, used to generate layer activations (Nagel et al., 2020; Hubara et al., 2021). Calibration data for LLMs conventionally consists of text sampled from a curated training dataset (Frantar et al., 2023; Xiao et al., 2023; Frantar and Alistarh, 2023; Sun et al.,

2024; Lin et al., 2024). In practice, the exact model training data may not be publicly available (Table 1). Consequently, large scale web text datasets (e.g. C4; Raffel et al., 2020) are ordinarily used as an approximation of the pre-training distribution. Recent work has questioned the performance impact of the calibration data used for LLM compression (Jaiswal et al., 2024; Williams and Aletras, 2024; Zeng et al., 2024). Synthetic data presents a promising avenue towards alleviating such concerns, including the varied quality of web text examples (Dodge et al., 2021). However, synthetic calibration data for post-training LLM compression has yet to be systematically explored.

Synthetic data for model compression has been previously explored in computer vision, regularly motivated by privacy and security concerns arising from sensitive training images (e.g. medical contexts). Haroush et al. (2020) and Cai et al. (2020) proposed approaches for data-free quantization (Nagel et al., 2019), allowing the model itself to synthesize input data for quantization. Fundamentally, these approaches generate images matching the learned statistics from batch normalization layers (Zhang et al., 2021; Li et al., 2023a), which are notably absent in LLMs (Wang et al., 2022).

2.3 Synthetic Data with Language Models

Synthetic data refers to artificial data that has been created with the aim of imitating real-world data (Liu et al., 2024). In the context of language models, supervised training of classification models with synthetic labeled data has been widely explored (Kumar et al., 2020; Schick and Schütze, 2021; Sahu et al., 2022; Meng et al., 2022; Chung et al., 2023; Li et al., 2023c). Similarly, synthetic data has seen broad use for supervised instruction fine-tuning (Wang et al., 2023; Ding et al., 2023; Xu et al., 2024). Most recently, partially or entirely synthetic datasets have been used for pre-training (Gunasekar et al., 2023; Li et al., 2023b; Maini et al., 2024; Ben Allal et al., 2024), although they may still deviate from the pre-training distribution of other LLMs if used for calibration data selection.

3 Self-calibration

When the exact training data for a model is unavailable, sampling calibration data from an alternative distribution offers an approximation at best. Even if the exact training data is available, individual examples may be noisy and deviate from the overall distribution. To address these limitations, we propose self-calibration, a general purpose adaptation to model compression that relies on calibration data from the model itself. Our hypothesis is that sampling from the learned posterior distribution, which approximates the training data, offers more representative calibration examples. In turn, we expect that such calibration examples will enable greater preservation of downstream performance following model compression.

3.1 Synthesizing Calibration Data

We formulate the synthesis of calibration examples as an open-ended text generation problem for a specific language model that we wish to compress. Crucially, we aim to generate synthetic data that is as representative as possible with respect to the training distribution. To achieve this, we refrain from using external data, which introduces assumptions about the training data distribution.

Fundamentally, text generation consists of predicting the next token in a sequence. Formally, we compute a probability distribution over the vocabulary $\mathcal{V}$ for the next token w_i, given context $w_{1:i-1}$. Taking the context as input, a language model generates the output logits, $u_{1:|\mathcal{V}|}$. The probability distribution is then formed by normalizing the logits with the softmax function.

To generate calibration data that reflects the model training data distribution, we condition generation upon only the beginning of sequence token (e.g. <s> or <|start_of_text|>). We continue to generate tokens until either the end of sequence token or maximum sequence length is reached. In the event that a generation does not reach the desired length, we simply concatenate additional generations. As a prefix or prompt would introduce bias and require external data, we do not directly condition generation. Instead, we rely upon scheduled temperature sampling to guide generation.

3.2 Temperature Scheduling

The softmax function can be additionally parameterized with a temperature t, to control the sharpness of the probability distribution (Ackley et al., 1985; Hinton et al., 2015). A lower temperature concentrates the probability mass on more likely tokens, while a higher temperature disperses the probability mass more uniformly. In practice, the temperature influences characteristics of the generated text, often improving its quality and diversity compared to greedy decoding (Holtzman et al.,

2020; Meister et al., 2023).

When generating text without context, we hypothesize that the first few generated tokens are crucial, influencing the content and coherence. To explore a variety of prefixes, we propose the use of a temperature schedule, inspired by Carlini et al. (2021). Formally, we define the probability of a token as:

$$P(w_i \mid w_{1:i-1}) = \frac{\exp(u_i/t_i)}{\sum_{j=1}^{|\mathcal{V}|} \exp(u_j/t_i)}$$

where t_i scales linearly from t_{initial} at the start of generation to t_{final}, across n token generation steps:

$$t_i = \begin{cases} t_{\text{initial}} + \frac{i}{n}(t_{\text{final}} - t_{\text{initial}}) & \text{if } i \leq n, \\ t_{\text{final}} & \text{if } i > n. \end{cases}$$

In practice, a temperature schedule enables us to experiment with a variety of generation strategies. For example, we are able to generate a diverse prefix (i.e. $t_{\text{initial}} > 1$) followed by a more confident continuation (i.e. $t_{\text{final}} \leq 1$), as well as a high-likelihood prefix followed by a creative continuation. We provide a comprehensive ablation of these parameters choices in §6.2. For comparison, we also present results with greedy decoding and standard sampling (i.e. without temperature).

4 Experimental Setup

4.1 Baseline Calibration Data

Real data. To evaluate the performance of self-calibration for LLM compression, we first consider real-world datasets that are conventionally used for LLM compression (Frantar et al., 2023).

- **C4** (Raffel et al., 2020): The Colossal Clean Crawled Corpus is routinely used as a source of calibration data (§2.2). This consists of web-text that has been deduplicated and filtered to maximize high-quality natural language text.

- **WikiText** (Merity et al., 2017): The WikiText dataset consists of a high quality encyclopedic text from Wikipedia. Notably, this includes only articles highlighted as 'Good' or 'Featured' by human editors. The review process assesses accuracy and writing quality, amongst other factors.

Synthetic data. Separately, we compare the performance of self-calibration with synthetic data generated (1) without a language model, and (2) with a substantially larger external model.

- **Vocabulary**: As a simple baseline, we create examples consisting of tokens randomly sampled from the model vocabulary. We assume a uniform distribution over the vocabulary, however we exclude special purpose tokens (e.g. <unk>).

- **Cosmopedia** (Ben Allal et al., 2024): The Cosmopedia dataset consists of a broad range of synthetic text, including textbooks, blog posts, and stories. These were created by prompting Mixtral 8x7B Instruct (Jiang et al., 2024) with a variety of high-quality topics selected from real data.

Sampling. Following convention, we randomly sample 128 calibration examples consisting of 2,048 tokens each (Frantar et al., 2023; Frantar and Alistarh, 2023; Sun et al., 2024). Although the aim of random sampling is to avoid selection bias, it could produce a sample that is less representative of the source dataset. Consequently, we repeat the sampling process to create five distinct calibration sets for each source dataset. We present an ablation of the quantity of calibration data used in §6.1.

Certain models (Gemma, Mistral, and Llama) were trained using multilingual and/or code data, which is reflected when sampling from these models. To enable a fair comparison with our English-only calibration datasets and evaluation tasks, we promote the generation of English-language text for these models. Specifically, we constrain only the first generation step to a pre-defined list of English stop words curated by Honnibal et al. (2020).

4.2 Models

We experiment with popular 'open source' LLMs from five different families: (1) **Gemma 2B** (Gemma Team et al., 2024), (2) **Phi-2** (2.7B) (Javaheripi et al., 2023), (3) **OPT 6.7B** (Zhang et al., 2022), (4) **Mistral 7B** v0.3 (Jiang et al., 2023), and (5) **Llama 3.1 8B** (Dubey et al., 2024).[1]

With the exception of OPT, which was pre-trained using only publicly available datasets, limited details surrounding the training data distribution have been disclosed. The training data for all models is reported to include public web documents. However, the training data for Phi-2 notably relies upon a substantial proportion of synthetic data generated with GPT-3.5 (Ouyang et al., 2022).

[1]We note that Gemma Team et al. (2024) use a naming scheme that excludes embedding parameters. For comparison, Gemma 2B has 2.5B trainable parameters. We also note that the embedding parameters are shared (Press and Wolf, 2017).

Method	Calibration Dataset		Gemma 2B	Phi-2 2.7B	OPT 6.7B	Mistral 7B	Llama 3.1 8B
-	-		60.7	65.8	57.4	67.4	67.8
AWQ	Real	C4	$59.5_{0.2}$	$65.4_{0.2}$	$57.6_{0.1}$	$67.1_{0.0}$	$66.9_{0.2}$
		WikiText	$59.5_{0.2}$	$65.4_{0.2}$	$57.5_{0.1}$	$67.1_{0.1}$	$67.1_{0.1}$
	Synthetic	Vocabulary	$59.3_{0.2}$	$64.5_{0.2}$	$56.6_{0.3}$	$66.5_{0.1}$	$66.0_{0.3}$
		Cosmopedia	$59.8_{0.2}$	$65.3_{0.2}$	$57.6_{0.1}$	$67.0_{0.2}$	$66.9_{0.3}$
		Self-calibration (Ours)	**59.8**$_{0.4}$	**65.4**$_{0.2}$	**57.6**$_{0.1}$	$67.0_{0.2}$	$66.6_{0.3}$
GPTQ	Real	C4	$58.7_{0.4}$	$64.7_{0.3}$	$56.8_{0.2}$	$66.8_{0.3}$	$66.9_{0.3}$
		WikiText	$58.6_{0.3}$	$64.6_{0.2}$	$56.9_{0.1}$	$66.9_{0.3}$	$66.6_{0.3}$
	Synthetic	Vocabulary	$57.9_{0.3}$	$64.3_{0.2}$	$56.6_{0.3}$	$66.0_{0.1}$	$65.7_{0.1}$
		Cosmopedia	$58.5_{0.3}$	$64.3_{0.1}$	$56.8_{0.1}$	$66.6_{0.2}$	$66.9_{0.1}$
		Self-calibration (Ours)	**59.9**$_{0.3}$	**65.0**$_{0.3}$	**56.9**$_{0.2}$	$65.9_{0.2}$	$66.1_{0.3}$
SparseGPT	Real	C4	$49.7_{0.8}$	$54.3_{0.3}$	$52.8_{0.2}$	$57.3_{0.3}$	$54.8_{0.3}$
		WikiText	$48.3_{0.2}$	$53.3_{0.5}$	$51.6_{0.2}$	$55.5_{0.3}$	$52.6_{0.4}$
	Synthetic	Vocabulary	$43.4_{0.3}$	$50.1_{0.2}$	$47.7_{0.2}$	$53.0_{0.4}$	$47.3_{0.4}$
		Cosmopedia	$47.7_{0.3}$	$52.3_{0.2}$	$50.9_{0.2}$	$55.1_{0.3}$	$50.9_{0.3}$
		Self-calibration (Ours)	**50.8**$_{0.2}$	**56.4**$_{0.3}$	$52.7_{0.3}$	$56.8_{0.3}$	$53.8_{0.4}$
Wanda	Real	C4	$44.2_{0.2}$	$50.4_{0.4}$	$50.6_{0.2}$	$53.7_{0.3}$	$49.0_{0.3}$
		WikiText	$44.8_{0.4}$	$49.9_{0.2}$	$49.2_{0.2}$	$53.4_{0.2}$	$49.2_{0.1}$
	Synthetic	Vocabulary	$42.1_{0.4}$	$47.0_{0.3}$	$43.2_{0.1}$	$48.4_{0.2}$	$44.7_{0.3}$
		Cosmopedia	$44.5_{0.2}$	$49.4_{0.4}$	$48.7_{0.2}$	$52.7_{0.2}$	$47.7_{0.2}$
		Self-calibration (Ours)	**45.2**$_{0.3}$	**51.5**$_{0.7}$	**50.7**$_{0.2}$	$53.5_{0.1}$	$49.1_{0.1}$

Table 2: Average task accuracy across five calibration sets for all models, with standard deviation denoted in subscript. Highlighted values indicate that self-calibration (ours) matches or exceeds the performance of all synthetic datasets. **Bold** values additionally indicate that self-calibration matches or exceeds the highest performing dataset overall, including the real datasets. Self-calibration temperature is fixed at 1.0 to enable fair comparison.

4.3 Model Compression

As it is not possible to experiment with every existing model compression approach, we select four of the most widely adopted methods. We report the implementation details in Appendix A and complete hyperparameter selection in Appendix C.

Quantization. For quantization, we trial **AWQ** (Lin et al., 2024) and **GPTQ** (Frantar et al., 2023). In both cases, we use 4-bit weight quantization, which sees minimal performance degradation while enabling efficient inference (Frantar et al., 2024).

Pruning. For pruning, we employ **SparseGPT** (Frantar and Alistarh, 2023) and **Wanda** (Sun et al., 2024). In both cases, we focus on the 2:4 semi-structured (50%) sparsity setting, which enables inference speedups on GPUs (Mishra et al., 2021).

4.4 Evaluation Tasks

To offer an impartial selection of evaluation tasks, we adopt all zero-shot tasks used in the original work to evaluate AWQ, GPTQ, SparseGPT, and Wanda. Namely, ARC (easy and challenge sets) (Clark et al., 2018), BoolQ (Clark et al., 2019), HellaSwag (Zellers et al., 2019), LAMBADA (Paperno et al., 2016), OpenBookQA (Banerjee et al., 2019), PIQA (Bisk et al., 2020), RTE (Dagan et al., 2006), StoryCloze (Mostafazadeh et al., 2016), and WinoGrande (Sakaguchi et al., 2021).

5 Results

Table 2 presents the average performance across all downstream tasks (§4.4) for every model tested (§4.2).[2] For self-calibration, we set t_{initial} and t_{final} as 1.0 (i.e. temperature sampling), to enable a fair comparison between models. However, we emphasize that the careful selection of these parameters can offer further performance improvements. We provide a deeper analysis on the impact of the temperature schedule in §6.2.

Self-calibration outperforms other synthetic datasets. We first observe that self-calibration matches or exceeds other synthetic datasets in 17 out 20 instances. For example, when quantizing Gemma 2B with GPTQ, self-calibration records a mean accuracy of 59.9, compared to 58.5 for Cosmopedia and 57.9 with random vocabulary. Similarly, when pruning Llama 3.1 8B with SparseGPT, self-calibration offers a 2.9 point increase in mean accuracy compared to Cosmopedia (53.8 versus 50.9). This suggests that *self-calibration may produce calibration data that are more representative*

[2]For detailed task performance please see Appendix D

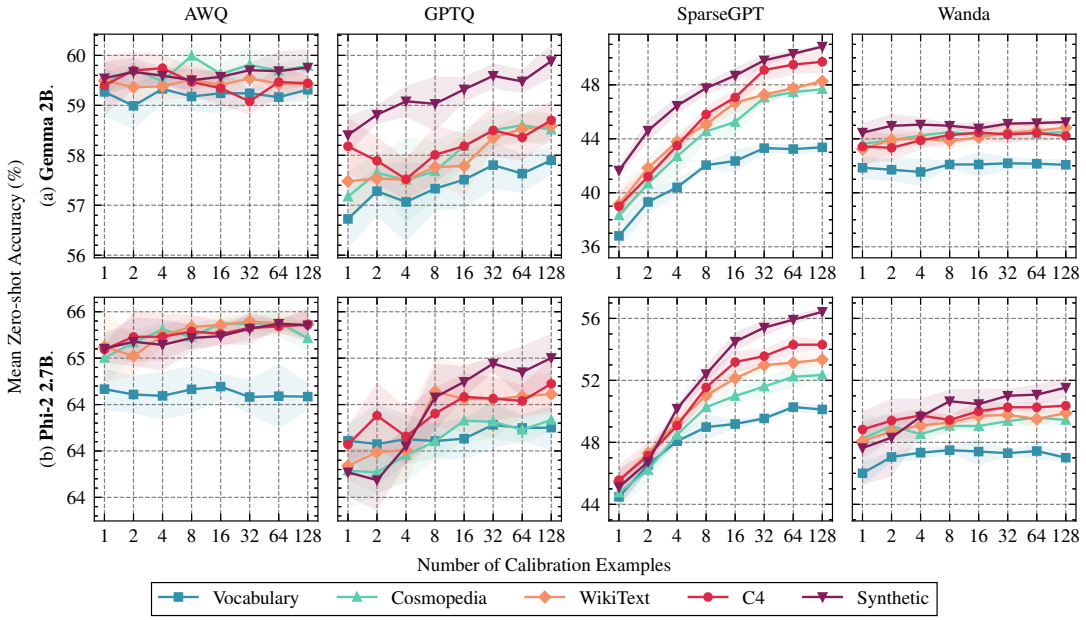

Figure 2: The mean zero-shot accuracy when compressing Gemma 2B and Phi-2 with each method. We present the mean value and standard deviation (shaded) across five distinct calibration sets sampled from each data source.

of the training distribution of each model, compared to using other synthetic datasets.

Self-calibration can outperform real world-data. Our results show that for Phi-2, Gemma 2B, and OPT 6.7B, self-calibration achieves the highest mean accuracy compared to all other datasets in all but one instance. The only exception is when pruning OPT 6.7B with SparseGPT, where self-calibration ranks second to C4 (52.7 with self-calibration versus 52.8 with C4). With Mistral 7B and Llama 3.1 8B, we observe that despite self-calibration not outperforming real data, performance is as competitive with real data as Cosmopedia (i.e. is comparable or outperforms Cosmopedia in five out eight instances). These outcomes suggest that *using self-calibration for model compression, results in downstream performance that is at least comparable to that of real data.*

Pruning benefits the most from self-calibration. Across all models and both pruning methods, self-calibration results in higher mean accuracy compared to other synthetic data. For example, when pruning Llama 3.1 8B with Wanda, self-calibration is only second to WikiText by a 0.1 point difference (49.1 compared to 49.2 with WikiText) whilst also being 1.4 points higher than Cosmopedia.

We also observe that quantization methods are less sensitive to the calibration data, making the dataset choice less critical. For example, the difference between the best and worst performing calibration data source for Gemma 2B is 0.6% with AWQ and 2.0% with GPTQ. In contrast, there is a range of 7.5% with SparseGPT and 3.2% with Wanda. We note that this corroborates earlier findings from Williams and Aletras (2024).

Random vocabulary consistently underachieves. For every model and compression method, we observe that random calibration data (i.e. Vocabulary) produces the lowest performance. In comparison to C4, compressing Phi-2 with this random synthetic calibration data degrades performance by 0.9% for AWQ, 0.5% for GPTQ, 4.2% for SparseGPT, and 3.4% for Wanda. This illustrates that purely random synthetic data is suboptimal for calibration, even for quantization which has lower sensitivity.

6 Analysis

6.1 Calibration Data Quantity Ablation

Methodology. To assess how the quantity of calibration data impacts performance, we experiment with calibration sets of different sizes. For each calibration set, we trial subsets of n examples, where $n \in \{1, 2, 4, 8, 16, 32, 64, 128\}$. We repeat this process across five distinct calibration sets sampled from each source of calibration data.[3]

Self-calibration may be more sample efficient. In the case of pruning, self-calibration may offer

[3]We perform this ablation using smaller models (Gemma 2B and Phi-2) due to computational resource constraints.

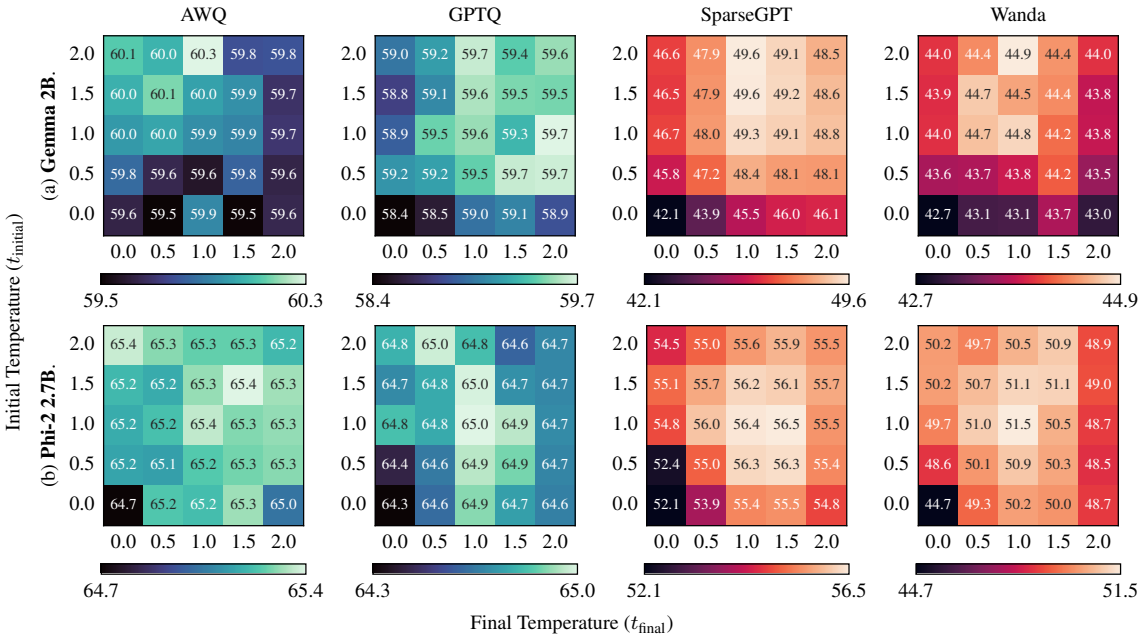

Figure 3: A joint parameter search for t_{initial} and t_{final} using $n = 10$ (§3.1) with Gemma 2B and Phi-2. We report the mean task accuracy across five distinct calibration sets.

comparable or greater performance with fewer examples. le with Phi-2 and SparseGPT, using 16 calibration examples achieves a mean accuracy of 54.5 compared to 54.3 with 128 examples of C4. While the same trend is visible for GPTQ, the performance margin between data sources is too small to reach the same conclusion. Data efficiency has the additional benefit of reducing the computational cost of compression (Frantar and Alistarh, 2023), i.e. enabling compression with fewer forward passes or the use of a larger batch size.

6.2 Sampling Strategy Ablation

Methodology. To investigate how the parameters of our sampling strategy (§3.1) impact performance, we explore a broad range of values: $t_{\text{initial}}, t_{\text{final}} \in \{0.0, 0.5, 1.0, 1.5, 2.0\}$. We emphasize that certain subsets of these values are equivalent to several standard decoding strategies:

- **Greedy decoding.** When both $t_{\text{initial}} = 0$ and $t_{\text{final}} = 0$, this is equivalent to selecting the token with the highest probability at every timestep.
- **Standard sampling.** Using a combination of $t_{\text{initial}} = 1$ and $t_{\text{final}} = 1$ is equivalent to applying softmax without a temperature parameter.
- **Temperature sampling.** When $t_{\text{initial}} = t_{\text{final}}$, a constant temperature is maintained throughout generation, equivalent to temperature sampling.

Sampling strategy can be influential for pruning. Figure 3 presents the influence of the sampling strategy parameters upon mean task accuracy. For SparseGPT and Wanda, the careful selection of sampling can offer improved performance. For example, Gemma 2B sees elevated performance when using a higher initial temperature and moderate final temperature. Conversely, using both a low initial temperature and final temperature for generation leads to substantially lower performance.

Selecting sampling parameters is not essential. We observe that it is possible to achieve within 0.4% of the maximum performance through using only standard sampling (i.e. $t_{\text{initial}} = t_{\text{final}} = 1$). This suggests that self-calibration can achieve reasonable performance with little attention towards the specific parameters used. Consequently, we suspect that using the model itself to generate calibration data is a relatively stable and reliable method.

6.3 Calibration Data Analysis

Methodology. The content and style of text can vary markedly between calibration data sources. Consequently, we seek to analyze how the text characteristics differ between them. To this end, we employ a variety of automatic metrics to assess various text characteristics of the calibration sets.

- **Perplexity.** As an indirect indicator of text quality, we calculate the average perplexity across

Dataset	PPL	Rep.	Cov.	Div.	Zipf
		Phi-2			
C4	$13.01_{0.59}$	$0.65_{0.01}$	$0.44_{0.01}$	$0.63_{0.01}$	$1.16_{0.02}$
WikiText	$10.32_{0.10}$	$0.65_{0.00}$	$0.40_{0.01}$	$0.65_{0.00}$	$1.12_{0.01}$
Vocabulary	1.98×10^{5}	$0.02_{0.00}$	$0.99_{0.00}$	$0.87_{0.00}$	$0.66_{0.00}$
Cosmopedia	$4.32_{0.08}$	$0.59_{0.01}$	$0.40_{0.01}$	$0.65_{0.00}$	$1.16_{0.01}$
Self-calibration	$4.40_{0.09}$	$0.65_{0.00}$	$0.31_{0.00}$	$0.57_{0.00}$	$1.31_{0.00}$
		Llama 3.1 8B			
Dataset	PPL	Rep.	Cov.	Div.	Zipf
C4	$8.65_{0.50}$	$0.64_{0.00}$	$0.18_{0.00}$	$0.62_{0.01}$	$1.16_{0.02}$
WikiText	$6.75_{0.11}$	$0.65_{0.00}$	$0.16_{0.00}$	$0.65_{0.00}$	$1.12_{0.01}$
Vocabulary	7.61×10^{5}	$0.01_{0.00}$	$0.87_{0.00}$	$0.91_{0.00}$	$0.49_{0.00}$
Cosmopedia	$3.37_{0.16}$	$0.55_{0.02}$	$0.18_{0.01}$	$0.65_{0.01}$	$1.17_{0.01}$
Self-calibration	$6.29_{0.09}$	$0.66_{0.00}$	$0.15_{0.00}$	$0.58_{0.00}$	$1.24_{0.00}$

Table 3: Text characteristics for all calibration sets for Phi-2 and Llama 3.1 8B, with standard deviation denoted in subscript: perplexity (PPL), repetitions (Rep.), vocabulary coverage (Cov.), n-gram diversity (Div.) and Zipf's coefficient (Zipf).

#	Generated Text
	Phi-2
1	<|endoftext|>The World Bank today approved US$5.7 million in fast-track financing to support the United Republic of Sudan's plans...
2	<|endoftext|>There are many considerations when getting a new pet. What animal is best for the family? What will work with our...
3	<|endoftext|>A new study shows most dogs are happy to see their humans even if they've been away for only an hour.\nDo you...
	Mistral 7B
1	<s>What with the heat of the summer and a seemingly endless amount of time spent outside in awe at the scenery and the local wildlife...
2	<s>As usual, there was a lot to like and a lot to dislike about last week's collection of new comic books. The biggest winner in this...
3	<s>One thing I miss from my former life, or so my wife has pointed out on a couple of occasions, is that I used to be an avid reader. I...

Table 4: The starting segment from the first three self-calibration examples generated using Phi-2 and Mistral 7B (t_{initial} and t_{final} both set to 1.0).

examples in the calibration set for a given model.

- **Repetitions.** Following Welleck et al. (2020), we report the average fraction of repeated tokens per sequence. More formally, this is computed across each sequence w of length L in dataset $\mathcal{D}$, where $\mathbb{I}$ denotes the binary indicator function:

$$R = \frac{1}{|\mathcal{D}|L} \sum_{w \in \mathcal{D}} \sum_{i=1}^{L} \mathbb{I}(w_i \in w_{1:i-1})$$

- **Vocabulary coverage.** To assess the lexical diversity of the calibration sets, we report the vocabulary coverage. We define this as the ratio between the subword tokens present in the calibration set and in the model vocabulary.

- **N-gram diversity.** Following Meister et al. (2023), we report the average fraction of unique n-grams ($n \in \{1, 2, 3, 4\}$) in the calibration set.

- **Zipf's coefficient.** Finally, we examine the extent to which the calibration set follows Zipf's law. Specifically, we calculate the fit of the exponent corresponding to the calibration set. Natural language text tends to have a value close to one.

Self-calibration data may be more consistent. Table 3 presents the results for each text metric across all datasets for Gemma 2B and Llama 3.1 8B.[4] Compared to real data sources (i.e. C4 and WikiText), self-calibration data differs across various metrics. For example with Llama 3.1 8B, self-calibration data has a lower vocabulary coverage (0.15 versus 0.16-0.18) and n-gram diversity (0.58 versus 0.62-0.65), indicating lower text diversity. However, self-calibration data has a comparable fraction of repetitions (0.66 versus 0.65). In combination with a higher Zipf's coefficient (1.24 versus 1.12-1.17), it is reasonable to conclude that self-calibration data is less diverse and more uniform.

Self-calibration data is ordinarily coherent. Table 4 presents self-calibration examples for Phi-2 and Mistral 7B. For brevity, we select the first three examples from the first calibration set. We observe that the self-generated text is typically coherent and fluent for both models. Moreover, the content is routinely semantically plausible. These properties are somewhat supported by the perplexity results in Table 3, with self-calibration demonstrating substantially lower perplexity than real data.

7 Conclusion

In this paper, we proposed self-calibration for LLM quantization and pruning as a solution to mitigate concerns about the availability, quality and representativeness of training data. Our proposed approach is intuitive and requires no external data sources, relying on the model itself. We empirically demonstrated that self-calibration maintains comparable or better downstream task performance across a variety of models and compression methods. Surprisingly, our results also revealed that self-calibration can enable higher downstream task performance than using real data. We hope that our study will inspire further work on the application of synthetic data to LLM compression.

[4] We observe similar results in other models (Appendix E).

Limitations

In this study, we experimented with English models and evaluation tasks, and therefore only English calibration data. However, Zeng et al. (2024) illustrate the importance of multilingual calibration data for the compression of multilingual models. Although we anticipate that our approach will generalize to multilingual models, we hope to explore this matter further in a future work.

Ethical Considerations

Language models are capable of generating text that is incorrect, biased, and harmful (Weidinger et al., 2022). To compress a given model, our approach requires the unsupervised generation of calibration data using the model. Consequently, the calibration data may contain material that is problematic. However, we note that our approach is unlikely to introduce new safety issues in the compressed model. For the generated calibration data to contain problematic content, it must have already been encoded in the weights of the original model.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines. *Cognitive Science*, 9(1):147–169.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *Preprint*, arXiv:2311.16867.

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku.

Pratyay Banerjee, Kuntal Kumar Pal, Arindam Mitra, and Chitta Baral. 2019. Careful selection of knowledge to solve open book question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6120–6129, Florence, Italy. Association for Computational Linguistics.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Yonatan Bisk, Rowan Zellers, Ronan Le bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439.

Yaohui Cai, Zhewei Yao, Zhen Dong, Amir Gholami, Michael W. Mahoney, and Kurt Keutzer. 2020. Zeroq: A novel zero shot quantization framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

John Chung, Ece Kamar, and Saleema Amershi. 2023. Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.

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. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting

large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew

Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou, and Iryna Gurevych. 2018. A legal perspective on training models for natural language processing. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Elias Frantar and Dan Alistarh. 2022. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, pages 4475–4488. Curran Associates, Inc.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10323–10337. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Elias Frantar, Roberto L. Castro, Jiale Chen, Torsten Hoefler, and Dan Alistarh. 2024. MARLIN: Mixed-precision auto-regressive parallel inference on large language models. *Preprint*, arXiv:2408.11743.

Elias Frantar, Eldar Kurtic, and Dan Alistarh. 2021. M-fac: Efficient matrix-free approximations of second-order information. In *Advances in Neural Information Processing Systems*, volume 34, pages 14873–14886. Curran Associates, Inc.

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, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

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, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara

von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gau-

tam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Car-

oline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski,

Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on Gemini research and technology. *Preprint*, arXiv:2403.08295.

Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A survey of quantization methods for efficient neural network inference. *Preprint*, arXiv:2103.13630.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand. Association for Computational Linguistics.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Matan Haroush, Itay Hubara, Elad Hoffer, and Daniel Soudry. 2020. The knowledge within: Methods for data-free model compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Babak Hassibi, David Stork, and Gregory Wolff. 1993. Optimal brain surgeon: Extensions and performance comparisons. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Itay Hubara, Yury Nahshan, Yair Hanani, Ron Banner, and Daniel Soudry. 2021. Accurate post training quantization with small calibration sets. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 4466–4475. PMLR.

Ajay Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. 2024. Compressing LLMs: The truth is rarely pure and never simple. In *The Twelfth International Conference on Learning Representations*.

Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro

Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, Suriya Gunasekar, Mojan Javaheripi, Piero Kauffmann, Yin Tat Lee, Yuanzhi Li, Anh Nguyen, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Michael Santacroce, Harkirat Singh Behl, Adam Taumann Kalai, Xin Wang, Rachel Ward, Philipp Witte, Cyril Zhang, and Yi Zhang. 2023. Phi-2: The surprising power of small language models.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. In *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*, pages 18–26, Suzhou, China. Association for Computational Linguistics.

Eldar Kurtic, Daniel Campos, Tuan Nguyen, Elias Frantar, Mark Kurtz, Benjamin Fineran, Michael Goin, and Dan Alistarh. 2022. The optimal BERT surgeon: Scalable and accurate second-order pruning for large language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4163–4181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Huantong Li, Xiangmiao Wu, Fanbing Lv, Daihai Liao, Thomas H. Li, Yonggang Zhang, Bo Han, and Mingkui Tan. 2023a. Hard sample matters a lot in zero-shot quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24417–24426.

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

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023c. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. In *Proceedings of Machine Learning and Systems*, volume 6, pages 87–100.

Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. In *First Conference on Language Modeling*.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, Bangkok, Thailand. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2023. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11:102–121.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 462–477. Curran Associates, Inc.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. 2021. Accelerating sparse deep neural networks. *Preprint*, arXiv:2104.08378.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. 2020. Up or down? adaptive rounding for post-training quantization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Markus Nagel, Mart van Baalen, Tijmen Blankevoort, and Max Welling. 2019. Data-free quantization through weight equalization and bias correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

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, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022.

Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Ofir Press and Lior Wolf. 2017. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, Dublin, Ireland. Association for Computational Linguistics.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. WinoGrande: An adversarial winograd schema challenge at scale. *Commun. ACM*, 64(9):99–106.

Victor Sanh, Thomas Wolf, and Alexander Rush. 2020. Movement pruning: Adaptive sparsity by fine-tuning. In *Advances in Neural Information Processing Systems*, volume 33, pages 20378–20389. Curran Associates, Inc.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023a. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Marcos Treviso, Ji-Ung Lee, Tianchu Ji, Betty van Aken, Qingqing Cao, Manuel R. Ciosici, Michael Hassid, Kenneth Heafield, Sara Hooker, Colin Raffel, Pedro H. Martins, André F. T. Martins, Jessica Zosa Forde, Peter Milder, Edwin Simpson, Noam Slonim, Jesse Dodge, Emma Strubell, Niranjan Balasubramanian, Leon Derczynski, Iryna Gurevych, and Roy Schwartz. 2023. Efficient methods for natural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 11:826–860.

Zhongwei Wan, Xin Wang, Che Liu, Samiul Alam, Yu Zheng, Jiachen Liu, Zhongnan Qu, Shen Yan, Yi Zhu, Quanlu Zhang, Mosharaf Chowdhury, and Mi Zhang. 2024. Efficient large language models: A survey. *Transactions on Machine Learning Research*. Survey Certification.

Jiaxi Wang, Ji Wu, and Lei Huang. 2022. Understanding the failure of batch normalization for transformers in nlp. In *Advances in Neural Information Processing Systems*, volume 35, pages 37617–37630. Curran Associates, Inc.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, page 214–229, New York, NY, USA. Association for Computing Machinery.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Miles Williams and Nikolaos Aletras. 2024. On the impact of calibration data in post-training quantization and pruning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10100–10118, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Carole-Jean Wu, Ramya Raghavendra, Udit Gupta, Bilge Acun, Newsha Ardalani, Kiwan Maeng, Gloria Chang, Fiona Aga, Jinshi Huang, Charles Bai, Michael Gschwind, Anurag Gupta, Myle Ott, Anastasia Melnikov, Salvatore Candido, David Brooks, Geeta Chauhan, Benjamin Lee, Hsien-Hsin Lee, Bugra Akyildiz, Maximilian Balandat, Joe Spisak, Ravi Jain, Mike Rabbat, and Kim Hazelwood. 2022. Sustainable AI: Environmental implications, challenges and opportunities. In *Proceedings of Machine Learning and Systems*, volume 4, pages 795–813.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38087–38099. PMLR.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2024. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11794–11812, Torino, Italia. ELRA and ICCL.

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. 2022. OPT: Open pre-trained transformer language models. *Preprint*, arXiv:2205.01068.

Xiangguo Zhang, Haotong Qin, Yifu Ding, Ruihao Gong, Qinghua Yan, Renshuai Tao, Yuhang Li, Fengwei Yu, and Xianglong Liu. 2021. Diversifying sample generation for accurate data-free quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15658–15667.

A Infrastructure

We use the model implementations and prepared datasets from the Hugging Face Transformers (Wolf et al., 2020) and Datasets (Lhoest et al., 2021) libraries, respectively. For pruning with SparseGPT and Wanda, we adopt the implementation from Sun et al. (2024). For quantization with AWQ and GPTQ, we use the NVIDIA TensorRT Model Optimizer and AutoGPTQ libraries, respectively.[5] To enable reproducible model evaluations, we use the EleutherAI Language Model Evaluation Harness (Gao et al., 2023). All experiments are conducted using a single NVIDIA A100 GPU.

[5] See `https://nvidia.github.io/TensorRT-Model-Optimizer` and `https://github.com/AutoGPTQ/AutoGPTQ`.

B Evaluation Datasets

Table 5 lists the number of examples used from the relevant split in each evaluation task. This is either the validation or test split, as implemented by Gao et al. (2023).

Dataset	# Examples
ARC-Easy (Clark et al., 2018)	2,376
ARC-Challenge (Clark et al., 2018)	1,172
BoolQ (Clark et al., 2019)	3,270
HellaSwag (Zellers et al., 2019)	10,042
LAMBADA (Paperno et al., 2016)	5,153
OpenBookQA (Banerjee et al., 2019)	500
PIQA (Bisk et al., 2020)	1,838
RTE (Dagan et al., 2006)	277
StoryCloze (Mostafazadeh et al., 2016)	1,511
WinoGrande (Sakaguchi et al., 2021)	1,267

Table 5: Number of examples in each evaluation task.

C Hyperparameters

Table 6 presents the hyperparameters used in all experiments. For SparseGPT and Wanda, we adopt the hyperparameters used in the original work. For AWQ and GPTQ, we use the hyperparameters from the respective implementations, NVIDIA TensorRT Model Optimizer and AutoGPTQ (§A).

Method	Hyperparameter	Value
AWQ	Bits per Weight	4
	Clip Step Size	0.05
	Group Size	128
	Maximum Clip Tokens	64
	Minimum Clip Ratio	0.5
	Scale Step Size	0.1
GPTQ	Bits per Weight	4
	Dampening	0.01
	Descending Activation Order	Yes
	Group Size	128
	Symmetric Quantization	Yes
	True Sequential Quantization	Yes
SparseGPT	Dampening	0.01
	Group Size	128
	Sparsity	2:4
Wanda	Group Size	1
	Sparsity	2:4

Table 6: The hyperparameters used in all experiments.

D Complete Results

In addition to the summarized results (Table 2, we present the task performance across compression methods and calibration data sources for each model: Gemma 2B (Table 7), Phi 2 (Table 8), OPT 6.7B (Table 9), Mistral 7B (Table 10), and Llama 3.1 8B (Table 11).

E Calibration Data Analysis

Supplementary to the text characteristics results for Phi-2 and Llama 3.1 8B presented in §6.3, we present the results for Gemma 2B (Table 12), OPT 6.7B (Table 13), Mistral 7B (Table 14). Finally, we additionally present self-calibration examples for the remaining models (Gemma 2B, OPT 6.7B, and Llama 3.1 8B) in Table 15.

Method	Dataset	ARC-e	ARC-c	BoolQ	HS	LMBD	OBQA	PIQA	RTE	SC	WG	Mean
-	-	74.1	40.4	69.7	52.8	58.3	30.8	77.2	64.3	74.9	64.6	60.7
AWQ	C4	73.4$_{0.1}$	39.4$_{0.6}$	67.8$_{0.3}$	51.4$_{0.1}$	59.6$_{1.1}$	29.2$_{0.6}$	76.8$_{0.1}$	59.1$_{2.3}$	74.0$_{0.4}$	64.1$_{0.4}$	59.5$_{0.2}$
	WikiText	73.4$_{0.4}$	39.5$_{0.8}$	68.0$_{1.5}$	51.5$_{0.0}$	59.1$_{1.2}$	29.4$_{0.6}$	76.6$_{0.3}$	58.6$_{1.7}$	74.0$_{0.3}$	64.7$_{0.5}$	59.5$_{0.2}$
	Vocabulary	72.1$_{0.3}$	39.6$_{0.4}$	67.1$_{1.0}$	51.2$_{0.1}$	57.3$_{0.1}$	29.0$_{0.8}$	75.9$_{0.5}$	61.2$_{1.3}$	74.5$_{0.2}$	64.5$_{0.2}$	59.3$_{0.2}$
	Cosmopedia	73.4$_{0.4}$	39.5$_{0.6}$	68.3$_{0.9}$	51.3$_{0.1}$	60.5$_{0.8}$	29.4$_{1.0}$	76.8$_{0.2}$	60.0$_{1.5}$	74.1$_{0.3}$	64.7$_{0.5}$	59.8$_{0.2}$
	Self-calibration	73.3$_{0.2}$	40.6$_{0.4}$	68.6$_{0.2}$	51.9$_{0.2}$	59.5$_{0.6}$	28.8$_{0.2}$	76.6$_{0.3}$	60.5$_{3.0}$	74.1$_{0.3}$	63.6$_{0.3}$	59.8$_{0.4}$
GPTQ	C4	72.8$_{0.7}$	38.4$_{0.8}$	68.7$_{1.4}$	50.9$_{0.2}$	54.9$_{1.7}$	27.8$_{1.1}$	76.1$_{0.3}$	60.6$_{2.8}$	73.6$_{0.4}$	63.1$_{0.7}$	58.7$_{0.4}$
	WikiText	71.3$_{0.8}$	37.2$_{0.4}$	67.3$_{0.8}$	51.3$_{0.2}$	55.3$_{0.3}$	29.0$_{0.9}$	76.0$_{0.4}$	61.8$_{1.8}$	73.4$_{0.4}$	63.4$_{0.4}$	58.6$_{0.3}$
	Vocabulary	70.8$_{1.0}$	36.3$_{0.7}$	69.1$_{0.7}$	50.3$_{0.3}$	53.2$_{1.8}$	29.1$_{0.9}$	75.9$_{0.3}$	58.3$_{1.4}$	72.3$_{0.5}$	63.6$_{0.7}$	57.9$_{0.3}$
	Cosmopedia	72.8$_{0.9}$	38.1$_{0.5}$	68.1$_{0.6}$	51.2$_{0.4}$	54.0$_{1.2}$	29.2$_{2.1}$	75.3$_{0.6}$	59.0$_{2.3}$	73.9$_{0.5}$	63.6$_{1.0}$	58.5$_{0.3}$
	Self-calibration	73.5$_{0.8}$	40.0$_{0.4}$	68.8$_{0.8}$	52.0$_{0.1}$	57.6$_{1.2}$	29.6$_{0.7}$	76.6$_{0.3}$	61.7$_{2.5}$	74.0$_{0.5}$	65.1$_{0.7}$	59.9$_{0.3}$
SparseGPT	C4	60.2$_{1.1}$	25.9$_{1.2}$	63.4$_{0.9}$	39.9$_{0.3}$	39.7$_{2.2}$	21.3$_{1.2}$	70.0$_{0.3}$	55.7$_{2.9}$	64.0$_{0.3}$	57.0$_{1.1}$	49.7$_{0.8}$
	WikiText	58.0$_{0.9}$	25.5$_{0.6}$	62.8$_{0.6}$	37.6$_{0.1}$	37.0$_{1.4}$	20.8$_{0.9}$	67.2$_{0.5}$	55.7$_{0.5}$	62.3$_{0.5}$	55.8$_{1.0}$	48.3$_{0.2}$
	Vocabulary	52.0$_{0.9}$	21.1$_{0.5}$	61.8$_{0.4}$	33.5$_{0.1}$	16.5$_{0.4}$	18.1$_{0.9}$	66.5$_{0.6}$	53.0$_{0.5}$	56.5$_{0.3}$	54.6$_{1.1}$	43.4$_{0.3}$
	Cosmopedia	60.1$_{1.0}$	25.7$_{0.7}$	62.2$_{0.1}$	37.8$_{0.3}$	29.6$_{1.4}$	19.8$_{0.9}$	68.1$_{0.5}$	56.3$_{1.9}$	61.2$_{0.6}$	55.9$_{0.9}$	47.7$_{0.3}$
	Self-calibration	63.0$_{0.5}$	28.0$_{0.8}$	62.7$_{0.3}$	40.5$_{0.2}$	38.1$_{1.2}$	22.1$_{0.5}$	70.7$_{0.7}$	57.5$_{1.5}$	66.7$_{0.2}$	59.0$_{0.7}$	50.8$_{0.2}$
Wanda	C4	54.9$_{0.6}$	24.6$_{0.6}$	53.7$_{2.8}$	36.4$_{0.1}$	19.8$_{0.3}$	17.0$_{0.2}$	66.5$_{0.5}$	54.6$_{1.6}$	59.1$_{0.3}$	55.5$_{0.3}$	44.2$_{0.2}$
	WikiText	54.4$_{0.5}$	23.9$_{0.5}$	60.6$_{1.4}$	35.8$_{0.2}$	20.2$_{0.9}$	17.2$_{0.9}$	66.4$_{0.2}$	55.6$_{1.4}$	58.9$_{0.3}$	55.4$_{0.5}$	44.8$_{0.4}$
	Vocabulary	51.0$_{0.4}$	22.1$_{0.3}$	54.4$_{2.7}$	33.8$_{0.1}$	13.4$_{0.4}$	16.3$_{1.1}$	65.8$_{0.2}$	51.6$_{1.8}$	57.7$_{0.3}$	54.7$_{0.6}$	42.1$_{0.4}$
	Cosmopedia	54.4$_{0.8}$	24.4$_{0.6}$	62.1$_{0.2}$	35.8$_{0.2}$	16.6$_{0.3}$	16.8$_{0.6}$	66.4$_{0.5}$	55.1$_{0.5}$	57.5$_{0.3}$	55.9$_{0.6}$	44.5$_{0.2}$
	Self-calibration	56.4$_{0.2}$	25.6$_{0.4}$	51.8$_{1.4}$	37.2$_{0.1}$	23.1$_{0.3}$	19.6$_{0.5}$	67.5$_{0.4}$	53.9$_{1.0}$	61.2$_{0.3}$	56.1$_{0.6}$	45.2$_{0.3}$

Table 7: Task accuracy across five calibration sets for Gemma 2B, with standard deviation denoted in subscript.

Method	Dataset	ARC-e	ARC-c	BoolQ	HS	LMBD	OBQA	PIQA	RTE	SC	WG	Mean
-	-	79.8	53.0	83.4	55.8	49.8	40.2	78.6	62.5	79.3	75.8	65.8
AWQ	C4	80.2$_{0.3}$	51.7$_{0.4}$	82.3$_{0.2}$	54.8$_{0.1}$	47.6$_{0.4}$	39.8$_{0.6}$	78.9$_{0.3}$	65.5$_{0.7}$	77.7$_{0.3}$	75.8$_{0.6}$	65.4$_{0.2}$
	WikiText	80.4$_{0.2}$	51.4$_{0.7}$	83.1$_{0.2}$	54.6$_{0.1}$	47.1$_{0.5}$	39.0$_{0.8}$	78.9$_{0.1}$	65.1$_{1.2}$	77.7$_{0.2}$	76.2$_{0.4}$	65.4$_{0.2}$
	Vocabulary	80.1$_{0.2}$	50.8$_{0.4}$	78.7$_{0.5}$	54.0$_{0.1}$	45.9$_{0.3}$	39.5$_{0.8}$	78.6$_{0.4}$	65.9$_{1.3}$	76.6$_{0.2}$	75.4$_{0.7}$	64.5$_{0.2}$
	Cosmopedia	80.2$_{0.2}$	51.0$_{0.5}$	81.7$_{0.5}$	54.7$_{0.1}$	46.8$_{0.1}$	39.5$_{0.6}$	78.3$_{0.0}$	66.8$_{1.4}$	77.7$_{0.2}$	75.8$_{0.6}$	65.3$_{0.2}$
	Self-calibration	80.6$_{0.3}$	51.1$_{0.3}$	82.9$_{0.1}$	54.7$_{0.1}$	47.5$_{0.2}$	39.3$_{0.3}$	78.1$_{0.2}$	65.8$_{0.8}$	78.1$_{0.5}$	75.6$_{0.7}$	65.4$_{0.2}$
GPTQ	C4	79.6$_{0.1}$	50.9$_{0.8}$	82.3$_{0.7}$	54.5$_{0.1}$	46.8$_{0.6}$	38.9$_{0.7}$	78.4$_{0.3}$	62.1$_{1.6}$	78.2$_{0.4}$	75.6$_{0.9}$	64.7$_{0.3}$
	WikiText	79.5$_{0.3}$	50.4$_{0.6}$	80.8$_{0.6}$	54.2$_{0.1}$	46.7$_{0.4}$	39.1$_{0.7}$	78.0$_{0.6}$	64.0$_{1.3}$	78.0$_{0.4}$	75.3$_{0.6}$	64.6$_{0.2}$
	Vocabulary	79.3$_{0.3}$	50.3$_{0.9}$	80.1$_{1.5}$	53.9$_{0.2}$	45.6$_{0.6}$	38.5$_{1.6}$	77.9$_{0.3}$	64.1$_{1.0}$	77.7$_{0.2}$	75.1$_{0.8}$	64.3$_{0.2}$
	Cosmopedia	79.5$_{0.4}$	50.4$_{0.3}$	80.9$_{1.1}$	54.4$_{0.1}$	45.8$_{0.6}$	38.6$_{0.6}$	78.2$_{0.5}$	63.4$_{0.9}$	77.5$_{0.5}$	74.7$_{0.8}$	64.3$_{0.1}$
	Self-calibration	79.6$_{0.3}$	51.6$_{0.6}$	82.0$_{0.7}$	54.6$_{0.2}$	46.9$_{0.8}$	39.0$_{0.4}$	77.9$_{0.3}$	64.5$_{0.8}$	78.2$_{0.6}$	75.6$_{0.7}$	65.0$_{0.3}$
SparseGPT	C4	69.3$_{0.6}$	35.1$_{0.8}$	67.5$_{1.1}$	42.1$_{0.4}$	32.6$_{0.6}$	27.7$_{0.9}$	72.0$_{1.0}$	59.2$_{2.0}$	69.0$_{0.3}$	68.6$_{0.4}$	54.3$_{0.3}$
	WikiText	69.6$_{0.7}$	35.1$_{1.0}$	63.1$_{0.7}$	40.6$_{0.3}$	33.3$_{0.3}$	27.4$_{0.4}$	71.3$_{0.7}$	57.4$_{5.1}$	68.2$_{0.2}$	67.5$_{0.8}$	53.3$_{0.5}$
	Vocabulary	67.1$_{0.3}$	32.3$_{0.5}$	64.4$_{0.6}$	37.8$_{0.1}$	20.8$_{0.7}$	23.3$_{0.8}$	71.2$_{0.5}$	57.7$_{2.2}$	64.0$_{0.2}$	62.7$_{1.3}$	50.1$_{0.2}$
	Cosmopedia	70.5$_{1.0}$	37.2$_{0.8}$	64.6$_{0.9}$	40.6$_{0.3}$	24.0$_{0.5}$	27.6$_{1.2}$	71.2$_{0.4}$	56.0$_{1.9}$	66.0$_{0.4}$	65.8$_{0.9}$	52.3$_{0.2}$
	Self-calibration	71.2$_{0.3}$	37.7$_{0.5}$	73.4$_{1.0}$	41.6$_{0.3}$	31.6$_{0.8}$	32.1$_{0.8}$	72.0$_{0.5}$	65.5$_{2.8}$	71.0$_{0.4}$	68.2$_{0.5}$	56.4$_{0.3}$
Wanda	C4	68.1$_{0.4}$	33.7$_{0.6}$	64.6$_{2.3}$	39.0$_{0.2}$	18.9$_{0.6}$	25.4$_{0.7}$	70.6$_{0.3}$	50.5$_{2.0}$	66.0$_{0.3}$	66.9$_{0.6}$	50.4$_{0.4}$
	WikiText	67.2$_{0.2}$	33.3$_{0.5}$	64.0$_{1.8}$	38.1$_{0.1}$	18.9$_{0.8}$	26.4$_{0.7}$	70.1$_{0.2}$	51.0$_{0.6}$	65.6$_{0.4}$	64.3$_{0.5}$	49.9$_{0.2}$
	Vocabulary	65.4$_{0.4}$	31.7$_{0.5}$	56.0$_{1.7}$	36.6$_{0.2}$	13.0$_{0.4}$	24.5$_{0.6}$	69.7$_{0.6}$	51.6$_{1.8}$	62.2$_{0.6}$	59.5$_{0.6}$	47.0$_{0.3}$
	Cosmopedia	66.0$_{0.9}$	31.5$_{0.7}$	66.6$_{2.1}$	38.2$_{0.2}$	16.6$_{0.7}$	23.5$_{0.5}$	69.5$_{0.4}$	53.9$_{2.1}$	64.3$_{0.5}$	64.4$_{0.4}$	49.4$_{0.4}$
	Self-calibration	67.6$_{0.3}$	35.1$_{0.6}$	68.8$_{2.0}$	39.5$_{0.1}$	18.7$_{0.5}$	25.8$_{0.4}$	70.1$_{0.3}$	59.1$_{3.5}$	65.7$_{0.3}$	64.9$_{1.3}$	51.5$_{0.7}$

Table 8: Task accuracy across five calibration sets for Phi-2 2.7B, with standard deviation denoted in subscript.

Method	Dataset	ARC-e	ARC-c	BoolQ	HS	LMBD	OBQA	PIQA	RTE	SC	WG	Mean
-	-	65.6	30.5	66.1	50.5	63.3	27.6	76.3	55.2	73.6	65.2	57.4
AWQ	C4	65.6$_{0.1}$	30.8$_{0.2}$	65.7$_{0.5}$	50.1$_{0.0}$	63.5$_{0.1}$	27.4$_{0.2}$	76.8$_{0.2}$	56.7$_{0.9}$	74.0$_{0.3}$	65.0$_{0.3}$	57.6$_{0.1}$
	WikiText	65.5$_{0.2}$	30.8$_{0.4}$	65.9$_{0.5}$	50.1$_{0.0}$	63.8$_{0.2}$	27.1$_{0.4}$	76.4$_{0.1}$	56.0$_{0.9}$	74.1$_{0.2}$	65.1$_{0.5}$	57.5$_{0.1}$
	Vocabulary	64.9$_{0.4}$	31.0$_{0.3}$	62.5$_{2.2}$	49.8$_{0.1}$	59.3$_{1.8}$	27.6$_{0.4}$	76.2$_{0.3}$	57.0$_{1.4}$	73.4$_{0.2}$	64.3$_{0.5}$	56.6$_{0.3}$
	Cosmopedia	65.4$_{0.1}$	31.0$_{0.2}$	66.0$_{0.5}$	50.0$_{0.1}$	64.0$_{0.3}$	27.5$_{0.4}$	76.8$_{0.3}$	56.3$_{0.4}$	74.1$_{0.2}$	65.1$_{0.5}$	57.6$_{0.1}$
	Self-calibration	65.8$_{0.2}$	30.9$_{0.3}$	65.6$_{0.4}$	50.2$_{0.1}$	63.6$_{0.2}$	26.9$_{0.5}$	77.1$_{0.2}$	56.8$_{0.7}$	73.9$_{0.2}$	64.9$_{0.4}$	57.6$_{0.1}$
GPTQ	C4	64.9$_{0.2}$	30.4$_{0.3}$	65.3$_{1.0}$	49.6$_{0.1}$	62.8$_{0.3}$	26.5$_{0.5}$	75.9$_{0.1}$	54.2$_{1.1}$	73.2$_{0.2}$	64.7$_{0.5}$	56.8$_{0.2}$
	WikiText	64.7$_{0.3}$	30.6$_{0.2}$	65.5$_{0.4}$	49.7$_{0.1}$	62.8$_{0.2}$	26.9$_{0.4}$	76.1$_{0.3}$	55.1$_{0.7}$	73.2$_{0.3}$	64.6$_{0.4}$	56.9$_{0.1}$
	Vocabulary	65.0$_{0.5}$	30.7$_{0.4}$	64.4$_{2.0}$	49.8$_{0.1}$	60.4$_{1.8}$	27.2$_{0.4}$	76.1$_{0.3}$	55.5$_{1.2}$	72.7$_{0.3}$	64.1$_{0.6}$	56.6$_{0.3}$
	Cosmopedia	65.1$_{0.3}$	30.2$_{0.6}$	64.8$_{0.7}$	49.6$_{0.1}$	62.5$_{0.4}$	27.2$_{0.4}$	75.5$_{0.3}$	55.3$_{0.8}$	73.3$_{0.3}$	64.5$_{0.4}$	56.8$_{0.1}$
	Self-calibration	65.5$_{0.2}$	30.3$_{0.7}$	65.1$_{0.4}$	49.8$_{0.0}$	62.3$_{0.5}$	26.9$_{0.4}$	76.0$_{0.3}$	55.2$_{1.2}$	73.2$_{0.2}$	64.7$_{0.5}$	56.9$_{0.2}$
SparseGPT	C4	59.6$_{0.3}$	25.4$_{0.7}$	63.0$_{0.4}$	43.2$_{0.1}$	55.2$_{0.8}$	23.9$_{0.5}$	72.4$_{0.6}$	53.1$_{0.4}$	70.0$_{0.3}$	61.8$_{0.5}$	52.8$_{0.2}$
	WikiText	59.1$_{0.9}$	26.2$_{0.5}$	62.1$_{0.1}$	41.3$_{0.2}$	50.6$_{0.5}$	24.4$_{0.4}$	70.1$_{0.6}$	52.9$_{0.5}$	68.1$_{0.2}$	61.3$_{1.5}$	51.6$_{0.2}$
	Vocabulary	54.4$_{0.5}$	22.8$_{0.4}$	62.4$_{0.3}$	38.4$_{0.1}$	38.1$_{1.2}$	17.7$_{0.5}$	70.3$_{0.5}$	52.6$_{0.5}$	63.9$_{0.5}$	56.3$_{1.1}$	47.7$_{0.2}$
	Cosmopedia	59.9$_{0.6}$	26.3$_{0.6}$	62.2$_{0.0}$	42.3$_{0.3}$	41.6$_{0.7}$	24.8$_{0.6}$	71.9$_{0.4}$	53.1$_{0.4}$	67.4$_{0.7}$	60.0$_{0.8}$	50.9$_{0.2}$
	Self-calibration	58.6$_{0.4}$	25.8$_{0.8}$	65.3$_{0.9}$	42.2$_{0.2}$	55.6$_{0.5}$	23.9$_{0.4}$	71.9$_{0.6}$	52.6$_{1.1}$	69.7$_{0.4}$	60.9$_{0.6}$	52.7$_{0.3}$
Wanda	C4	56.7$_{0.5}$	24.7$_{0.4}$	62.3$_{0.1}$	41.6$_{0.1}$	43.9$_{0.2}$	23.2$_{0.9}$	71.2$_{0.3}$	53.7$_{0.3}$	68.4$_{0.4}$	60.2$_{0.7}$	50.6$_{0.2}$
	WikiText	56.0$_{0.1}$	24.8$_{0.4}$	62.2$_{0.0}$	39.6$_{0.2}$	40.2$_{0.4}$	21.5$_{0.8}$	69.8$_{0.4}$	53.1$_{0.3}$	66.4$_{0.4}$	58.7$_{0.4}$	49.2$_{0.2}$
	Vocabulary	47.4$_{0.2}$	20.4$_{0.4}$	62.2$_{0.0}$	33.4$_{0.1}$	22.4$_{0.4}$	14.4$_{0.1}$	66.6$_{0.1}$	53.9$_{1.4}$	58.7$_{0.3}$	52.8$_{0.7}$	43.2$_{0.1}$
	Cosmopedia	57.0$_{0.1}$	24.7$_{0.3}$	62.2$_{0.0}$	40.7$_{0.2}$	31.0$_{0.7}$	23.0$_{0.5}$	70.9$_{0.4}$	52.7$_{0.0}$	66.0$_{0.5}$	58.5$_{0.4}$	48.7$_{0.2}$
	Self-calibration	56.3$_{0.2}$	24.6$_{0.3}$	64.1$_{0.4}$	41.3$_{0.1}$	45.7$_{0.5}$	21.5$_{0.3}$	70.8$_{0.4}$	53.9$_{0.3}$	68.3$_{0.3}$	60.1$_{0.7}$	50.7$_{0.2}$

Table 9: Task accuracy across five calibration sets for OPT 6.7B, with standard deviation denoted in subscript.

Method	Dataset	ARC-e	ARC-c	BoolQ	HS	LMBD	OBQA	PIQA	RTE	SC	WG	Mean
-	-	79.6	48.7	82.4	60.9	69.2	33.6	80.3	67.9	78.3	73.6	67.4
AWQ	C4	79.3$_{0.0}$	48.1$_{0.2}$	81.6$_{0.4}$	59.9$_{0.1}$	68.1$_{0.4}$	33.6$_{0.6}$	79.7$_{0.2}$	69.2$_{0.4}$	78.3$_{0.1}$	72.9$_{0.2}$	67.1$_{0.0}$
	WikiText	79.4$_{0.3}$	48.0$_{0.6}$	81.8$_{0.4}$	60.0$_{0.1}$	68.1$_{0.2}$	33.7$_{0.4}$	79.9$_{0.2}$	69.1$_{0.8}$	78.5$_{0.3}$	72.4$_{0.5}$	67.1$_{0.1}$
	Vocabulary	79.3$_{0.3}$	48.3$_{0.2}$	79.3$_{0.5}$	59.9$_{0.1}$	67.1$_{0.6}$	33.4$_{0.5}$	79.7$_{0.1}$	67.4$_{0.5}$	77.8$_{0.4}$	72.3$_{0.7}$	66.5$_{0.1}$
	Cosmopedia	79.3$_{0.3}$	48.6$_{0.1}$	81.6$_{0.4}$	60.0$_{0.1}$	67.6$_{0.2}$	34.0$_{0.5}$	79.4$_{0.2}$	67.5$_{1.2}$	78.9$_{0.2}$	72.8$_{0.4}$	67.0$_{0.2}$
	Self-calibration	79.1$_{0.2}$	47.9$_{0.9}$	81.8$_{0.3}$	60.0$_{0.1}$	68.0$_{0.2}$	34.2$_{0.5}$	80.1$_{0.2}$	68.2$_{1.2}$	78.5$_{0.2}$	72.7$_{0.3}$	67.0$_{0.2}$
GPTQ	C4	79.0$_{0.3}$	48.0$_{0.5}$	81.8$_{0.7}$	60.0$_{0.2}$	68.0$_{0.6}$	32.7$_{0.2}$	80.1$_{0.3}$	67.1$_{2.1}$	78.3$_{0.3}$	73.1$_{0.5}$	66.8$_{0.3}$
	WikiText	79.1$_{0.4}$	47.9$_{0.5}$	82.1$_{0.5}$	60.1$_{0.1}$	68.0$_{0.5}$	32.2$_{0.7}$	80.0$_{0.3}$	67.5$_{1.7}$	78.4$_{0.4}$	73.4$_{0.4}$	66.9$_{0.3}$
	Vocabulary	78.4$_{0.4}$	47.1$_{1.0}$	81.6$_{0.3}$	59.9$_{0.1}$	67.0$_{0.6}$	32.5$_{0.6}$	79.7$_{0.2}$	63.9$_{1.7}$	77.2$_{0.2}$	72.5$_{0.5}$	66.0$_{0.1}$
	Cosmopedia	79.1$_{0.3}$	47.1$_{0.5}$	81.5$_{0.6}$	60.0$_{0.1}$	67.9$_{0.2}$	32.0$_{0.4}$	80.2$_{0.3}$	66.7$_{2.1}$	78.1$_{0.4}$	73.1$_{0.4}$	66.6$_{0.2}$
	Self-calibration	78.3$_{0.3}$	46.8$_{0.5}$	80.7$_{0.9}$	59.9$_{0.2}$	66.9$_{0.6}$	32.0$_{0.6}$	79.4$_{0.4}$	63.0$_{0.7}$	78.3$_{0.2}$	73.1$_{0.4}$	65.9$_{0.2}$
SparseGPT	C4	67.4$_{0.7}$	34.3$_{0.8}$	75.2$_{0.9}$	46.7$_{0.3}$	53.9$_{0.6}$	23.9$_{0.5}$	73.3$_{0.7}$	60.3$_{1.9}$	71.8$_{0.5}$	66.3$_{0.9}$	57.3$_{0.3}$
	WikiText	67.2$_{0.4}$	33.3$_{0.5}$	64.3$_{0.2}$	45.2$_{0.2}$	54.0$_{0.5}$	23.1$_{0.4}$	71.3$_{0.3}$	60.1$_{2.5}$	70.4$_{0.4}$	66.3$_{0.6}$	55.5$_{0.3}$
	Vocabulary	62.5$_{1.4}$	30.0$_{0.8}$	71.0$_{0.6}$	44.5$_{0.2}$	42.7$_{0.4}$	20.2$_{0.6}$	71.5$_{0.3}$	56.9$_{2.9}$	68.6$_{0.3}$	62.2$_{0.9}$	53.0$_{0.4}$
	Cosmopedia	69.5$_{0.4}$	35.4$_{0.7}$	66.4$_{0.8}$	45.6$_{0.3}$	42.9$_{0.8}$	24.2$_{0.7}$	71.7$_{0.2}$	60.7$_{3.7}$	69.5$_{0.5}$	64.9$_{0.5}$	55.1$_{0.3}$
	Self-calibration	65.9$_{0.5}$	32.2$_{0.8}$	76.0$_{0.9}$	46.7$_{0.1}$	51.7$_{0.8}$	23.2$_{0.5}$	73.1$_{0.6}$	60.5$_{1.0}$	72.5$_{0.2}$	66.5$_{0.4}$	56.8$_{0.3}$
Wanda	C4	64.3$_{0.4}$	30.5$_{0.6}$	70.4$_{0.6}$	44.3$_{0.1}$	42.3$_{0.3}$	21.2$_{0.6}$	71.9$_{0.3}$	56.4$_{2.0}$	70.8$_{0.3}$	64.5$_{0.5}$	53.7$_{0.3}$
	WikiText	64.8$_{0.6}$	31.0$_{0.5}$	66.1$_{0.8}$	43.5$_{0.1}$	44.5$_{0.2}$	21.6$_{0.4}$	70.7$_{0.2}$	58.6$_{1.1}$	70.0$_{0.2}$	63.6$_{0.5}$	53.4$_{0.2}$
	Vocabulary	58.4$_{0.5}$	26.1$_{0.2}$	64.3$_{0.7}$	39.5$_{0.2}$	29.8$_{0.3}$	17.8$_{0.6}$	69.7$_{0.1}$	54.6$_{1.4}$	64.4$_{0.2}$	59.0$_{0.7}$	48.4$_{0.2}$
	Cosmopedia	65.5$_{0.2}$	32.4$_{0.2}$	65.1$_{0.6}$	43.6$_{0.1}$	37.6$_{0.3}$	21.0$_{0.7}$	70.7$_{0.3}$	58.1$_{1.6}$	69.3$_{0.2}$	63.4$_{0.5}$	52.7$_{0.2}$
	Self-calibration	63.7$_{0.3}$	30.0$_{0.6}$	68.6$_{1.4}$	44.3$_{0.1}$	41.7$_{0.3}$	20.6$_{0.7}$	71.6$_{0.4}$	58.9$_{0.6}$	70.8$_{0.3}$	64.8$_{0.4}$	53.5$_{0.1}$

Table 10: Task accuracy across five calibration sets for Mistral 7B, with standard deviation denoted in subscript.

Method	Dataset	ARC-e	ARC-c	BoolQ	HS	LMBD	OBQA	PIQA	RTE	SC	WG	Mean
-	-	81.4	51.5	82.2	60.0	67.1	33.4	80.0	70.0	78.2	74.0	67.8
AWQ	C4	80.7$_{0.6}$	50.3$_{0.6}$	79.9$_{0.6}$	58.6$_{0.1}$	66.1$_{0.8}$	34.6$_{0.4}$	79.5$_{0.4}$	68.3$_{0.9}$	77.3$_{0.3}$	73.4$_{0.4}$	66.9$_{0.2}$
	WikiText	80.4$_{0.3}$	50.4$_{0.9}$	81.3$_{0.8}$	58.8$_{0.1}$	65.6$_{0.2}$	34.5$_{0.3}$	79.5$_{0.2}$	68.6$_{0.5}$	77.7$_{0.5}$	73.8$_{0.3}$	67.1$_{0.1}$
	Vocabulary	80.3$_{0.4}$	49.0$_{1.0}$	80.0$_{0.5}$	58.2$_{0.2}$	63.4$_{0.4}$	34.1$_{0.8}$	79.2$_{0.2}$	66.3$_{0.9}$	76.6$_{0.2}$	72.3$_{0.2}$	66.0$_{0.3}$
	Cosmopedia	81.1$_{0.1}$	50.3$_{0.7}$	81.0$_{0.5}$	58.8$_{0.1}$	65.4$_{0.5}$	34.7$_{0.5}$	79.6$_{0.2}$	67.1$_{2.3}$	77.6$_{0.4}$	73.0$_{0.5}$	66.9$_{0.3}$
	Self-calibration	80.8$_{0.3}$	50.0$_{0.6}$	80.6$_{0.2}$	58.7$_{0.1}$	65.1$_{0.4}$	34.5$_{0.6}$	79.6$_{0.3}$	66.1$_{2.1}$	77.3$_{0.3}$	73.7$_{0.3}$	66.6$_{0.3}$
GPTQ	C4	80.5$_{0.4}$	48.7$_{0.8}$	80.5$_{0.4}$	58.8$_{0.3}$	65.4$_{0.7}$	32.7$_{1.2}$	79.6$_{0.3}$	71.9$_{1.4}$	78.1$_{0.3}$	72.8$_{1.0}$	66.9$_{0.3}$
	WikiText	80.4$_{0.4}$	48.5$_{1.0}$	80.4$_{0.5}$	58.8$_{0.1}$	65.3$_{0.2}$	33.2$_{0.8}$	79.3$_{0.1}$	69.5$_{1.6}$	77.7$_{0.4}$	72.9$_{0.2}$	66.6$_{0.3}$
	Vocabulary	80.0$_{0.5}$	47.9$_{1.2}$	80.9$_{0.6}$	58.0$_{0.2}$	62.8$_{0.5}$	33.8$_{0.4}$	79.2$_{0.5}$	65.1$_{1.0}$	76.7$_{0.3}$	72.6$_{0.5}$	65.7$_{0.1}$
	Cosmopedia	80.8$_{0.7}$	49.1$_{0.4}$	81.1$_{0.7}$	58.9$_{0.1}$	64.7$_{0.7}$	34.0$_{0.9}$	79.2$_{0.3}$	70.3$_{1.0}$	77.9$_{0.4}$	73.4$_{0.4}$	66.9$_{0.1}$
	Self-calibration	79.7$_{0.5}$	47.4$_{0.8}$	81.4$_{0.3}$	58.5$_{0.3}$	64.9$_{0.3}$	30.8$_{0.8}$	79.2$_{0.3}$	69.5$_{1.7}$	77.7$_{0.3}$	72.3$_{0.4}$	66.1$_{0.3}$
SparseGPT	C4	63.4$_{0.9}$	31.0$_{1.1}$	71.3$_{1.8}$	43.9$_{0.4}$	50.9$_{0.6}$	23.0$_{1.0}$	70.7$_{0.6}$	57.4$_{1.5}$	70.4$_{0.4}$	65.9$_{0.5}$	54.8$_{0.3}$
	WikiText	62.1$_{1.0}$	30.0$_{1.3}$	66.6$_{2.2}$	41.4$_{0.1}$	49.1$_{1.4}$	22.6$_{1.3}$	68.3$_{0.4}$	53.5$_{0.8}$	68.7$_{0.5}$	63.9$_{1.3}$	52.6$_{0.4}$
	Vocabulary	57.0$_{0.5}$	25.6$_{1.1}$	65.3$_{1.4}$	37.2$_{0.2}$	28.9$_{1.2}$	20.1$_{0.7}$	69.2$_{0.5}$	52.6$_{0.3}$	61.3$_{0.5}$	56.1$_{0.8}$	47.3$_{0.4}$
	Cosmopedia	64.6$_{1.0}$	31.9$_{1.1}$	64.8$_{1.1}$	41.6$_{0.4}$	34.3$_{0.9}$	21.8$_{1.3}$	68.9$_{0.5}$	53.6$_{0.5}$	66.4$_{0.7}$	61.6$_{1.1}$	50.9$_{0.3}$
	Self-calibration	64.5$_{0.8}$	32.3$_{1.0}$	70.7$_{0.9}$	42.8$_{0.2}$	43.3$_{1.5}$	22.0$_{0.9}$	69.5$_{0.3}$	58.9$_{2.4}$	69.6$_{0.5}$	64.1$_{0.5}$	53.8$_{0.4}$
Wanda	C4	57.7$_{0.6}$	26.8$_{0.3}$	66.9$_{1.6}$	38.2$_{0.1}$	33.9$_{0.5}$	19.3$_{0.7}$	68.6$_{0.2}$	53.5$_{0.9}$	65.6$_{0.2}$	59.8$_{0.4}$	49.0$_{0.3}$
	WikiText	57.9$_{0.4}$	28.2$_{0.5}$	66.9$_{0.3}$	37.8$_{0.2}$	34.6$_{0.4}$	20.4$_{0.3}$	67.8$_{0.3}$	53.1$_{0.3}$	65.2$_{0.3}$	59.8$_{0.5}$	49.2$_{0.1}$
	Vocabulary	53.6$_{0.4}$	22.9$_{0.6}$	62.3$_{0.2}$	34.0$_{0.1}$	20.0$_{1.2}$	18.1$_{0.9}$	66.1$_{0.5}$	52.1$_{1.3}$	60.4$_{0.4}$	57.3$_{1.1}$	44.7$_{0.3}$
	Cosmopedia	57.7$_{0.5}$	26.1$_{0.2}$	65.7$_{0.6}$	37.2$_{0.2}$	26.6$_{0.4}$	20.0$_{0.8}$	68.6$_{0.4}$	52.4$_{0.7}$	64.0$_{0.3}$	58.7$_{0.6}$	47.7$_{0.2}$
	Self-calibration	58.1$_{0.4}$	27.5$_{0.2}$	67.7$_{0.4}$	37.8$_{0.2}$	31.8$_{0.4}$	19.9$_{0.7}$	69.0$_{0.4}$	54.3$_{1.3}$	65.7$_{0.4}$	59.2$_{0.5}$	49.1$_{0.1}$

Table 11: Task accuracy across five calibration sets for Llama 3.1 8B, with standard deviation denoted in subscript.

Dataset	PPL	Rep.	Cov.	Div.	Zipf
C4	$19.30_{1.06}$	$0.66_{0.01}$	$0.10_{0.00}$	$0.63_{0.01}$	$1.16_{0.01}$
WikiText	$14.93_{0.58}$	$0.68_{0.00}$	$0.09_{0.00}$	$0.65_{0.00}$	$1.12_{0.01}$
Vocabulary	4.31×10^{6}	$0.00_{0.00}$	$0.64_{0.00}$	$0.96_{0.00}$	$0.27_{0.00}$
Cosmopedia	$6.49_{0.22}$	$0.59_{0.01}$	$0.09_{0.00}$	$0.65_{0.01}$	$1.19_{0.01}$
Self-calibration	$7.22_{0.15}$	$0.68_{0.00}$	$0.07_{0.00}$	$0.59_{0.00}$	$1.25_{0.01}$

Table 12: Text characteristics for all Gemma 2B calibration sets, with standard deviation denoted in subscript.

Dataset	PPL	Rep.	Cov.	Div.	Zipf
C4	$11.80_{0.46}$	$0.65_{0.01}$	$0.44_{0.01}$	$0.63_{0.01}$	$1.16_{0.01}$
WikiText	$11.05_{0.16}$	$0.65_{0.00}$	$0.40_{0.01}$	$0.65_{0.00}$	$1.12_{0.01}$
Vocabulary	2.02×10^{5}	$0.02_{0.00}$	$0.99_{0.00}$	$0.87_{0.00}$	$0.66_{0.00}$
Cosmopedia	$5.76_{0.13}$	$0.60_{0.01}$	$0.40_{0.01}$	$0.65_{0.01}$	$1.15_{0.01}$
Self-calibration	$7.66_{0.16}$	$0.67_{0.00}$	$0.30_{0.00}$	$0.55_{0.01}$	$1.27_{0.00}$

Table 13: Text characteristics for all OPT 6.7B calibration sets, with standard deviation denoted in subscript.

Dataset	PPL	Rep.	Cov.	Div.	Zipf
C4	$7.81_{0.17}$	$0.65_{0.01}$	$0.47_{0.01}$	$0.63_{0.00}$	$1.15_{0.01}$
WikiText	$5.81_{0.07}$	$0.67_{0.00}$	$0.42_{0.01}$	$0.65_{0.00}$	$1.10_{0.01}$
Vocabulary	1.64×10^{5}	$0.03_{0.00}$	$0.98_{0.00}$	$0.89_{0.00}$	$0.55_{0.00}$
Cosmopedia	$3.07_{0.03}$	$0.53_{0.01}$	$0.46_{0.00}$	$0.67_{0.01}$	$1.15_{0.01}$
Self-calibration	$5.79_{0.15}$	$0.66_{0.00}$	$0.41_{0.00}$	$0.59_{0.00}$	$1.24_{0.00}$

Table 14: Text characteristics for all Mistral 7B calibration sets, with standard deviation denoted in subscript.

#	Generated Text
	Gemma 2B
1	<bos>The <b>G36S<b> is an assault rifle created for the German Army from 1997 to 2010 by Heckler & Koch. It is a simplified...
2	<bos>I have a long story to share. Long story short, I've learned that I have to have a higher IQ. You must have an IQ at least 142 as...
3	<bos>You have come to the right place to learn about the different types of floor drain and how to select what's best for you. We have...
	OPT 6.7B
1	<s>It's an interesting concept, but there's no way anyone can get past the cost. I can't see this going anywhere.\nWell this is what's...
2	<s>A lot of things are not on par with the other versions. I love the game, but there are some major differences, so if there are people...
3	<s>A new study on the use of blockchain technology – an online ledger – to help banks make smart lending decision and...
	Llama 3.1 8B
1	<|begin_of_text|>You are at:Home»Lifestyle»Food»I have a problem. . . and it's called peanut butter!\nI have a problem. . . and it's...
2	<|begin_of_text|>The American\nOmnithologists' Union\n**Chilson in Africa: New Records for the Birds of...
3	<|begin_of_text|>The following is a partial list of the current year's and past year's notable events in the life of the Anglican...

Table 15: The starting segment from the first three self-calibration examples generated using Gemma 2B, OPT 6.7B, and Llama 3.1 8B (t_{initial} and t_{final} both set to 1.0).