

Self-calibration for Language Model Quantization and Pruning

Miles Williams^{◇♣} George Chrysostomou[♣] Nikolaos Aletras[◇]

[◇]University of Sheffield

[♣]Enterprise AI Services, AstraZeneca

{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, this is randomly sampled web text, 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, with a view to better approximating 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

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; Mesnard et al., 2024). However, they demand extensive computational resources for inference (Wu et al., 2022; Luccioni et al., 2023), presenting a limiting factor for 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 while retaining downstream task performance (Wan et al., 2024).

Quantization and pruning have emerged as prominent model compression approaches for

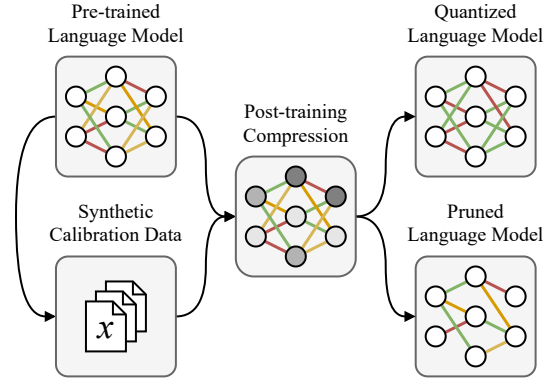


Figure 1: Self-calibration for the post-training quantization and pruning of language models.

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).

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; Sun et al., 2024; Lin et al., 2024), aiming to reflect the model training data distribution.

However, recent work has questioned the influence of calibration data in LLM compression. Jaiswal et al. (2024) hint that the careful selection of calibration data may benefit high-sparsity pruning. Concurrently, Williams and Aletras (2024) illustrate the impact of calibration data in quantization and pruning. Finally, Zeng et al. (2024) and Kurz et al. (2024) highlight the role of language-specific calibration data for multilingual models.

¹<https://github.com/mlsw/llm-compression-calibration>

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

Model	Reference	Open Source	
		Weights	Data
GPT-4	Achiam 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	Anil et al. (2024)	✗	✗
OLMo	Groeneveld et al. (2024)	✓	✓
Claude 3	Anthropic (2024)	✗	✗
Gemma	Mesnard 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.

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.

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). However, the distribution of such datasets may deviate from the pre-training distribution of other LLMs.

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 pro-

pose 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 task 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 through 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 the most 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:

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

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](#); [Chrysostomou 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 study on the quantity of calibration data used in §6.1.

Certain models (Gemma, Mistral, and Llama) were trained using multilingual 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 model families: (1) **Gemma 2B** ([Mesnard et al., 2024](#)), (2) **Phi-2 2.7B** ([Jawaheripi 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](#)).²

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](#)).

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

Method	Type	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).³ For self-calibration, we set t_{initial} and t_{final} as 1.0 (i.e. standard sampling), to enable a fair comparison between models. However, we emphasize that the careful selection of these parameters could lead to further performance improvements. We provide a deeper analysis surrounding the impact of the temperature schedule in §6.2.

Self-calibration outperforms other synthetic datasets. We observe that the performance of self-calibration matches or exceeds other synthetic datasets in 17 out of 20 instances. For example, when quantizing Gemma 2B with GPTQ, self-calibration records a mean accuracy of 59.9%, compared to 58.5% with Cosmopedia and 57.9% with 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 is more representative of the training distribution of each model, compared to other synthetic datasets.

Self-calibration can outperform real-world data. Our results show that for Phi-2, Gemma 2B, and

³Complete results are presented in Appendix E.

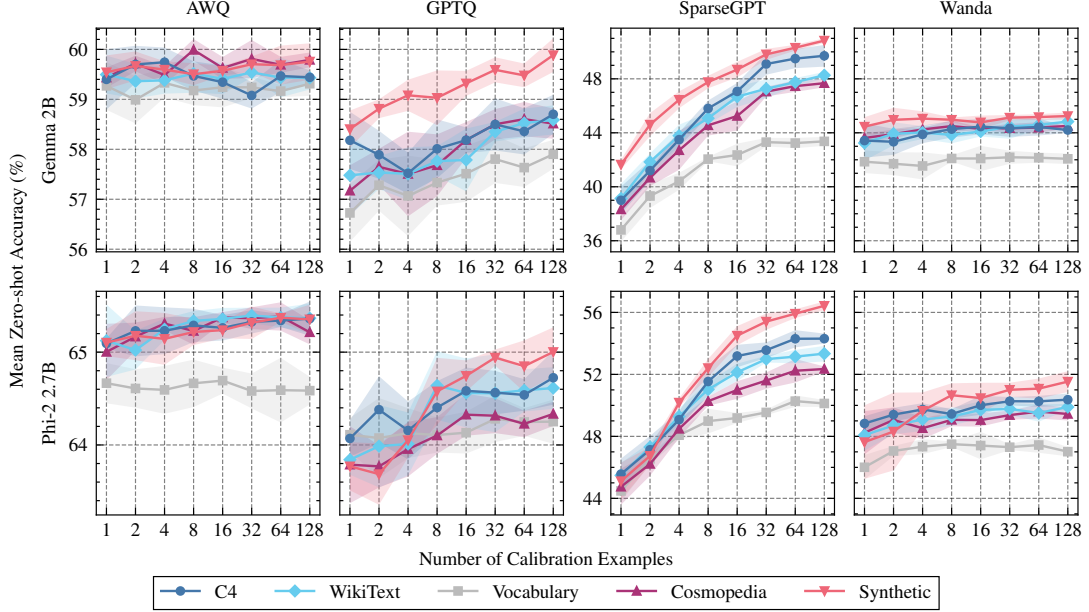


Figure 2: The mean zero-shot accuracy when compressing Gemma 2B and Phi-2 with each method. We present the mean value and standard deviation (shaded) across five distinct calibration sets sampled from each data source.

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). Although self-calibration does not outperform real data for Mistral 7B and Llama 3.1 8B, we observe that the performance is as competitive with real data as Cosmopedia (i.e. matches or outperforms Cosmopedia in five out of 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 second only 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 appear less sensitive to the calibration data. 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. This suggests that the choice of calibration dataset is less critical when applying quantization

to language models, corroborating 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 may be less sensitive.

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.⁴

Self-calibration may be more sample efficient.

In the case of pruning, self-calibration may offer comparable or greater performance with less data. For example, when pruning Phi-2 with SparseGPT,

⁴We perform this ablation using smaller models (Gemma 2B and Phi-2) due to computational resource constraints.

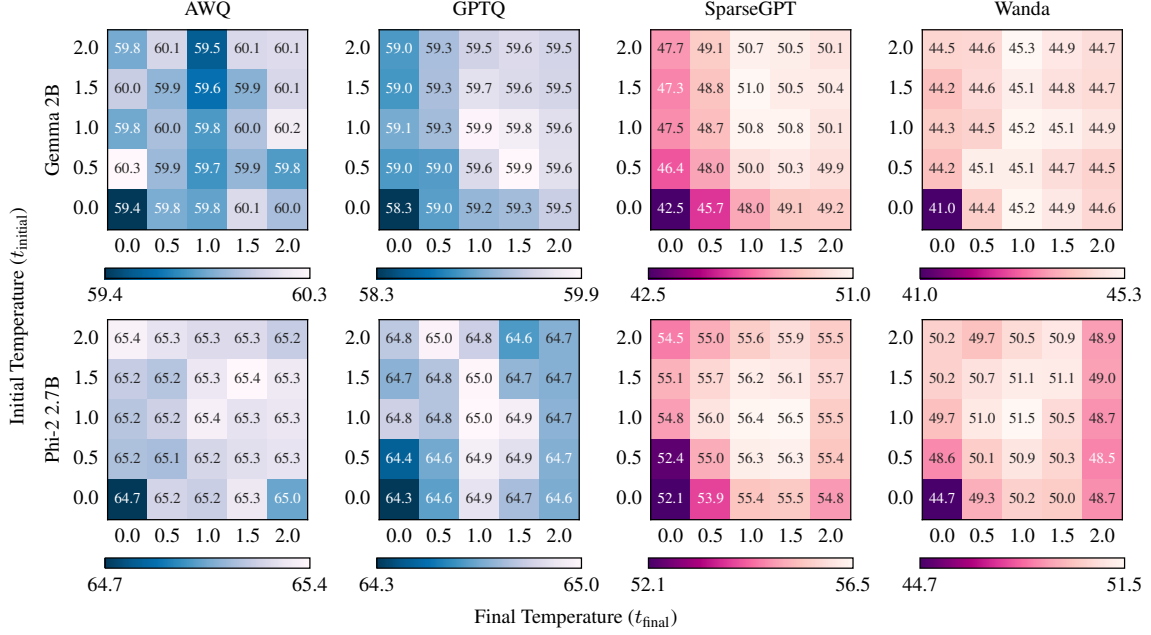


Figure 3: A joint parameter search for t_{initial} and t_{final} using $n = 10$ (§3.1) with Gemma 2B and Phi-2. We report the mean task accuracy across five distinct calibration sets.

C4 reaches a mean accuracy of 54.3% with 128 examples, while self-calibration achieves 54.5% with only 16 examples. While the same trend is visible for GPTQ, the performance margin between data sources is too small to draw the same conclusion. Finally, we note that improved sample efficiency can reduce the computational cost of model compression (Frantar and Alistarh, 2023). In practical terms, this can enable (1) fewer forward passes, as a direct result of fewer examples, or (2) an increased batch size, due to fewer intermediate activations.

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 influence performance.

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

Selecting sampling parameters is not essential.

We observe that it is possible to achieve within 0.5 points 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

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:

$$D = \frac{1}{N} \sum_{n=1}^N \frac{\# \text{ unique } n\text{-grams}}{\# \text{ total } n\text{-grams}}$$

- **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 is generally coherent text.

Table 3 presents self-calibration data from Gemma 2 and Llama 3.1 8B. For brevity, we select the first three texts generated by each model. We observe that the self-generated text is typically coherent and fluent in both models. Moreover, the content is routinely semantically plausible. These properties are somewhat supported by the perplexity results in Table 4, with self-calibration demonstrating substantially lower perplexity than real data.

Self-calibration may produce less diverse text.

Table 4 presents the text characteristics for Gemma 2B and Llama 3.1 8B across all datasets.⁵ Compared to real data sources (i.e. C4 and WikiText), self-calibration data differs across various metrics. For example, self-calibration data from Llama 3.1 8B has a lower vocabulary coverage (0.15 versus 0.16-0.18) and n -gram diversity (0.58 versus 0.62-0.65). However, self-calibration data has a higher Zipf’s coefficient (1.24 versus 1.12-1.16)

⁵We observe similar results in other models (Appendix D).

#	Generated Text
Gemma 2B	
1	<bos>The G36S is an assault rifle created for the German Army from 1997 to 2010 by Heckler & Koch. It is a simplified...
2	<bos>Are you considering making an investment in a used or new Mercedes-Benz S-Class? Make Mercedes-Benz of Houston your...
3	<bos>I recently created a poll to see what everyone thinks the best of the current generation of S13’s are. I have gotten some great...
Llama 3.1 8B	
1	< begin_of_text >You are at:Home»Lifestyle»Food»I have a problem... and it’s called peanut butter!«I have a problem... and it’s...
2	< begin_of_text >When we’re in the heat of our journey, when our plans and goals and hopes and dreams and desires are in control...
3	< begin_of_text >This article by David K. Li from NBC News on February 9, 2021, talks about the increase of the vaccine mandate...

Table 3: The starting segment of the first three synthetic texts generated by Gemma 2B and Llama 3.1 8B, using standard sampling.

Dataset	PPL	Rep.	Cov.	Div.	Zipf
Gemma 2B					
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}
Llama 3.1 8B					
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 4: Text characteristics across all calibration sets for Gemma 2B and Llama 3.1 8B, with standard deviation denoted in subscript.

and slightly elevated repetitions (0.66 versus 0.64-0.65). Overall, this suggests that self-calibration data is typically less diverse compared to 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, instead relying on the model itself. We empirically demonstrated that self-calibration maintains comparable or greater 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.

Limitations

In this study, we experimented with English models and evaluation tasks, and therefore only English calibration data. However, recent work has illustrated the importance of language-specific calibration data when compressing multilingual models (Zeng et al., 2024; Kurz et al., 2024). 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 from the model itself. Consequently, the calibration data may contain material that is problematic. However, we note that this 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.

Acknowledgments

We are grateful to Vladimir Poroshin, Vitor Jeronimo, Szymon Palucha, Christopher May, Mario Sanger, and the anonymous reviewers for their invaluable feedback. MW is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation grant EP/S023062/1. NA is supported by EPSRC grant EP/Y009800/1, part of the RAI UK Keystone projects.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, et al. 2023. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- 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, M  rouane Debbah,   tienne 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.
- 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, et al. 2024. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- 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 Tram  r, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song,   lfar 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.
- George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. [Investigating hallucinations in pruned large language models for abstractive summarization](#). *Transactions of the Association for Computational Linguistics*, 12:1163–1181.
- 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 Textual 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, et al. 2024. [The Llama 3 herd of models](#). Preprint, arXiv:2407.21783.
- Richard Eckart de Castilho, Giulia Dore, Thomas Marconi, 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 Hoefer, 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 Hoefer, 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, et al. 2023. [A framework for few-shot language model evaluation](#).
- 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 de-generation](#). 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, et al. 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, Léo 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, Léo Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, et al. 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.
- Simon Kurz, Jian-Jia Chen, Lucie Flek, and Zhixue Zhao. 2024. [Investigating language-specific calibration for pruning multilingual large language models](#). *Preprint*, arXiv:2408.14398.
- 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, Guntjan 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 Viguiet, 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*.
- 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, et al. 2024. [Gemma: Open models based on Gemini research and technology](#). *Preprint*, arXiv:2403.08295.
- 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)*.
- 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 Lazariidou, 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, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, et al. 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, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, et al. 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 Khoshabi, 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, et al. 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, et al. 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.⁶ 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 80GB GPU.

B Evaluation Datasets

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

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).

D Calibration Data Analysis

Supplementary to the text characteristic results for Gemma 2 and Llama 3.1 8B presented in §6.3, we present the results for Phi-2 2.7B, OPT 6.7B, and Mistral 7B in Table 8. Finally, we also present self-calibration examples for the remaining models (Phi-2 2.7B, OPT 6.7B, and Mistral 7B) in Table 7.

E 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 9), Phi-2 2.7B (Table 10), OPT 6.7B (Table 11), Mistral 7B (Table 12), and Llama 3.1 8B (Table 13).

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.

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.

#	Generated Text
Phi-2 2.7B	
1	< endoftext >\n\ndef simple_math_problem() -> int:\n Nathan collects all the leaves from his 8 bushes.\n Each bush has 16 plants...
2	< endoftext >\n\n## TAKING OWNERSHIP OF WORKPLACE FEARS \n\nGood morning everyone,\n\nI'm here today to talk...
3	< endoftext >\n\n## BOOSTING ANIMAL POPULATION IN INDIA\n\nIndia is known for its rich biodiversity, with a variety of...
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>You are here\n\nOlympics Day 10: US men, Phelps, Lochte & swimming's greatest\n\nUpdated: Wednesday, 20 August 2014...
3	<S>Tampa Bay Lightning\nI'm a simple man, I see Lightning, I read Stamkos.\nYeah, the Bolts are gonna be so fun to watch next year...
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>While working on an assignment on how to manage a conflict in our teams at University, I was inspired to do so in my own work...
3	<S>Using "the old reliable" "the old faithful" methods of lead generation can quickly become... well, let's say, repetitive, uninspired...

Table 7: The starting segment of the first three synthetic texts generated by Phi-2 2.7B, OPT 6.7B, and Mistral 7B, using standard sampling.

⁶See <https://nvidia.github.io/TensorRT-Model-Optimizer> and <https://github.com/AutoGPTQ/AutoGPTQ>.

Dataset	PPL	Rep.	Cov.	Div.	Zipf
Phi-2 2.7B					
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	2.41 _{0.02}	0.67 _{0.00}	0.33 _{0.00}	0.57 _{0.00}	1.24 _{0.00}
OPT 6.7B					
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	7.06 _{0.21}	0.66 _{0.00}	0.32 _{0.01}	0.57 _{0.01}	1.25 _{0.01}
Mistral 7B					
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 8: Text characteristics across all calibration sets for Phi-2 2.7B, OPT 6.7B, and Mistral 7B, with standard deviation denoted in subscript.

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 Self-calibration	73.4 _{0.4} 73.3 _{0.2}	39.5 _{0.6} 40.6 _{0.4}	68.3 _{0.9} 68.6 _{0.2}	51.3 _{0.1} 51.9 _{0.2}	60.5 _{0.8} 59.5 _{0.6}	29.4 _{1.0} 28.8 _{0.2}	76.8 _{0.2} 76.6 _{0.3}	60.0 _{1.5} 60.5 _{3.0}	74.1 _{0.3} 74.1 _{0.3}	64.7 _{0.5} 63.6 _{0.3}	59.8 _{0.2} 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 Self-calibration	72.8 _{0.9} 73.5 _{0.8}	38.1 _{0.5} 40.0 _{0.4}	68.1 _{0.6} 68.8 _{0.8}	51.2 _{0.4} 52.0 _{0.1}	54.0 _{1.2} 57.6 _{1.2}	29.2 _{2.1} 29.6 _{0.7}	75.3 _{0.6} 76.6 _{0.3}	59.0 _{2.3} 61.7 _{2.5}	73.9 _{0.5} 74.0 _{0.5}	63.6 _{1.0} 65.1 _{0.7}	58.5 _{0.3} 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 Self-calibration	60.1 _{1.0} 63.0 _{0.5}	25.7 _{0.7} 28.0 _{0.8}	62.2 _{0.1} 62.7 _{0.3}	37.8 _{0.3} 40.5 _{0.2}	29.6 _{1.4} 38.1 _{1.2}	19.8 _{0.9} 22.1 _{0.5}	68.1 _{0.5} 70.7 _{0.7}	56.3 _{1.9} 57.1 _{1.5}	61.2 _{0.6} 66.7 _{0.2}	55.9 _{0.9} 59.0 _{0.7}	47.7 _{0.3} 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 Self-calibration	54.4 _{0.8} 56.4 _{0.2}	24.4 _{0.6} 25.6 _{0.4}	62.1 _{0.2} 51.8 _{1.4}	35.8 _{0.2} 37.2 _{0.1}	16.6 _{0.3} 23.1 _{0.3}	16.8 _{0.6} 19.6 _{0.5}	66.4 _{0.5} 67.5 _{0.4}	55.1 _{0.5} 53.9 _{1.0}	57.5 _{0.3} 61.2 _{0.3}	55.9 _{0.6} 56.1 _{0.6}	44.5 _{0.2} 45.2 _{0.3}

Table 9: 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 Self-calibration	80.2 _{0.2} 80.6 _{0.3}	51.0 _{0.5} 51.1 _{0.3}	81.7 _{0.5} 82.9 _{0.1}	54.7 _{0.1} 54.7 _{0.1}	46.8 _{0.1} 47.5 _{0.2}	39.5 _{0.6} 39.3 _{0.3}	78.3 _{0.0} 78.1 _{0.2}	66.8 _{1.4} 65.8 _{0.8}	77.7 _{0.2} 78.1 _{0.5}	75.8 _{0.6} 75.6 _{0.7}	65.3 _{0.2} 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 Self-calibration	79.5 _{0.4} 79.6 _{0.3}	50.4 _{0.3} 51.6 _{0.6}	80.9 _{1.1} 82.0 _{0.7}	54.4 _{0.1} 54.6 _{0.2}	45.8 _{0.6} 46.9 _{0.8}	38.6 _{0.6} 39.0 _{0.4}	78.2 _{0.5} 77.9 _{0.3}	63.4 _{0.9} 64.5 _{0.8}	77.5 _{0.5} 78.2 _{0.6}	74.7 _{0.8} 75.6 _{0.7}	64.3 _{0.1} 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 Self-calibration	70.5 _{1.0} 71.2 _{0.3}	37.2 _{0.8} 37.7 _{0.5}	64.6 _{0.9} 73.4 _{1.0}	40.6 _{0.3} 41.6 _{0.3}	24.0 _{0.5} 31.6 _{0.8}	27.6 _{1.2} 32.1 _{0.8}	71.2 _{0.4} 72.0 _{0.5}	56.0 _{1.9} 65.5 _{2.8}	66.0 _{0.4} 71.0 _{0.4}	65.8 _{0.9} 68.2 _{0.5}	52.3 _{0.2} 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 Self-calibration	66.0 _{0.9} 67.6 _{0.3}	31.5 _{0.7} 35.1 _{0.6}	66.6 _{2.1} 68.8 _{2.0}	38.2 _{0.2} 39.5 _{0.1}	16.6 _{0.7} 18.7 _{0.5}	23.5 _{0.5} 25.8 _{0.4}	69.5 _{0.4} 70.1 _{0.3}	53.9 _{2.1} 59.1 _{3.5}	64.3 _{0.5} 65.7 _{0.3}	64.4 _{0.4} 64.9 _{1.3}	49.4 _{0.4} 51.5 _{0.7}

Table 10: 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}
GPTQ	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}
	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}
SparseGPT	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}
	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}
Wanda	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}
	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}
Wanda	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 11: 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}
GPTQ	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}
	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}
SparseGPT	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}
	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}
Wanda	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}
	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}
Wanda	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 12: 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 13: Task accuracy across five calibration sets for Llama 3.1 8B, with standard deviation denoted in subscript.