7.1 Strided Log-Likelihoods

When computing the statistic, we compute log probabilties under the model of sequences longer than the context length. To do this, we use a strided window approach and use a stride equal to half the context length. We find that this optimal because lower strides provide diminishing returns given that decreasing the stride increases the compute time required to compute model log probabilties.

7.2 Pretraining Details

In this section, we provide additional details on the training procedure for our 1.4B language model trained from scratch on Wikitext with intentional contamination. This model uses a GPT-2 architecture with 1.4B parameters, with the architecture hyperparameters given by a hidden dimension of 1536, 24 heads, 48 layers, sequence length of 2048. The training batch size was 256 and based on the number of training tokens, sequence length, and training batch size, we trained this model for 46000 steps as the total tokens should equal the product of the number of train steps, training batch size, and sequence length. Finally, the model is optimized with an AdamW Optimizer with a learning rate of $1E-4$ and weight decay of 0.1. To train the model, we used the Levanter framework on X TPUs for Y days (CITE).

7.3 10 Canary Datasets

In this section we provide additional details on the 10 canary datasets that were injected into the standard pretraining data (Wikitext, taken from the RedPajama corpus). For BoolQ[1] (Clark et al., 2019), HellaSwag[2] (Zellers et al., 2019), MNLI[3] (Williams et al., 2018), Natural Questions[4] (Kwiatkowski et al., 2019), TruthfulQA[5] (Lin et al., 2022), PIQA[6] (Bisk et al., 2019), we sample a random subset of 1000 examples. For OpenbookQA[7] (Mihaylov et al., 2018), because of its smaller test set of size n=500, we used all 500 examples. Finally, for MMLU[8] (Hendrycks et al., 2021), we selected the subsets that did not contain multi-line examples and had more examples, specifically Professional Psychology (n=611), MMLU Professional Law (n=1000), MMLU High School Psychology (n=544). Finally, we shuffle the examples for all datasets to make them exchangable. In Table 3, we provide additional information about the injected datasets including number of examples, average words per example, and number of tokens per dataset. For each duplication rate in the high duplication rate settin (1, 10, 50, 100) we included a short, medium and longer dataset for 1, 10, and 50. We compute the average words per example mulitply it by the number of examples to estimate the total tokens per dataset. Based on the total tokens, we estimate the dataset length and duplicate it by a certain amount. For pretraining dataset with high duplication rates, the total token count is 19.235M tokens calculated by multiplying the duplication rate of the dataset and the number of tokens per instance. This means that the injected dataset is 0.1% of the entire pre-training dataset.

7.4 Expanded LLaMA2 and MMLU Results

We list the results of our test on 49 of 58 test sets in MMLU. We find p-values lower than 0.05 on 12 test sets, but rule out 10 of these as invalid due to suspected non-exchangeability.

[1] https://github.com/google-research-datasets/boolean-questions

[2] https://rowanzellers.com/hellaswag/

[3] https://cims.nyu.edu/~sbowman/multinli/

[4] https://github.com/google-research-datasets/natural-questions

[5] https://github.com/sylinrl/TruthfulQA/blob/main/data/finetune_truth.jsonl

[6] https://yonatanbisk.com/piqa/

[7] https://allenai.org/data/open-book-qa

[8] https://github.com/hendrycks/test

Table 3: We report the information about the injected datasets as this informed how often we duplicated each dataset in the pretraining data.

Name	Examples	Avg Words/Ex	Tokens	Dup Rate (High)	Dup Rate (Low)
BoolQ	1000	110	110k	1	1
HellaSwag	1000	185	185k	1	1
OpenbookQA	500	40	20k	1	2
Natural Questions	1000	32	32k	10	2
MNLI	1000	235	235k	10	4
TruthfulQA	1000	25	25k	10	4
PIQA	1000	50	50k	50	7
MMLU Pro. Law	1000	2000	200k	50	7
MMLU Pro. Psych	611	50	30k	50	10
MMLU H.S. Psych	544	37	20k	100	10

Table 4: Significant Results on LlaMA2 with MMLU.

Dataset	LLaMA2 p-value
college-computer-science-test	7.35e-08
college-mathematics-test	5.16e-04
econometrics-test	5.28e-04
formal-logic-test	1.73e-06
high-school-computer-science-test	2.99e-09
high-school-european-history-test	1.64e-10
high-school-us-history-test	1.25e-08
high-school-world-history-test	2.30e-06
jurisprudence-test	9.48e-03
nutrition-test	1e-38

Table 5: Non-Significant Results on LlaMA2 with MMLU.

Dataset	LLaMA2 p-value
abstract-algebra-test	1.03e-01
anatomy-test	5.86e-01
astronomy-test	5.50e-01
business-ethics-test	9.36e-01
clinical-knowledge-test	1.99e-01
college-biology-test	9.30e-02
college-chemistry-test	4.82e-01
college-medicine-test	1.49e-01
college-physics-test	6.94e-01
computer-security-test	1.18e-01
conceptual-physics-test	5.54e-01
electrical-engineering-test	2.66e-01
global-facts-test	7.79e-01
high-school-biology-test	8.18e-01
high-school-chemistry-test	2.29e-01
high-school-geography-test	1.94e-01
high-school-government-and-politics-test	3.81e-01
high-school-macroeconomics-test	5.43e-01
high-school-mathematics-test	4.73e-01
high-school-microeconomics-test	9.38e-01
high-school-physics-test	1.70e-01
high-school-psychology-test	8.54e-01
high-school-statistics-test	2.05e-01
human-aging-test	8.82e-01
human-sexuality-test	8.07e-01
international-law-test	6.12e-02
logical-fallacies-test	3.88e-01
machine-learning-test	5.03e-01
management-test	5.16e-01
marketing-test	8.74e-01
medical-genetics-test	5.01e-01
miscellaneous-test	1.24e-01
moral-disputes-test	3.04e-01
moral-scenarios-test	6.52e-01
philosophy-test	1.84e-01
prehistory-test	3.25e-01
professional-accounting-test	5.12e-01