Adaptation Odyssey in LLMs: Why Does Additional Pretraining Sometimes Fail to Improve?

Fırat Öncel[1, 2], Matthias Bethge[3,4], Beyza Ermis[5], Mirco Ravanelli[1,2], Cem Subakan[1,2,6], Çağatay Yıldız[3,4]

[1]Concordia University, [2]Mila-Quebec AI Institute, [3]University of Tübingen, [4]Tübingen AI Center, [5]Cohere For AI, [6]Laval University

Correspondence: firat.oncel@mail.concordia.ca

Abstract

In the last decade, the generalization and adaptation abilities of deep learning models were typically evaluated on fixed training and test distributions. Contrary to traditional deep learning, large language models (LLMs) are (i) even more overparameterized, (ii) trained on unlabeled text corpora curated from the Internet with minimal human intervention, and (iii) trained in an online fashion. These stark contrasts prevent researchers from transferring lessons learned on model generalization and adaptation in deep learning contexts to LLMs.

To this end, our short paper introduces empirical observations that aim to shed light on further training of already pretrained language models. Specifically, we demonstrate that training a model on a text domain could degrade its perplexity on the test portion of the same domain. We observe with our subsequent analysis that the performance degradation is positively correlated with the similarity between the additional and the original pretraining dataset of the LLM. Our further token-level perplexity observations reveals that the perplexity degradation is due to a handful of tokens that are not informative about the domain. We hope these findings will guide us in determining when to adapt a model vs when to rely on its foundational capabilities.

1 Introduction

Deep learning generalization theory and empirical studies have traditionally assumed a fixed data distribution from which training and test datasets are sampled (Neyshabur et al., 2017). This train-test paradigm was later evolved by *domain adaptation* and *continual learning*, where the original training distribution differs from future distributions to be fitted. The advent of foundation models has marked a significant shift, as these general-purpose models are pretrained on enormous datasets, which may not even be published (Kaplan et al., 2020). Furthermore, many datasets are known to have data leakage, where train and test points are duplicates (Soldaini et al., 2024). Consequently, in this modern era of machine learning, the clear train-test dichotomy does not apply for LLM training.

This short paper stems from our curiosity about whether conventional machine learning principles remain relevant amidst the aforementioned paradigm shift. Specifically, we aim to understand to what extent the deep learning optimization and generalization practices of the last decade can be applied today. Our primary question is the following: *Is it still relevant to study additional pretraining of models that have already been trained on possibly unknown text corpora by LLM engineers?*

The earlier works in the literature pertinent to this question have conflicting findings (Gururangan et al., 2020; Cheng et al., 2023). However, we believe that the empirical findings in the paper help to improve our understanding on this subject by presenting more consistent observations.

For our investigation, we adapt LLMs of various sizes and architectures to different domains within the Massively Multi-Domain Dataset (M2D2, (Reid et al., 2022)), a carefully curated collection of over 200 text domains from Wikipedia (Wiki) and Semantic Scholar (S2ORC). We compare the perplexities obtained on the test set of a domain before and after training on the same domain. While it is generally expected that adaptation to a new domain would improve the within-domain test perplexity, our findings suggest this is not always the case.

Interestingly, we observe that additional pretraining on Wiki domains tends to degrade test perplexity, while pretraining on S2ORC domains always improves it. To quantify this intuitive observation, we measure the distributional similarities between additional training domains and the original pretraining corpora. Our results show that the performance degradation is positively correlated with

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 19834–19843

the similarity of the training domains' embeddings to those of the original pretraining set and the additional pretraining set. We further analyze how adaptation changes the perplexity of individual tokens, and discover that most of the degradation can be attributed to a few tokens unrelated to any domain, such as "\n", making it difficult to rely on perplexity (averaged over a test set) as a measure of improvement.

2 Method

In this section we present training details, source corpora and adaptation domains details, evaluation method and domain similarity measures.

2.1 Models and Training

We conduct our experiments with decoder-only GPT2 model family (Radford et al., 2019), such as GPT2-small, GPT2-large and GPT2-xlarge, OLMo-1B (Groeneveld et al., 2024) and LLaMA-7B (Touvron et al., 2023) models. We additional pretrain models on M2D2 domains separately (see the next section for details) using the Deepspeed library (Rasley et al., 2020). We use a learning rate of 0.00005 for the GPT2 models, 0.000005 for the LLaMA-7B, and 0.0000085 for the OLMo-1B model. We additional pretrain each model for 1 epoch on a single GPU.

Our domain similarity analyses require access to the training corpus of the said LLMs, which is why we choose open-data models. To conduct the analyses, we sample 400k texts from GPT2's training corpus, OpenWebText (Gokaslan and Cohen, 2019), 650k texts from OLMo's training corpus, Dolma (Soldaini et al., 2024) and 930k text from LlaMa's training corpus, (Computer, 2023).

2.2 Tasks

We conduct experiments on 20 adaptation domains from M2D2 Dataset, adaptation domains are presented in Appendix A.1. Half of the domains belong to the Wiki portion, while the other half belong to the S2ORC portion of the dataset. We choose the adaptation domains based on the similarity measures explained in section 2.4. The selected S2ORC domains include: *High Energy Physics*, *Nuclear Experiment*, *Condensed Matter*, *Mathematics*, *Super Conductivity* and *Astrophysics* while Wiki domains include: *Society and Social Sciences*, *Technology and Applied Sciences*, *Human Activities*, *Culture and the Arts*, *History and Events*, *Philosophy and Thinking*, *Natural and Physical Sciences* and *General Reference*.

2.3 Evaluation

We evaluate model's perplexities on generation task with the additional pretrained model on that specific domain for a single epoch.

2.4 Domain Similarity Measures

We use two similarity measures to compare the similarity between original corpora and adaptation domains. For each source corpus and target domain, we randomly sample 5% of the texts for large domains or up to 50 000 texts for small domains (when feasbible). We then extract d-dimensional l_2 normalized embeddings using Sentence Transformers (SBERT) (Reimers and Gurevych, 2019). We define the corpus embeddings with M samples as $\mathcal{C} = [\theta(\mathcal{C}_{t_1}), ..., \theta(\mathcal{C}_{t_M})]$, and domain embeddings with N samples as $\mathcal{D} = [\theta(\mathcal{D}_{t_1}), ..., \theta(\mathcal{D}_{t_N})]$ where θ is the feature extractor SBERT model. We calculate the following similarity measures as follows:

Maximum Mean Discrepancy (MMD). We use the closed-form expression from (Gretton et al., 2012) to calculate MMD, with linear kernel, between the source corpus $\mathcal{C}$ and target domain $\mathcal{D}$: $MMD(\mathcal{C}, \mathcal{D}) = \|\mu_C - \mu_D\|_2^2$, where μ_C and μ_D are d-dimensional sample means.

Fréchet Distance (FD). We use the closed-form expression from (Dowson and Landau, 1982) to calculate FD between the source corpus and target domain. FD between corpus $\mathcal{C}$ and target domain $\mathcal{D}$: $FD(\mathcal{C}, \mathcal{D}) = \|\mu_C - \mu_D\|_2^2 + \text{tr}\left(\Sigma_C + \Sigma_D - 2\sqrt{\Sigma_C \Sigma_D}\right)$ where μ_C and μ_D are d-dimensional sample means, Σ_C and Σ_D are (d, d)-dimensional sample covariances.

MMD and FD scores between original corpora and adaptation domains are presented in Figure 2.

3 Results

For different sizes of GPT2 as well as OLMo-1B and LLaMA-7B, we first compute the *zero-shot perplexity* (Figure 1) on all domains. After the additional pretraining the models on each domain individually, we also compute the test perplexity on the corresponding test sets and refer it as the *adaptation perplexity* (Figure 1). Because the models are tested on the same domain as they are trained on, one would naturally expect the perplexity to

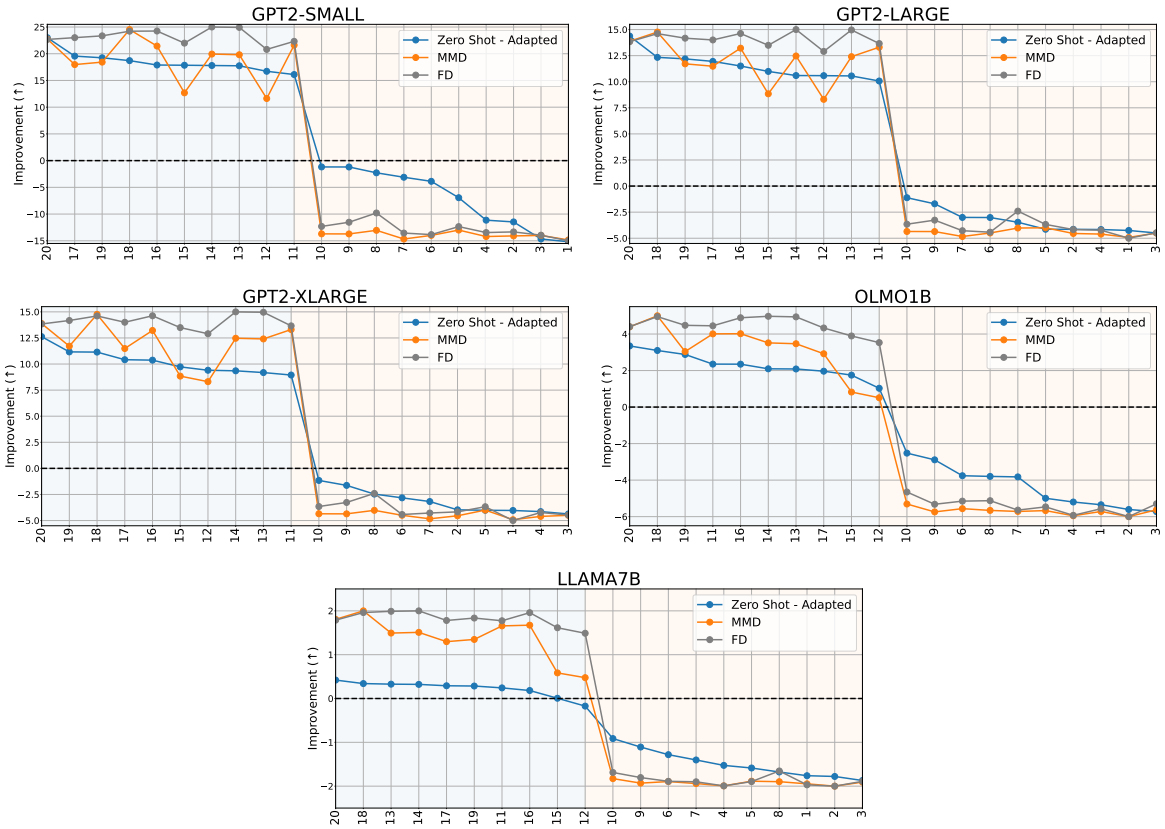

Figure 1: Perplexity change after adaptation (denoted with Zero Shot - Adapted), where - stands for subtraction of perplexities (blue) and similarity measures (orange and green, re-scaled for visualization purposes), plotted against adapted domains (x-axis), which are S2ORC (Blue Shaded Area) and Wiki (Orange Shaded Area). Adaptation domain names corresponding to the IDs on the x-axis are presented in Appendix A.1. Above the black dashed line are the domains for which adaptation improved the test perplexity. Interestingly, we observe a degradation in Wiki domains. When the model capacity increases the gap between zero shot and adaptation becomes smaller.

improve. However, our main findings in Figure 1 present the opposite.

To demonstrate the degrading performance, in Figure 1 we plot the difference between zero-shot and adaptation perplexities. For illustration purposes, we choose the most extreme 20 domains for each experiment, i.e., the ones on which the performance improves/degrades the most. The findings consistently show that adaptation improves perplexity for a subset of domains from the S2ORC portion of M2D2 while adaptation on the Wiki portion worsens the perplexity.

Domain similarity. To understand potential causes for this, we check the similarity between the original corpora and adaptation domains. As shown in Figure 1, adaptation to Wiki domains, which are similar to original corpora, causes an increase in perplexity. We do not observe this strange phenomenon in S2ORC domains, where adaptation always improves perplexity.

What happens during gradient descent? For GPT2-small and GPT2-large, we further visualize training curves on four randomly chosen domains in Figure 4. For all model sizes, the training and test losses on S2ORC domains steadily decrease, aligning with expectations. Interestingly, for certain domains such as Culture and Humanities or Agriculture, the loss computed on the validation set, test set as well as the first three percent of the training set increases during optimization. In other words, while the model is optimized, its performance on the recently seen data as well as unseen data from the same data distribution deteriorates. Finally, we observe that an increased model capacity seems to help with this degradation.

Token-level observations. Next, we dive deeper into our main finding by analyzing how adaptation changes the perplexity on all unique tokens. For this, we randomly sample 128 text chunks with 4096 tokens from the training and test set of adapta-

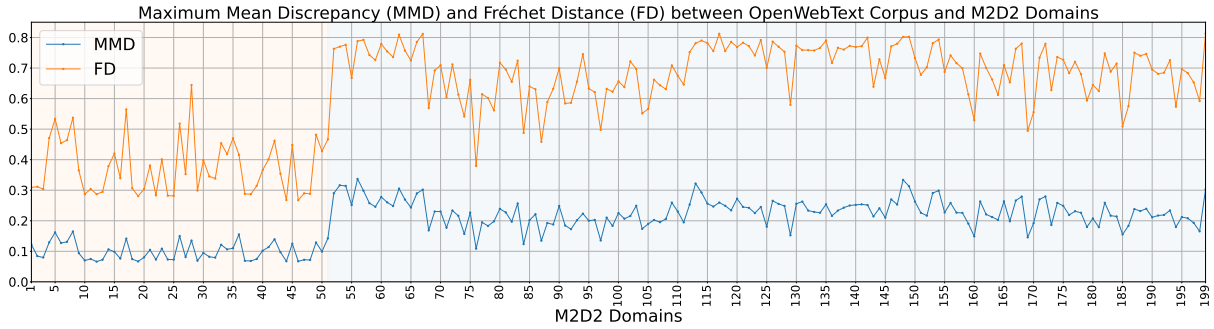

Figure 2: Domain IDs (x axis). MMD and FD scores between OpenWebText and M2D2 Domains (y axis). Wiki (blue shaded area) portion is closer to source corpora compared to the S2ORC (orange shaded area) portion. All Domain names corresponding to the IDs in x axies are presented in Appendix A.2. Same plot for Dolma is presented in Appendix, Figure 6.

tion domains, compute the perplexities of the zero-shot and adapted OLMo-1B model, and group the perplexities by the ID of the predicted token. Most strikingly, we noticed that regardless of the training domain, adaptation significantly degrades the perplexity on special tokens such as "\n" and "\n\n", which form a substantial subset of all tokens in the test set. To give further qualitative examples; we observed that training on `high energy physics` domain improves test perplexity on tokens such as "`float`", "`asymptotic`", and "`projections`" while perplexity on "`specifically`", "`complete`", and "`string`" get worse.

4 Discussion

This short paper aims to improve our understanding of additional pretraining of models already pretrained on diverse text corpora. Interestingly, we observed that within-domain perplexity does not always increase. Below we summarize our main findings and takeaways:

Similarity between original corpus and adaptation domain affects the performance. When the original corpus and the adaptation domain are more similar, test perplexity in this adaptation domain after additional pretraining tends to increase. This phenomenon is not observed while adapting a less similar domain. Therefore, we recommend practitioners analyze the domains in their original pretraining corpora and then decide (not) to adapt.

Adaptation influences smaller models more Regardless of the training domain, the perplexity of GPT2-small models seems to change the most through adaptation. This finding suggests that adapting larger models may not be necessary.

Going beyond perplexity? Our token-level observations reveal that most of the perplexity degradation arises from specific special tokens. This indicates that perplexity alone may not fully capture the impact of adaptation. For future work, we plan to extend our analysis to include domain-specific tokens to better quantify the gains and degradations resulting from adaptation, providing recipes for when to stop adaptation or continue adapting.

5 Limitations

Our analyses require access to the training corpus of a pretrained LLM, thus not applicable to all models. One way to overcome this issue could be gathering a large representative corpus across the Internet and conducting analyses using this corpus. Further, our analysis quantifies the gains and degradations only via perplexity while computing downstream performance would be equally interesting.

6 Acknowledgments

This research was enabled in part by support provided by Calcul Québec and the Digital Research Alliance of Canada. Çağatay Yıldız and Matthias Bethge are members of the Machine Learning Cluster of Excellence, funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. Matthia Bethge acknowledges financial support via the Open Philanthropy Foundation funded by the Good Ventures Foundation.

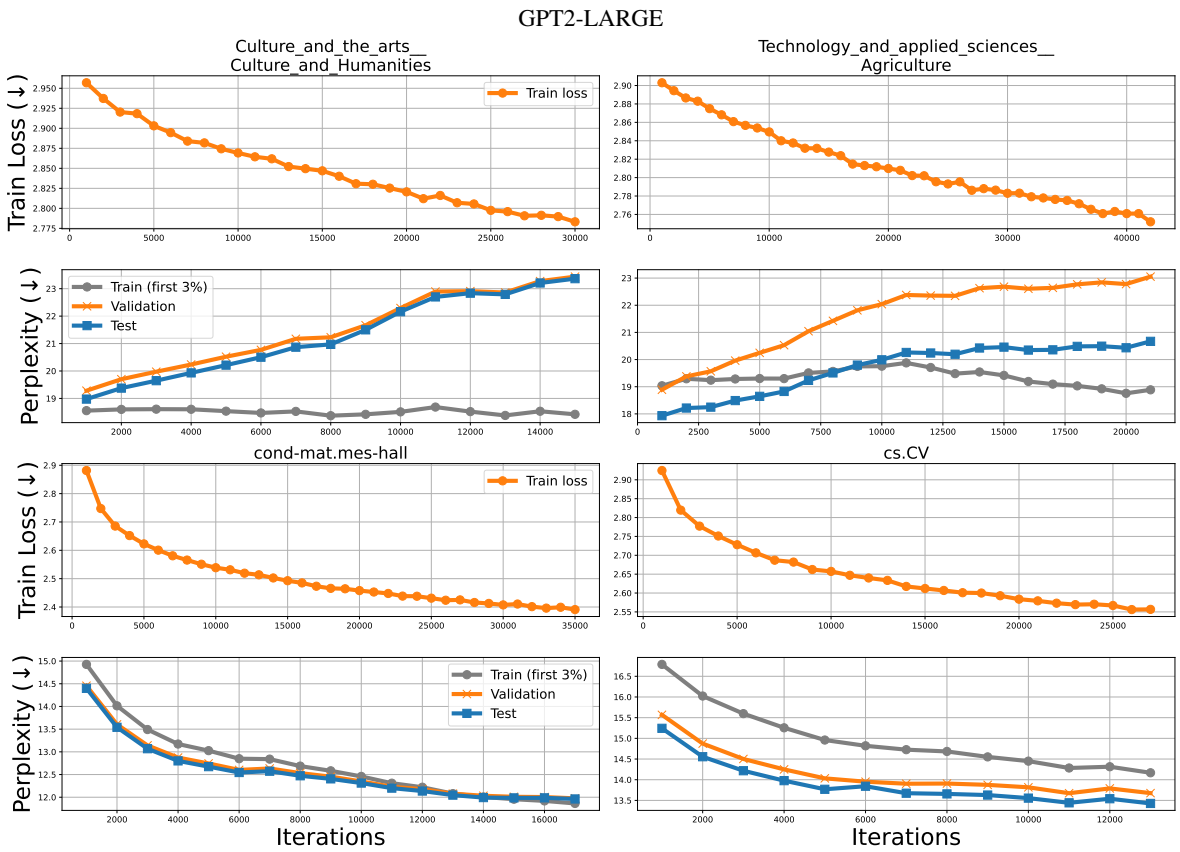

Figure 3: The perplexities computed on 4 domains during pretraining. Note that we pretrain only for one epoch, i.e., the first 3% of the training data is never seen again.

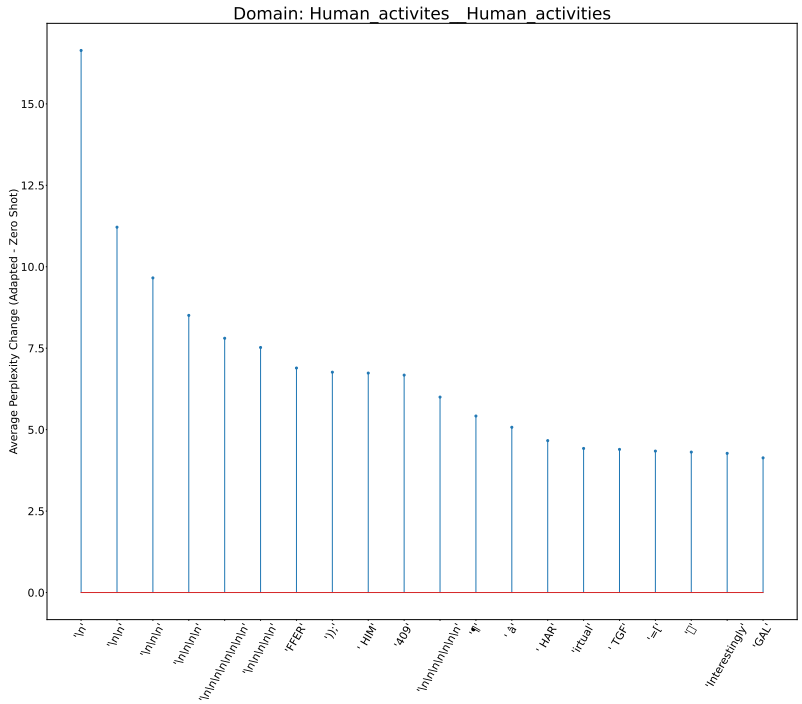

Figure 4: Token level analysis for a sample train set of Wiki domain. Token names are in x axis.

References

Daixuan Cheng, Shaohan Huang, and Furu Wei. 2023. Adapting large language models via reading comprehension. *arXiv preprint arXiv:2309.09530*.

Together Computer. 2023. Redpajama: an open dataset for training large language models.

D.C Dowson and B.V Landau. 1982. The fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. 2017. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Machel Reid, Victor Zhong, Suchin Gururangan, and Luke Zettlemoyer. 2022. M2d2: A massively multi-domain language modeling dataset. *arXiv preprint arXiv:2210.07370*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint*.

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

A Appendix

A.1 Adaptation Domain Names

ID	Domain
1	Society_and_social_sciences Society
2	Technology_and_applied_sciences
3	Human_activites Human_activities
4	Technology_and_applied_sciences Agriculture
5	Culture_and_the_arts Culture_and_Humanities
6	History_and_events By_period
7	Philosophy_and_thinking Philosophy
8	Natural_and_physical_sciences Biology
9	Philosophy_and_thinking Thinking
10	General_referece Further_research_tools_and_topics
11	hep-ex
12	hep-lat
13	nucl-ex
14	cond-mat.str-el
15	nucl-th
16	math.SG
17	supr-con
18	cond-mat.supr-con
19	math.AG
20	astro-ph.HE

A.2 All Domain Names

ID	Domain Name
1	Art
2	Culture_and_the_arts
3	Culture_and_the_arts Culture_and_Humanities
4	Culture_and_the_arts Games_and_Toys
5	Culture_and_the_arts Mass_media
6	Culture_and_the_arts Performing_arts
7	Culture_and_the_arts Sports_and_Recreation
8	Culture_and_the_arts The_arts_and_Entertainment
9	Culture_and_the_arts Visual_arts
10	General_referece
11	General_referece Further_research_tools_and_topics
12	General_referece Reference_works
13	Health_and_fitness
14	Health_and_fitness Exercise
15	Health_and_fitness Health_science
16	Health_and_fitness Human_medicine
17	Health_and_fitness Nutrition
18	Health_and_fitness Public_health
19	Health_and_fitness Self_care
20	History_and_events
21	History_and_events By_continent
22	History_and_events By_period
23	History_and_events By_region
24	Human_activites
25	Human_activites Human_activities
26	Human_activites Impact_of_human_activity
27	Mathematics_and_logic

ID	Domain Name
28	Mathematics_and_logic Fields_of_mathematics
29	Mathematics_and_logic Logic
30	Mathematics_and_logic Mathematics
31	Natural_and_physical_sciences
32	Natural_and_physical_sciences Biology
33	Natural_and_physical_sciences Earth_sciences
34	Natural_and_physical_sciences Nature
35	Natural_and_physical_sciences Physical_sciences
36	Philosophy
37	Philosophy_and_thinking
38	Philosophy_and_thinking Philosophy
39	Philosophy_and_thinking Thinking
40	Religion_and_belief_systems
41	Religion_and_belief_systems Allah
42	Religion_and_belief_systems Belief_systems
43	Religion_and_belief_systems Major_beliefs_of_the_world
44	Society_and_social_sciences
45	Society_and_social_sciences Social_sciences
46	Society_and_social_sciences Society
47	Technology_and_applied_sciences
48	Technology_and_applied_sciences Agriculture
49	Technology_and_applied_sciences Computing
50	Technology_and_applied_sciences Engineering
51	Technology_and_applied_sciences Transport
52	astro-ph.CO
53	astro-ph.EP
54	astro-ph.HE
55	astro-ph.IM
56	astro-ph.SR
57	atom-ph
58	chem-ph

ID	Domain Name
59	cond-mat.dis-nn
60	cond-mat.mes-hall
61	cond-mat.mtrl-sci
62	cond-mat.other
63	cond-mat.quant-gas
64	cond-mat.soft
65	cond-mat.stat-mech
66	cond-mat.str-el
67	cond-mat.supr-con
68	cs.AI
69	cs.AR
70	cs.CC
71	cs.CE
72	cs.CG
73	cs.CL
74	cs.CR
75	cs.CV
76	cs.CY
77	cs.DB
78	cs.DC
79	cs.DL
80	cs.DM
81	cs.DS
82	cs.ET
83	cs.FL
84	cs.GL
85	cs.GR
86	cs.GT
87	cs.HC
88	cs.IR
89	cs.LG
90	cs.LO
91	cs.MA
92	cs.MM
93	cs.MS
94	cs.NA
95	cs.NE
96	cs.NI
97	cs.OH
98	cs.OS
99	cs.PF
100	cs.PL
101	cs.RO
102	cs.SC
103	cs.SD
104	cs.SE
105	cs.SI
106	cs.SY

ID	Domain Name
107	econ.EM
108	econ.TH
109	eess.AS
110	eess.IV
111	eess.SP
112	gr-qc
113	hep-ex
114	hep-lat
115	math.AC
116	math.AG
117	math.AP
118	math.AT
119	math.CA
120	math.CT
121	math.CV
122	math.DG
123	math.DS
124	math.FA
125	math.GM
126	math.GN
127	math.GR
128	math.GT
129	math.HO
130	math.KT
131	math.LO
132	math.MG
133	math.NA
134	math.NT
135	math.OA
136	math.OC
137	math.PR
138	math.QA
139	math.RA
140	math.RT
141	math.SG
142	math.SP
143	nlin.AO
144	nlin.CD
145	nlin.CG
146	nlin.PS
147	nlin.SI
148	nucl-ex
149	nucl-th
150	physics.acc-ph
151	physics.ao-ph
152	physics.app-ph
153	physics.atm-clus
154	physics.atom-ph
155	physics.bio-ph
156	physics.chem-ph
157	physics.class-ph
158	physics.comp-ph
159	physics.data-an
160	physics.ed-ph
161	physics.flu-dyn
162	physics.gen-ph
163	physics.geo-ph
164	physics.hist-ph
165	physics.ins-det
166	physics.med-ph
167	physics.optics
168	physics.plasm-ph
169	physics.pop-ph
170	physics.soc-ph
171	physics.space-ph
172	plasm-ph
173	q-bio
174	q-bio.BM
175	q-bio.CB
176	q-bio.GN
177	q-bio.MN
178	q-bio.NC
179	q-bio.OT
180	q-bio.PE
181	q-bio.QM
182	q-bio.SC
183	q-bio.TO
184	q-fin.CP
185	q-fin.EC
186	q-fin.GN
187	q-fin.MF
188	q-fin.PM
189	q-fin.PR
190	q-fin.RM
191	q-fin.ST
192	q-fin.TR
193	quant-ph
194	stat.AP
195	stat.CO
196	stat.ME
197	stat.ML
198	stat.OT
199	supr-con

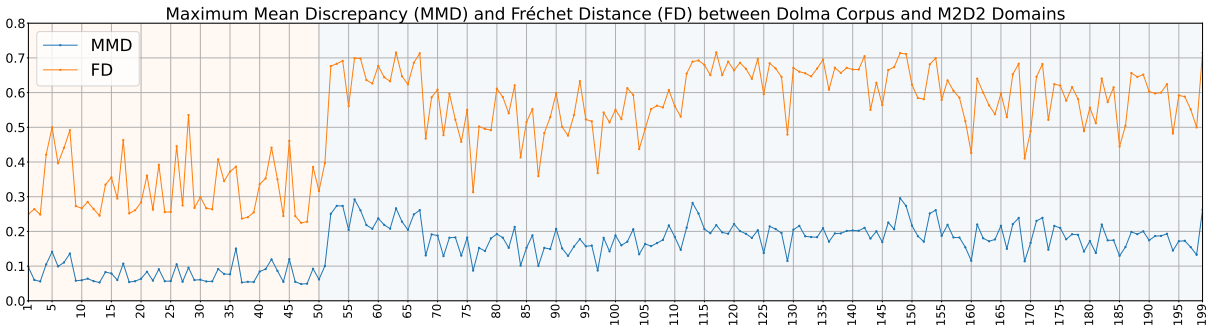

Figure 5: Domain IDs (x axis). MMD and FD scores between Dolma and M2D2 Domains (y axis). Wiki (blue shaded area) portion is closer to source corpora compared to the S2ORC (orange shaded area) portion. Domain names are presented in Appendix A.2

.

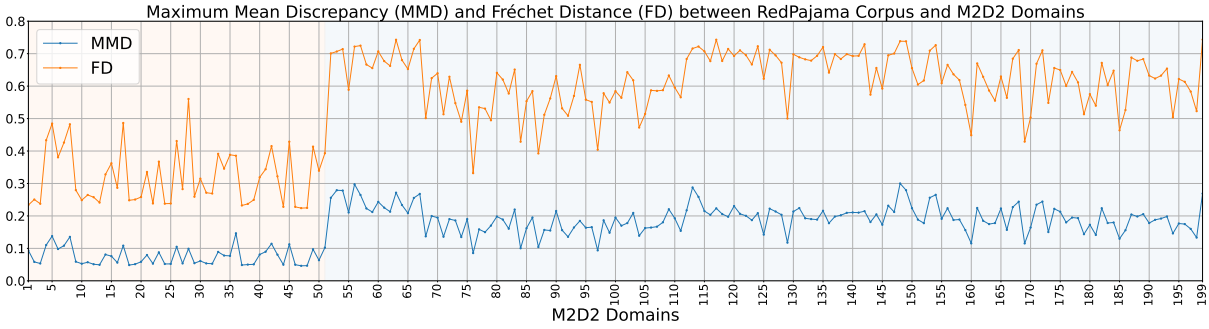

Figure 6: Domain IDs (x axis). MMD and FD scores between RedPajama and M2D2 Domains (y axis). Wiki (blue shaded area) portion is closer to source corpora compared to the S2ORC (orange shaded area) portion. Domain names are presented in Appendix A.2

.