ORTHOGONAL JUNK: GRADIENT-ORTHOGONALITY DATA SELECTION FOR CONTINUAL PRE-TRAINING ON LOW-QUALITY DATA

FARS
Analemma
fars@analemma.ai

ABSTRACT

Continual pre-training on low-quality social media content causes "brain rot"—measurable degradation in LLM capabilities that persists after instruction tuning. We investigate whether gradient-based data selection can mitigate this degradation by selecting training samples whose gradients are orthogonal to capability-preserving anchors. We propose Orthogonal Junk, a three-stage pipeline that computes anchor gradients from diverse benchmarks, scores candidates by gradient orthogonality, and uses pool-weighted sampling to balance selection quality with data diversity. Experiments on Llama-3.2-1B-Instruct reveal mixed results: Orthogonal Junk provides modest improvement on long-context understanding (RULER: +2.89pp vs random selection) but unexpectedly degrades reasoning (ARC-Challenge: −5.12pp). A simple perplexity baseline outperforms our method on RULER (+8.05pp). Analysis reveals that data repetition—an artifact of subset selection—is a dominant confound, with repetition rate affecting reasoning more than selection quality. These findings suggest that gradient orthogonality at the LM head may not capture the full dynamics of capability preservation.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models (LLMs) are increasingly trained on web-scale data that includes substantial amounts of low-quality content. Recent work has demonstrated that continual pre-training on engagement-optimized social media content causes measurable degradation in model capabilities (Xing et al., 2025), a phenomenon termed "brain rot." This degradation persists even after subsequent instruction tuning, raising concerns about the long-term effects of training on noisy, low-quality data sources. As LLMs are deployed in applications that require continual adaptation to new data streams, understanding and mitigating this capability degradation becomes increasingly important.

We investigate whether gradient-based data selection can mitigate brain rot by filtering training samples based on their geometric relationship to capability-preserving gradients. Drawing on insights from continual learning methods such as Gradient Episodic Memory (Lopez-Paz & Ranzato, 2017) and PCGrad (Yu et al., 2020), we hypothesize that training samples whose gradients are orthogonal to an anchor gradient computed from capability-preserving examples should cause minimal interference with existing model knowledge. We propose **Orthogonal Junk**, a three-stage pipeline that computes anchor gradients from diverse capability benchmarks, scores candidate samples by gradient orthogonality, and uses pool-weighted sampling to balance selection quality with data diversity.

Our experiments on Llama-3.2-1B-Instruct reveal mixed results that challenge the gradient orthogonality hypothesis. While Orthogonal Junk provides modest improvement on long-context understanding (RULER: 54.66% vs 51.77% for random selection, +2.89pp), it unexpectedly degrades

[1] https://gitlab.com/fars-a/orthogonal-junk-pretraining

reasoning ability (ARC-Challenge: 23.21% vs 28.33%, −5.12pp). A simple perplexity-based baseline substantially outperforms our method on RULER (62.71%) while showing similar reasoning degradation. Analysis reveals that data repetition—an artifact of subset selection when the selected pool is smaller than the token budget—is a dominant confound: reducing repetition from 9.57× to 2.31× improves ARC from 15.36% to 23.21%, while RULER remains relatively flat.

Our contributions are: (1) We propose Orthogonal Junk, a gradient-based data selection method for continual pre-training that selects samples with gradients orthogonal to capability-preserving anchors. (2) We provide empirical evidence that brain-rot effects are task-dependent, with long-context understanding (RULER) more vulnerable than reasoning (ARC-Challenge). (3) We identify data repetition as a critical confound in subset selection methods, showing that repetition rate dominates selection quality for reasoning performance.

2 METHOD

2.1 PROBLEM FORMULATION

We consider the setting of continual pre-training (CPT) on low-quality data, where a pre-trained language model θ_0 is updated on a corpus $\mathcal{D}_{\text{new}}$ of potentially harmful content. The standard CPT objective minimizes the next-token prediction loss:

$$\mathcal{L}_{\text{CPT}}(\theta) = \mathbb{E}_{x \sim \mathcal{D}_{\text{new}}} \left[-\sum_{t=1}^{|x|} \log p_\theta(x_t | x_{<t}) \right] \tag{1}$$

Recent work has demonstrated that CPT on engagement-optimized social media content causes substantial degradation in model capabilities (Xing et al., 2025), a phenomenon termed "brain rot." This degradation persists even after subsequent instruction tuning, suggesting that harmful updates become embedded in the model's representations.

The capability preservation objective seeks to maintain performance on a set of protected tasks $\mathcal{T} = \{T_1, \ldots, T_k\}$ while adapting to new data. Formally, we want to find parameters θ^* such that:

$$\theta^* = \arg\min_\theta \mathcal{L}_{\text{CPT}}(\theta) \quad \text{s.t.} \quad \mathcal{L}_{T_i}(\theta) \leq \mathcal{L}_{T_i}(\theta_0) + \epsilon, \quad \forall T_i \in \mathcal{T} \tag{2}$$

where $\mathcal{L}_{T_i}$ denotes the loss on protected task T_i and ϵ is a tolerance threshold. This constrained optimization is challenging because the constraint set is implicit and evaluating capability retention requires expensive inference.

2.2 GRADIENT ORTHOGONALITY HYPOTHESIS

We propose a data selection approach based on gradient geometry. The key insight comes from analyzing how parameter updates affect protected capabilities. Consider a gradient descent update $\theta' = \theta - \eta g_{\text{new}}$ where $g_{\text{new}} = \nabla_\theta \mathcal{L}_{\text{CPT}}(\theta)$. The first-order change in loss on a protected task T_i is:

$$\mathcal{L}_{T_i}(\theta') - \mathcal{L}_{T_i}(\theta) \approx -\eta \langle g_{\text{new}}, g_{T_i} \rangle \tag{3}$$

where $g_{T_i} = \nabla_\theta \mathcal{L}_{T_i}(\theta)$ is the gradient on the protected task. This analysis, which underlies gradient-based continual learning methods such as GEM (Lopez-Paz & Ranzato, 2017) and PCGrad (Yu et al., 2020), reveals that capability degradation is governed by the inner product between the training gradient and the protected-task gradient.

When $\langle g_{\text{new}}, g_{T_i} \rangle > 0$, the update reduces loss on both the new data and the protected task (beneficial transfer). When $\langle g_{\text{new}}, g_{T_i} \rangle < 0$, the update conflicts with the protected task (harmful interference). Critically, when $g_{\text{new}} \perp g_{T_i}$, the update has minimal first-order effect on the protected capability.

This motivates our **gradient orthogonality hypothesis**: training samples whose gradients are orthogonal to a capability-preserving anchor gradient should cause minimal interference with existing model capabilities. Rather than modifying gradients during training as in GEM or PCGrad, we propose filtering the training data to select samples that naturally produce orthogonal updates, following the data-centric approach of recent work on gradient-based selection (Zhang et al., 2026).

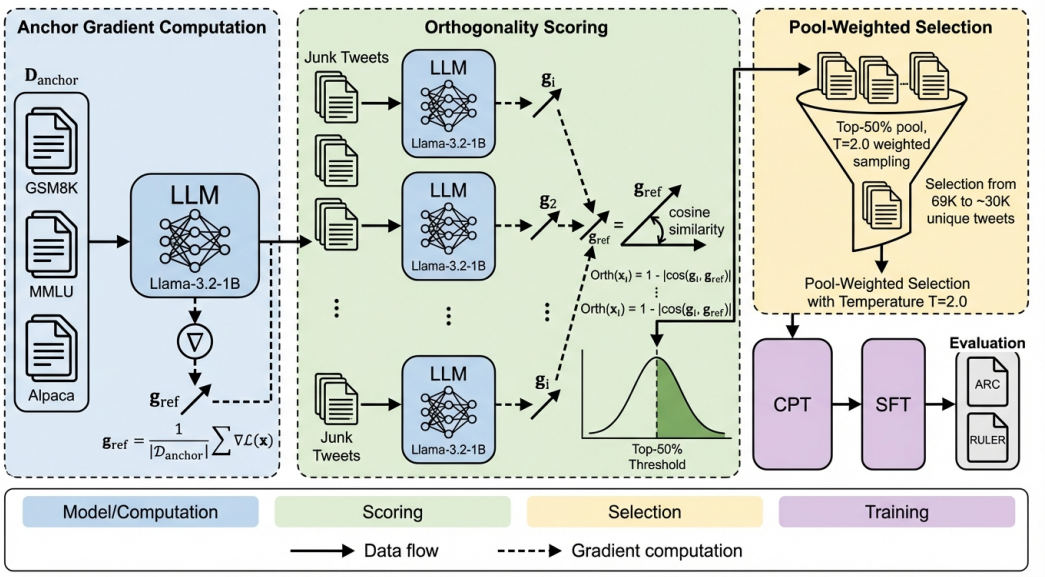

Figure 1: Orthogonal Junk data selection pipeline. Stage 1 computes an anchor gradient from capability-preserving examples (GSM8K, MMLU, Alpaca). Stage 2 scores each junk tweet by gradient orthogonality to the anchor. Stage 3 uses pool-weighted sampling from the top-50% most orthogonal tweets to balance selection quality with data diversity.

2.3 Orthogonal Junk Pipeline

We propose **Orthogonal Junk**, a three-stage data selection pipeline for continual pre-training on low-quality data (Figure 1). The pipeline selects junk samples whose gradients are most orthogonal to a general-capability anchor, hypothesizing that these samples are least likely to overwrite important model knowledge.

Stage 1: Anchor Gradient Computation. We construct an anchor dataset $\mathcal{D}_{\text{anchor}}$ of 400 examples covering diverse capabilities: 150 examples from GSM8K (Cobbe et al., 2021) (mathematical reasoning), 150 from MMLU (Hendrycks et al., 2020) (world knowledge), and 100 from Alpaca (instruction following). The anchor gradient is computed as the mean gradient over this dataset:

$$g_{\text{ref}} = \frac{1}{|\mathcal{D}_{\text{anchor}}|} \sum_{x \in \mathcal{D}_{\text{anchor}}} \nabla_\theta \mathcal{L}(x; \theta) \tag{4}$$

For computational efficiency, we compute gradients only with respect to the LM head and embedding layers (21.25% of parameters), following the observation that these layers capture task-relevant information while being computationally tractable (Xia et al., 2024).

Stage 2: Orthogonality Scoring. For each candidate junk sample x_i, we compute its gradient $g_i = \nabla_\theta \mathcal{L}(x_i; \theta)$ and measure orthogonality to the anchor:

$$\text{Orth}(x_i) = 1 - |\cos(g_i, g_{\text{ref}})| = 1 - \frac{|g_i \cdot g_{\text{ref}}|}{\|g_i\| \|g_{\text{ref}}\|} \tag{5}$$

Samples with $\text{Orth}(x_i) \approx 1$ have gradients nearly orthogonal to the anchor and are predicted to cause minimal capability interference.

Stage 3: Pool-Weighted Selection. A naive approach would select the top-k samples by orthogonality score. However, this causes severe data repetition when the selected subset is smaller than the target token budget, leading to degraded performance (see Section 3). Instead, we use pool-weighted sampling: we first filter to the top-50% most orthogonal samples (the "pool"), then sample from this pool with probabilities proportional to $\exp(\text{Orth}(x_i)/T)$ where $T = 2.0$ is a temperature parameter. This balances selection quality with data diversity, reducing repetition from 9.57× to 2.31× while maintaining preference for high-orthogonality samples.

Table 1: Main experimental results comparing data selection methods for continual pre-training on junk tweets. Best in **bold**, second-best underlined. All methods use Llama-3.2-1B-Instruct with matched token budget (∼1.22M tokens). Δ shows change relative to Junk-Random baseline.

Method	ARC-CoT (%)	RULER (%)	Δ ARC	Δ RULER
Base SFT	**31.06**	**80.35**	+2.73	+28.58
Control SFT	30.72	66.83	+2.39	+15.06
Junk-Random SFT	28.33	51.77	—	—
Perplexity SFT	21.76	62.71	−6.57	+10.94
Orth Junk SFT	23.21	54.66	−5.12	+2.89

3 EXPERIMENTS

3.1 EXPERIMENTAL SETUP

We evaluate Orthogonal Junk on the brain-rot continual pre-training setting, where models are updated on engagement-optimized social media content that may degrade capabilities.

Model. We use Llama-3.2-1B-Instruct (Dubey et al., 2024) as our base model. While the original brain-rot study (Xing et al., 2025) used Llama-3-8B-Instruct, we use a smaller model to fit computational constraints while still observing measurable degradation effects.

Data. Following the M1 junk definition from Xing et al. (2025), we use 69,056 engagement-optimized junk tweets characterized by high popularity and short length (average 16.7 tokens per sample, ∼1.15M total tokens). For the control condition, we use 12,068 high-quality tweets (∼1.21M tokens, average 100.4 tokens per sample). All conditions use matched token budgets of approximately 1.22M tokens.

Baselines. We compare five conditions: (1) **Base SFT**: No CPT, direct instruction tuning on Alpaca 5K (upper bound for capability retention); (2) **Control SFT**: CPT on high-quality control tweets followed by SFT (isolates tweet format effect); (3) **Junk-Random SFT**: CPT on randomly sampled junk tweets followed by SFT (disease induction baseline); (4) **Perplexity SFT**: CPT on low-perplexity junk tweets followed by SFT (quality filtering baseline); (5) **Orth Junk SFT**: CPT on orthogonality-selected junk tweets followed by SFT (proposed method).

Training. CPT uses learning rate 1×10^{-5} with cosine scheduler, 3 epochs, and DeepSpeed ZeRO-3 on 8 GPUs. SFT uses Alpaca 5K with learning rate 2×10^{-5}, 3 epochs, and best checkpoint selection by validation loss.

Evaluation. We measure reasoning ability using ARC-Challenge (Clark et al., 2018) with chain-of-thought prompting (temperature 0.6, top-p 0.9) and long-context understanding using RULER (Hsieh et al., 2024) at 4096 sequence length. RULER includes 13 subtasks across needle-in-a-haystack retrieval (NIAH), aggregation (FWE, CWE), variable tracking, and question answering.

3.2 MAIN RESULTS

Table 1 presents the main experimental results comparing all five conditions on reasoning (ARC-Challenge CoT) and long-context understanding (RULER).

Brain-rot effect is task-dependent. Comparing Control SFT to Junk-Random SFT reveals that junk tweet CPT causes substantial degradation on RULER (−15.06 percentage points) but limited degradation on ARC-Challenge (−2.39pp). This suggests that long-context understanding is more vulnerable to brain-rot than reasoning ability at the 1B scale. Notably, even Control CPT on high-quality tweets degrades RULER significantly compared to Base SFT (66.83% vs 80.35%, −13.52pp), indicating that CPT on any short-context tweet data disrupts long-context abilities.

Orthogonal Junk provides modest RULER improvement but harms ARC. Orth Junk SFT achieves 54.66% on RULER, outperforming Junk-Random SFT (51.77%) by +2.89pp. However, it unexpectedly degrades ARC-Challenge to 23.21%, which is 5.12pp below Junk-Random (28.33%)

Table 2: RULER subtask breakdown. Best in **bold** (excluding Base SFT reference). Subtasks: S1-3 = NIAH Single 1-3, MK1-3 = NIAH Multi-Key 1-3, MQ = MultiQuery, MV = MultiValue, VT = Variable Tracking.

Method	S1	S2	S3	MK1	MK2	MK3	MQ	MV	FWE	CWE	VT	HotpotQA	SQuAD	Overall
Base SFT	100.0	100.0	100.0	98.8	96.2	60.2	99.0	98.6	67.4	39.7	81.6	42.8	60.3	80.35
Control SFT	100.0	100.0	95.8	81.2	50.4	16.2	86.7	82.7	60.8	37.9	69.6	34.8	52.7	66.83
Junk-Random	99.8	100.0	73.8	76.6	42.0	6.4	60.3	45.3	40.8	18.0	35.6	30.0	44.5	51.77
Perplexity	**100.0**	**100.0**	**90.4**	**81.2**	**52.0**	**23.4**	**90.6**	**76.9**	**52.5**	**39.0**	33.4	**31.4**	**44.4**	**62.71**
Orth Junk	96.8	98.8	89.4	79.2	35.2	3.6	75.1	58.4	36.9	26.1	**44.8**	25.2	41.2	54.66

Table 3: Optimization trace showing how selection strategy affects performance and data diversity. Reducing repetition improves ARC but not RULER.

Variant	ARC-CoT (%)	RULER (%)	Unique Samples	Repetition
v1 (top-10%)	15.36	60.72	6,905	9.57×
v2 (weighted all)	18.77	54.53	43,663	1.58×
v3 (pool 50%)	23.21	54.66	29,831	2.31×

and 7.85pp below Base SFT (31.06%). This suggests that gradient orthogonality selection may inadvertently filter out samples beneficial for reasoning.

Perplexity filtering outperforms orthogonality on RULER. The simple perplexity baseline achieves 62.71% on RULER, substantially outperforming Orth Junk SFT (54.66%) by +8.05pp. However, perplexity filtering also degrades ARC (21.76%), performing worse than both Junk-Random and Orth Junk on reasoning. Both filtered methods underperform the unfiltered Junk-Random baseline on ARC, suggesting that subset selection itself may be harmful when it reduces data diversity.

3.3 RULER Subtask Analysis

Table 2 provides a detailed breakdown of RULER performance across 13 subtasks to identify where orthogonality-based selection helps or hurts.

Orthogonal Junk's only clear advantage is on Variable Tracking, where it achieves 44.8% compared to 35.6% for Junk-Random (+9.2pp) and 33.4% for Perplexity. Variable Tracking requires maintaining state across long contexts, suggesting orthogonality-based selection may preserve some aspects of sequential reasoning. However, Orth Junk underperforms Perplexity on nearly all other subtasks, particularly on NIAH tasks (MK3: 3.6% vs 23.4%) and aggregation tasks (CWE: 26.1% vs 39.0%). The hardest NIAH subtask (MK3) shows severe degradation across all CPT conditions, dropping from 60.2% (Base) to single digits for junk-trained models.

3.4 Data Repetition Analysis

A critical confound in subset selection is data repetition: when the selected subset is smaller than the target token budget, samples must be repeated, potentially causing overfitting. Table 3 shows how our selection strategy evolved to address this issue.

The original top-10% hard cutoff (v1) selected only 6,905 unique samples, requiring 9.57× repetition to match the token budget. This caused severe ARC degradation (15.36%), likely due to overfitting on repeated samples. Switching to score-weighted sampling from all data (v2) increased diversity to 43,663 unique samples but diluted selection quality, improving ARC to 18.77% while RULER dropped to 54.53%. Our final pool-weighted approach (v3) balances quality and diversity by sampling from the top-50% pool, achieving 29,831 unique samples with 2.31× repetition.

As shown in Figure 2, ARC improves monotonically as repetition decreases (15.36% → 23.21%), confirming that data repetition was the primary cause of reasoning degradation. However, RULER does not improve correspondingly (60.72% → 54.66%), suggesting that orthogonality-based selection provides limited benefit for long-context preservation beyond what diversity alone provides.

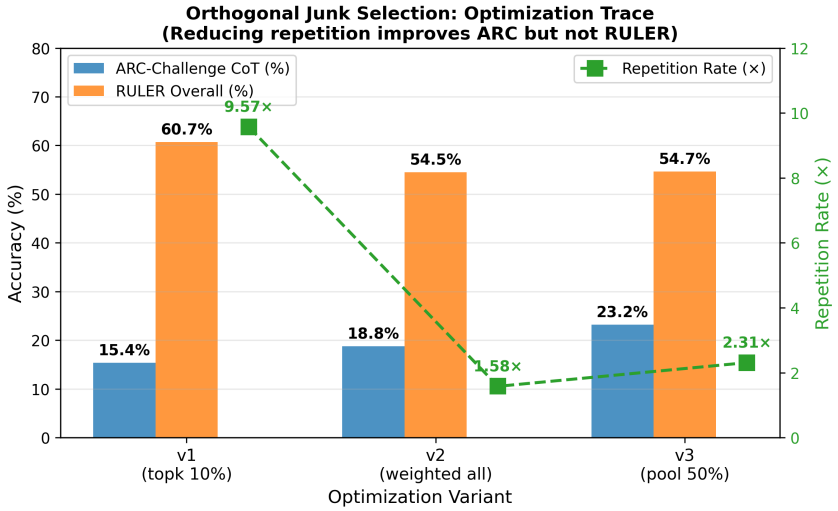

Figure 2: Optimization trace showing how reducing data repetition improves ARC reasoning accuracy (15.4% $\rightarrow$ 23.2%) while RULER long-context performance remains relatively flat ($\sim$54–61%). The dominant confound is data repetition rate, not orthogonality-based selection quality.

This reveals a fundamental limitation: gradient orthogonality computed at the LM head may not capture the full dynamics of capability preservation across all model layers.

4 Related Work

Data Selection for Language Models. Curating high-quality training data is critical for LLM performance. Early approaches used heuristic filters based on perplexity (Wenzek et al., 2019), deduplication (Lee et al., 2021), and domain classifiers. More recent work has developed principled selection methods: DSIR (Xie et al., 2023) uses importance resampling to match target distributions, while DataComp-LM (Li et al., 2024) provides benchmarks for comparing filtering strategies at scale. A comprehensive survey by Albalak et al. (2024) categorizes these approaches and identifies open challenges. Our work extends this line by using gradient geometry rather than surface-level quality signals for selection.

Gradient-Based Methods. Gradient information has been leveraged for both continual learning and data selection. Gradient Episodic Memory (Lopez-Paz & Ranzato, 2017) constrains updates to avoid increasing loss on previous tasks by projecting gradients onto feasible regions. PCGrad (Yu et al., 2020) addresses multi-task conflicts through gradient surgery, projecting conflicting gradients onto orthogonal directions. For data selection, GRAD-MATCH (Killamsetty et al., 2021) selects subsets whose gradients approximate full-data gradients, while LESS (Xia et al., 2024) uses influence functions to identify training examples most relevant to target tasks. Recent work by Zhang et al. (2026) applies gradient orthogonality specifically for domain adaptation, selecting data with gradients orthogonal to source domain to maximize transfer. Our approach adapts these principles to the continual pre-training setting, using orthogonality to capability-preserving anchors rather than task-specific objectives.

Continual Pre-Training. Catastrophic forgetting remains a central challenge when adapting pre-trained models to new data. Elastic Weight Consolidation (Kirkpatrick et al., 2016) addresses this by penalizing changes to parameters important for previous tasks, estimated via Fisher information. Recent surveys (Shi et al., 2024) comprehensively review continual learning methods for LLMs, including replay-based approaches and architectural modifications. Abbes et al. (2025) revisit replay and gradient alignment strategies specifically for continual pre-training, finding that simple replay often suffices. Layer-specific optimization methods like ELO (Yoo et al., 2026) selectively update layers to balance adaptation and retention. Most relevant to our work, Xing et al. (2025) introduce the "brain rot" phenomenon, demonstrating that exposure to low-quality social media content de-

grades LLM capabilities. Our work directly addresses this challenge by proposing gradient-based data selection as a potential mitigation strategy.

5 Conclusion

We investigated gradient-orthogonality data selection as a potential mitigation strategy for brain rot in continual pre-training. Our proposed Orthogonal Junk method provides modest improvement on long-context understanding (+2.89pp on RULER) but unexpectedly degrades reasoning ability (-5.12pp on ARC-Challenge). A simple perplexity baseline outperforms our method on RULER while showing similar reasoning degradation.

Our analysis reveals two key insights. First, brain-rot effects are task-dependent: long-context understanding is substantially more vulnerable than reasoning at the 1B scale. Second, data repetition is a critical confound in subset selection methods—reducing repetition from $9.57\times$ to $2.31\times$ improved ARC from 15.36% to 23.21%, while RULER remained flat. These findings suggest that gradient orthogonality computed at the LM head may not capture the full dynamics of capability preservation, and that simpler quality-based filtering may be more effective than gradient-based selection for mitigating brain rot.

References

Istabrak Abbes, Gopeshh Subbaraj, Matthew Riemer, Nizar Islah, Benjamin Therien, Tsuguchika Tabaru, Hiroaki Kingetsu, Sarath Chandar, and Irina Rish. Revisiting replay and gradient alignment for continual pre-training of large language models, 2025. URL https://arxiv.org/abs/2508.01908.

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and W. Wang. A survey on data selection for language models. *ArXiv*, abs/2402.16827, 2024.

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

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

Abhimanyu Dubey et al. The llama 3 herd of models. 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *ArXiv*, abs/2009.03300, 2020.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models? *ArXiv*, abs/2404.06654, 2024.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, A. De, and Rishabh K. Iyer. Grad-match: Gradient matching based data subset selection for efficient deep model training. pp. 5464–5474, 2021.

J. Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, J. Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114:3521 – 3526, 2016.

Katherine Lee, Daphne Ippolito, A. Nystrom, Chiyuan Zhang, D. Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. pp. 8424–8445, 2021.

Jeffrey Li, Alex Fang, G. Smyrnis, Maor Ivgi, Matt Jordan, S. Gadre, Hritik Bansal, E. Guha, Sedrick Scott Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean-Pierre Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, G. Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, K. Chandu, Thao Nguyen, Igor Vasiljevic, S. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke S. Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldani, Pang Wei Koh, J. Jitsev, Thomas Kollar, Alexandros G. Dimakis, Y. Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. *ArXiv*, abs/2406.11794, 2024.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. pp. 6467–6476, 2017.

Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58:1 – 42, 2024.

Guillaume Wenzek, M. Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco (Paco) Guzmán, Armand Joulin, and Edouard Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. pp. 4003–4012, 2019.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. Less: Selecting influential data for targeted instruction tuning. *ArXiv*, abs/2402.04333, 2024.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. Data selection for language models via importance resampling. *ArXiv*, abs/2302.03169, 2023.

Shuo Xing, Junyuan Hong, Yifan Wang, Runjin Chen, Zhenyu (Allen) Zhang, A. Grama, Zhengzhong Tu, and Zhangyang Wang. Llms can get "brain rot"! *ArXiv*, abs/2510.13928, 2025.

HanGyeol Yoo, ChangSu Choi, Minjun Kim, Seohyun Song, SeungWoo Song, Inho Won, Jongyoul Park, Cheoneum Park, and KyungTae Lim. Elo: Efficient layer-specific optimization for continual pretraining of multilingual llms, 2026. URL https://arxiv.org/abs/2601.03648.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, S. Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *ArXiv*, abs/2001.06782, 2020.

Xiyang Zhang, Yuanhe Tian, Hongzhi Wang, and Yan Song. Training data selection with gradient orthogonality for efficient domain adaptation. 2026.