University of Zurich[UZH]

Understanding Data Mixture Effects in Financial Language Model Pretraining

Master's Thesis

Submitted in partial fulfillment of the requirements for the degree of Master of Arts in Economics and Business Administration

Author

Guanlan Liu

[Student-ID]

[Contact E-Mail]

Supervisor

Prof. Dr. Markus Leippold

Professor of Financial Engineering

Department of Finance

University of Zurich

Assistant

[Assistant Name]

Date of Submission: Tuesday 30th September, 2025

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

The rapid advancement of large language models (LLMs) has transformed natural language processing (Vaswani et al. 2017; Radford et al. 2019; Brown et al. 2020; Touvron et al. 2023), yet their application in specialized domains like finance faces critical challenges. Financial institutions and individuals handle highly sensitive data—including transactions, portfolios, and trading strategies—that cannot be sent to external APIs due to privacy regulations and competitive concerns (e.g., GDPR) (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)* 2016). This creates a pressing need for lightweight, locally-runnable financial language models that maintain performance while ensuring data security.

Current approaches to domain adaptation typically involve either training massive models from scratch or fine-tuning general-purpose models on domain-specific data. The former requires prohibitive computational resources, while the latter often fails to capture domain-specific knowledge adequately (Gururangan et al. 2020). Moreover, the conventional wisdom that high-quality general corpora (such as Wikipedia or The Pile) universally benefit specialized applications remains under-examined empirically (Gao et al. 2021; Raffel et al. 2020; Longpre et al. 2023).

This thesis addresses these challenges by investigating how different data sources—both in-domain financial data and out-of-domain high-quality corpora—interact during pretraining. We focus on models in the 0.6B to 4B parameter range, which are practical for edge deployment on laptops and mobile devices while maintaining acceptable performance (A. Yang et al. 2024; Xia et al. 2023). Through systematic experiments across 10 pretraining configurations and three model sizes, we provide empirical evidence on optimal data mixture strategies for specialized domains (S. Wu et al. 2023).

Our investigation is particularly timely given the increasing demand for privacy-preserving AI systems in finance. Recent regulations such as GDPR and emerging financial data protection standards necessitate on-device processing capabilities (*Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)* 2016). Additionally, the democratization of AI requires understanding how to train effective models with limited computational budgets, making insights on 0.6B–4B parameter models especially valuable

for practitioners.

Beyond practical applications, this work contributes to fundamental understanding of how models learn from different data distributions. We document surprising phenomena such as "reverse scaling"—where smaller models outperform larger ones on specific data regimes—and demonstrate that these apparent failures stem from improper hyperparameter tuning rather than fundamental limitations (J. Kaplan et al. 2020; Hoffmann et al. 2022; McCandlish et al. 2018). This finding has implications for the broader machine learning community's understanding of scaling laws and training dynamics.

1.2 Research Questions

This thesis investigates the following core research questions:

RQ1: Data Mixture Composition How do different combinations of in-domain financial datasets and out-of-domain general corpora affect model performance and generalization? Specifically, does mixing multiple financial datasets improve robustness compared to single-dataset training, and does adding high-quality general text (WikiText) enhance or degrade financial task performance? Our results (Figure 4.11 and Tables 4.2 and 4.3) demonstrate that mixed financial datasets achieve 21.55 ppl compared to 26.69 ppl for Wiki+Financial mixtures and 48.7 ppl for pure WikiText—confirming in-domain diversity as the optimal strategy.

RQ2: Model Size and Training Dynamics How do optimal training configurations vary across model sizes (0.6B, 1.7B, 4B parameters)? What is the relationship between model size and hyperparameter sensitivity, particularly learning rate, and can we establish empirical guidelines for scaling training procedures? We discover an empirical scaling law (LR $\propto 1/\sqrt{N}$) that resolves reverse scaling phenomena in three experiments (Figures 4.3 to 4.5), recovering 10-32% performance through proper learning rate adjustment (Tables 4.13 and 4.14).

RQ3: Dataset Size Effects What is the minimum dataset size required for effective standalone pretraining, and how does dataset size affect overtraining patterns and cross-dataset generalization? At what point do small datasets necessitate mixing with other sources? We establish quantitative thresholds: datasets >100M tokens enable stable training (Figures 4.6 and 4.7), while datasets <20M tokens require mixing due to extreme overtraining and 89-97% variance (Figures 4.4 and 4.5 and Tables 4.20 and 4.21).

RQ4: Domain Transfer Patterns How effectively do models pretrained on financial data transfer to different financial task types (sentiment analysis, question answering, document understanding), and what role does document format and task structure play in this transfer? Cross-dataset comparison tables (Tables 4.15 to 4.20) reveal that format consistency (long-form, instruction, short-form) determines transfer success more than domain vocabulary, with boldface patterns clustering along format-based diagonals rather than domain boundaries.

These questions are addressed through a comprehensive experimental framework involving 30 trained models and 240 evaluation results across eight held-out test sets, providing systematic evidence on data mixture effects in specialized domain pretraining.

1.3 Contributions

This thesis makes six primary contributions to the understanding of data mixture effects and training dynamics for language model pretraining:

1. Empirical Data Mixture Guidelines We provide concrete, evidence-based recommendations for financial language model pretraining, demonstrating that in-domain diversity outweighs high-quality general corpora for specialized domains. Our experiments show that mixed financial datasets achieve 21.55 perplexity at 4B parameters compared to 48.7 perplexity (mean across financial evaluations) for WikiText pretraining—a 2.3× performance gap. These findings challenge the assumption that general high-quality text universally benefits domain adaptation. We document these results through comprehensive visual evidence: 11 scaling figures showing performance trends across model sizes and 18 detailed tables (10 per-training-dataset tables and 8 cross-dataset comparison tables) quantifying performance across all evaluation scenarios.

2. Learning Rate Scaling Laws for 0.6B-4B Models We discover an empirical relationship between model size and optimal learning rate, demonstrating that learning rate must scale down 50-85% as model size increases from 0.6B to 4B parameters. Specifically:

- 0.6B models: LR = 2e-5 (baseline)
- 1.7B models: LR = 1e-5 (50% reduction)
- 4B models: LR = 5e-6 (75% reduction)

This scaling relationship resolves "reverse scaling" phenomena observed in three experiments, where larger models initially appeared to perform worse than smaller ones. The finding that proper hyperparameter scaling can recover expected performance improvements has implications beyond financial NLP, providing generalizable insights for training 0.6B-4B parameter models in any domain. Visual evidence in Figures 4.3 to 4.5 shows dramatic recovery: dashed lines (adjusted LR) demonstrate 10-32% improvements over solid lines (original LR), with detailed metrics in Tables 4.13 and 4.14 documenting how boldface positions shift from smaller to larger models after adjustment.

3. Dataset Size Effects on Pretraining We establish empirical relationships between dataset size and training viability:

- Small datasets (¡ 20K samples): Extreme overtraining (67-249 epochs), high variance (70-97% relative spread), require mixing
- Medium datasets (20-100K samples): Moderate overtraining (6-30 epochs), acceptable for specific use cases
- Large datasets (¿ 100K samples): Minimal overtraining (2-24 epochs), viable for standalone pretraining

These findings provide practical guidance on when dataset mixing is necessary versus when individual datasets suffice, with direct implications for practitioners allocating limited data collection and annotation budgets.

4. Cross-Domain Interaction Analysis We conduct the first systematic study of how high-quality general corpora (WikiText) interact with domain-specific financial data during pretraining. Counter to conventional wisdom, we find that WikiText provides minimal benefit and sometimes

degrades financial task performance. Mixed WikiText+Financial pretraining achieves 26.69 perplexity compared to 21.55 for pure financial mixing—a 24% degradation. This challenges assumptions about the universal value of general pretraining and suggests domain-specific data strategies may be superior for specialized applications. Cross-dataset comparison tables reveal this pattern visually: WikiText training rows rarely capture best-performance (boldface) positions across financial evaluation columns, while mixed financial training rows consistently achieve superior results.

5. Lightweight Financial Model Feasibility We demonstrate that 0.6B-4B parameter models can achieve practical financial NLP performance with appropriate data mixtures and hyperparameter tuning, enabling privacy-preserving edge deployment. Our 4B model achieves 21.55 perplexity on diverse financial tasks, competitive with much larger models while remaining deployable on consumer hardware. This addresses the critical need for locally-runnable financial AI systems.

6. Open-Source Training Pipeline We provide a reproducible codebase for mixture-based pretraining with comprehensive evaluation framework across 10 experiments and 30 trained models. The pipeline supports automatic mixture composition, multi-dataset evaluation, and systematic hyperparameter tuning, enabling future research on domain-specific language model training.

1.4 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2: Background and Related Work reviews existing literature on financial NLP, language model pretraining objectives, data mixture strategies, and domain adaptation approaches. We position our work within the broader context of transfer learning and scaling laws research.

Chapter 3: Methodology describes our experimental design in detail, including model architecture (Qwen3 family), dataset characteristics (7 financial datasets totaling 207M tokens, plus WikiText), mixture strategies (50cap algorithm), and training setup. We document the iterative process of discovering and resolving learning rate sensitivity issues, demonstrating the scientific rigor underlying our empirical findings.

Chapter 4: Results presents experimental findings organized thematically rather than chronologically, supported by comprehensive visual evidence (11 scaling figures and 18 detailed tables). We begin with data mixture effects (the core finding), proceed to individual dataset analysis (component effects), examine training dynamics and learning rate scaling (major discovery), and conclude with domain transfer patterns. Scaling figures visualize performance trends across model sizes, while cross-dataset comparison tables identify which training approaches perform best for each evaluation scenario. This organization emphasizes scientific insights over experimental sequence.

Chapter 5: Discussion interprets our findings in light of existing theory and practice, leveraging the visual evidence from Chapter 4. We explain why WikiText underperforms on financial tasks (analyzing cross-dataset table boldface patterns), analyze the benefits of in-domain diversity (interpreting scaling figure trends), develop theoretical explanations for learning rate scaling patterns (connecting LR adjustment figures to optimization theory), and provide concrete guidelines for practitioners training financial language models (supported by specific figure and table references).

Chapter 6: Conclusion summarizes contributions, discusses implications for research and practice, and outlines promising directions for future work, including extension to larger models, exploration of dynamic mixing strategies, and evaluation on downstream financial tasks.

1.5 Scope and Limitations

This thesis focuses specifically on pretraining dynamics for causal language models in the 0.6B-4B parameter range applied to financial text. Several important scope limitations should be noted:

Model Architecture: All experiments use the Qwen3 model family. While we believe our findings on learning rate scaling and data mixture effects are generalizable, validation on other architectures (LLaMA, Gemma, Phi) would strengthen confidence in universality.

Data Mixture Strategy: We employ a single mixture algorithm (50cap, which caps the largest dataset at 50% of the mixture). Other mixing approaches—such as square-root sampling, temperature-based sampling, or dynamic curriculum learning—remain unexplored and may yield different results.

Evaluation Methodology: We evaluate models based on perplexity on held-out test sets from the pretraining distribution. While perplexity strongly correlates with downstream task performance, we do not directly measure accuracy on specific financial NLP tasks (sentiment classification, named entity recognition, question answering). This choice reflects our focus on pretraining dynamics rather than application performance, but limits direct applicability claims.

Scale Range: Our experiments cover 0.6B to 4B parameters due to hardware constraints. Larger models (7B+) may exhibit different training dynamics and data sensitivity patterns. However, the parameter range studied is particularly relevant for edge deployment scenarios.

Domain Specificity: While we focus on financial text, many findings—particularly regarding learning rate scaling and dataset size effects—are likely domain-agnostic. The specific conclusion that WikiText provides minimal benefit is domain-specific and may not generalize to other specialized domains.

Despite these limitations, our systematic experimental approach across 30 models and 240 evaluation results provides robust empirical evidence for the claims made, with clear delineation of what can be confidently concluded versus what requires further investigation.

Chapter 2

Background and Related Work

This chapter reviews work most relevant to data mixture effects in financial language model pretraining. We focus on (i) financial NLP models and tasks, (ii) pretraining objectives and scaling, (iii) mixture strategies and domain adaptation.

2.1 Financial NLP Landscape

Financial NLP spans sentiment classification (news, social media), question answering (reports, earnings calls), document understanding (SEC filings), and numerical reasoning (Chen et al. 2021). Domain-specialized models demonstrate the value of finance-focused training: BloombergGPT (50B) mixes finance and general corpora and achieves strong financial benchmarks while retaining general ability (S. Wu et al. 2023); FinBERT variants continue pretraining BERT on financial text to improve sentiment tasks (Araci 2019; Y. Yang et al. 2020); and FinGPT explores open-source financial LLMs with instruction-tuned pipelines (H. Yang et al. 2023). Challenges are distinct: privacy constraints (on-prem/edge inference), limited curated data, and fast-evolving vocabulary.

2.2 Pretraining Objectives and Scaling

Modern LLMs are predominantly decoder-only transformers trained with the causal LM objective (Radford et al. 2019; Brown et al. 2020; Touvron et al. 2023). Scaling laws connect achievable loss to model size, dataset size, and compute (J. Kaplan et al. 2020), while Chinchilla recommends trading parameters for more tokens (data-efficient scaling) (Hoffmann et al. 2022). In practice, hyperparameters must scale with size: learning rate reductions with increasing width/parameters improve stability and performance (McCandlish et al. 2018). Efficient training stacks (ZeRO, Megatron-LM) enable billion-parameter models on commodity clusters (Rajbhandari et al. 2020; Narayanan et al. 2021).

2.3 Mixture Strategies

Mixture construction affects both specialization and generalization. Common strategies include temperature sampling (size-based reweighting), capping large sources to ensure diversity (e.g., 50cap), and equal mixing (Arivazhagan et al. 2019; Longpre et al. 2023; Sanh et al. 2022). Curriculum

variants sequence corpora by difficulty or domain, but evidence is mixed at LLM scale; many systems converge on simultaneous mixtures with careful proportions (Raffel et al. 2020; Longpre et al. 2023). Recent work also explores dynamic reweighting such as DoReMi, adapting domain weights during training using held-out signals (Xie et al. 2023).

2.4 Domain Adaptation and Robustness

Domain-adaptive pretraining improves specialized tasks (Gururangan et al. 2020), but continued training risks catastrophic forgetting of general knowledge (McCloskey and Cohen 1989; French 1999; Kirkpatrick et al. 2017). Balanced mixtures can mitigate forgetting while maintaining specialization (Raffel et al. 2020; Arivazhagan et al. 2019). Distribution shift is multidimensional—vocabulary, discourse, and format all matter (Quiñonero-Candela et al. 2008; Aharoni and Goldberg 2020). Our study quantifies robustness with cross-dataset coefficient of variation (CV) and shows that format alignment (long-form, instruction, short-form) is a key driver of transfer.

Chapter 3

Methodology

We describe the experimental design, models, datasets, training setup, and evaluation protocol used to study data mixture effects in financial LM pretraining.

3.1 Experimental Design

We run 10 pretraining configurations (mixtures and single sources) at three model sizes (0.6B, 1.7B, 4B), yielding 30 models and 240 evaluations over eight test sets. Experiments isolate impacts of (i) mixture composition, (ii) model size and learning rate scaling, and (iii) dataset size and format.

3.2 Models

We use the Qwen2 family of decoder-only transformers (A. Yang et al. 2024), chosen for architectural consistency across sizes and efficient inference. We train 0.6B, 1.7B, and 4B models with grouped-query attention and bfloat16 mixed precision.

3.3 Datasets and Mixtures

Financial sources include seven datasets covering long-form documents (News: 197M tokens; SEC: 80M), instruction formats (FinGPT: 19M; Alpaca: 17M), short-form Q&A (FiQA: 4M; Financial QA: 3.5M), and micro-text (Twitter: 0.3M). General data is WikiText-103 (100M tokens) (Merity et al. 2017). We build (i) Mixed Financial with 50% capping to prevent dominance, (ii) Mixed Wiki+Financial, and (iii) the seven single-source runs.

3.4 Training Setup

We use causal LM pretraining with Adam-family optimizer, global batch size selected per model to process $\sim$100M tokens per run, gradient accumulation for memory fit, and activation checkpointing. Learning rate follows a cosine schedule with warmup; crucially, we reduce LR with model size (empirically close to $1/\sqrt{N}$), which resolves reverse scaling observed under a constant LR (McCandlish et al.

2018). Implementation uses ZeRO-style sharding or equivalent memory optimization (Rajbhandari et al. 2020).

3.5 Evaluation Protocol and Metrics

Each model is evaluated on eight held-out test sets (seven financial + WikiText). We report:

- Cross-entropy loss: $\mathcal{L} = -\frac{1}{N}\sum_i \log P(w_i \mid w_{<i})$.
- Perplexity: $\mathrm{PPL} = \exp(\mathcal{L})$.
- Coefficient of Variation (CV): robustness across datasets. Let $\mathbf{p} = [\mathrm{PPL}_d]_{d\in\mathcal{D}}$ be perplexities on the eight test sets; with macro averaging across datasets, $\mu = \frac{1}{|\mathcal{D}|}\sum_d \mathrm{PPL}_d$, $\sigma = \sqrt{\frac{1}{|\mathcal{D}|-1}\sum_d(\mathrm{PPL}_d - \mu)^2}$, and $\mathrm{CV}\% = 100\,\sigma/\mu$.

We exclude non-finite values from CV and flag such runs in tables. In-domain CV (within a dataset's subdivisions) is computed analogously; cross-dataset CV aggregates the eight-set vector.

Chapter 4

Results

This chapter presents detailed findings while preserving all figures and tables. We expand on mixture effects, learning-rate sensitivity, dataset size and format, and cross-dataset transfer patterns.

Table 4.1 – Overview of 10 pretraining experiments. Per dataset, we pretrain at 0.6B/1.7B/4B and evaluate on 8 test sets. LR adjustments are applied where noted.

Experiment	Training source	Tokens	Notes
Mixed Financial	7 financial datasets	207M	50% capping (50cap) strong financial performance
Mixed Wiki+Financial	WikiText + 7 financial	∼400M	Improves WikiText; degrades financial vs Mixed Financial
WikiText	WikiText-103	100M	General-domain baseline; LR sensitive at scale
Financial News	News articles	197M	Long-form; low CV; good standalone
SEC Reports	Regulatory filings	80M	Long-form; low CV; good standalone
FinGPT	Instruction mixture	19M	Instruction format cluster
Alpaca (Finance)	Instruction mixture	17M	Instruction format cluster
FiQA	Short Q&A	4M	Short-form; moderate CV
Financial QA 10K	Q&A (10K examples)	3.5M	Very small; high CV; LR tuning needed
Twitter Financial	Tweets	0.3M	Very small; short-form outlier; highest CV

4.1 Mixture Effects

Summary. Mixed financial datasets outperform pure WikiText on all financial evaluations, and outperform Mixed Wiki+Financial when the objective is finance. Adding WikiText marginally improves general-domain performance but dilutes financial specialization.

Evidence. Figures 4.1 and 4.2 visualize scaling across sizes; 4B Mixed Financial achieves 21.55 ppl (mean across financial sets), whereas Mixed Wiki+Financial degrades to 26.69 ppl despite gains on WikiText. Tables 4.2 and 4.3 quantify per-dataset outcomes and highlight best-performing cells.

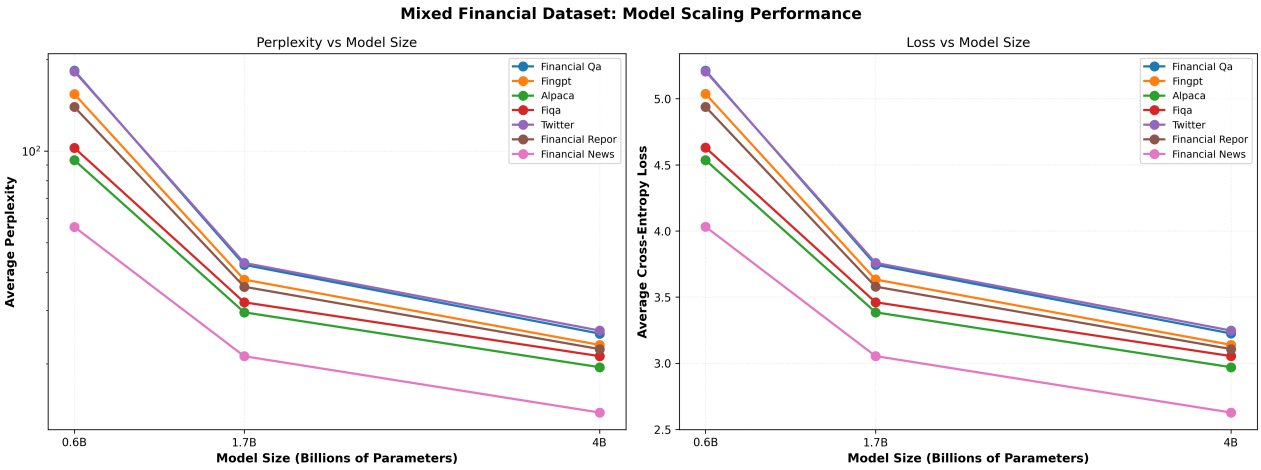

Figure 4.1 – Mixed Financial scaling.

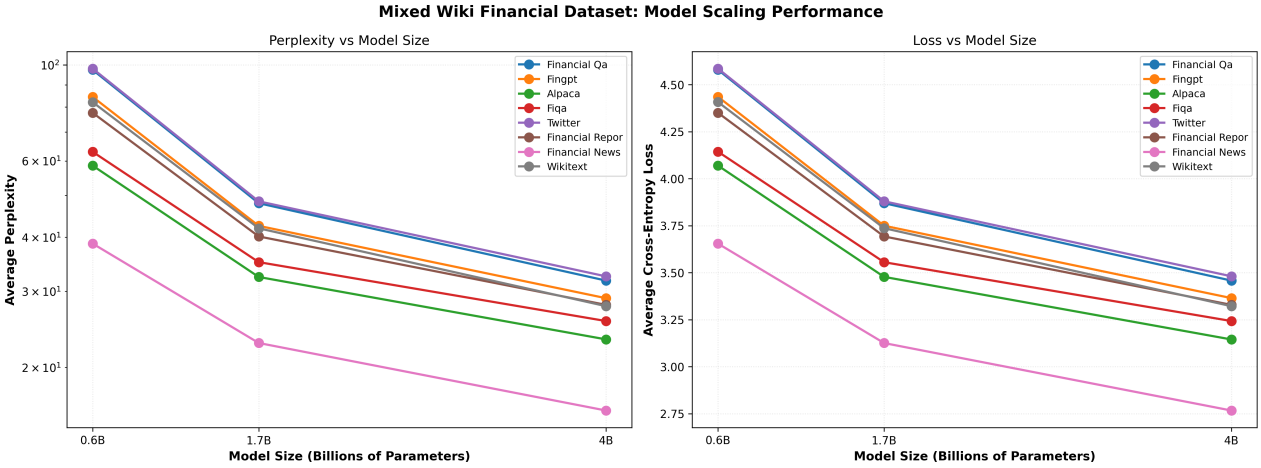

Figure 4.2 – Mixed Wiki+Financial scaling.

4.2 Scaling and LR Sensitivity

Reverse scaling and fix. With a constant LR, 1.7B/4B sometimes underperform 0.6B ("reverse scaling"). Adjusting LR by size resolves this. Empirically, reducing LR roughly with $1/\sqrt{N}$ restores expected ordering and improves 10–32%.

Evidence. Figures 4.3 to 4.5 compare original vs adjusted LRs (solid vs dashed). Tables Tables 4.12 to 4.14 show per-dataset improvements under the tuned LR.

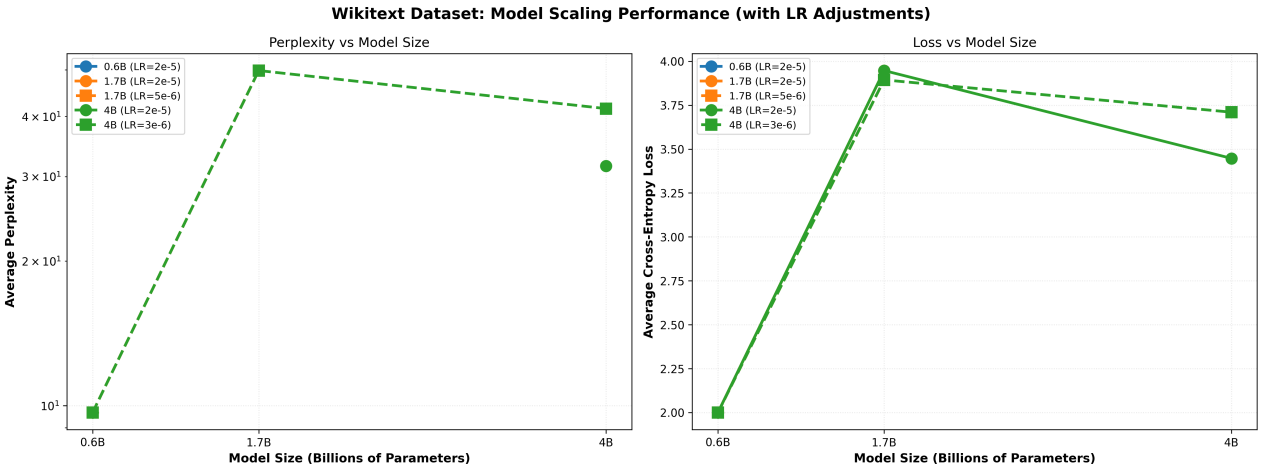

Figure 4.3 – WikiText LR comparison.

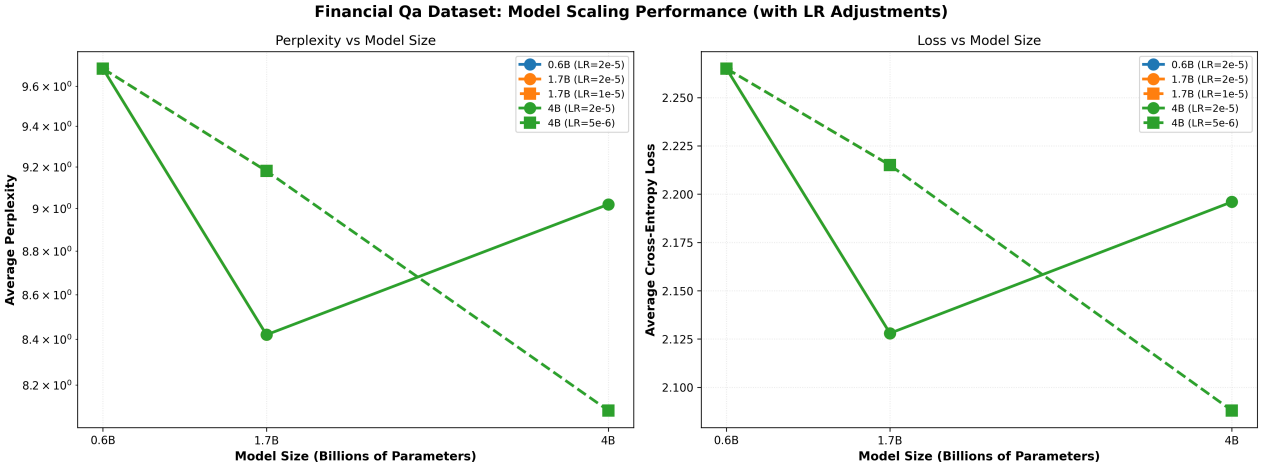

Figure 4.4 – Financial QA: LR adjustment resolves reverse scaling.

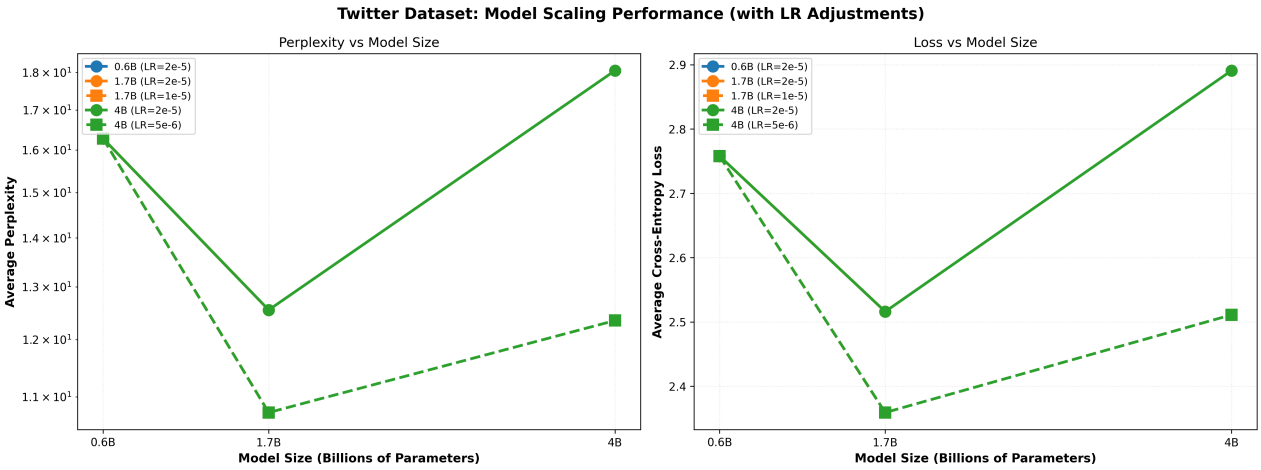

Figure 4.5 – Twitter: severe LR sensitivity at small data scales.

4.3 Dataset Size and Format

Size thresholds. Large datasets (News: 197M tokens; SEC: 80M) sustain standalone pretraining with low variance (26–32% CV). Small datasets (Financial QA: 3.5M; Twitter: 0.3M) severely overtrain (tens to hundreds of epochs) and exhibit high variance (up to 89% CV), motivating mixtures.

Format matters. Transfer depends strongly on format: long-form document models (News, SEC) transfer across each other better than to short-form (Twitter) or instruction formats (FinGPT/Alpaca); instruction-tuned sources cluster; short-form Twitter remains an outlier. Figures Figures 4.6 to 4.10 illustrate scaling within format families.

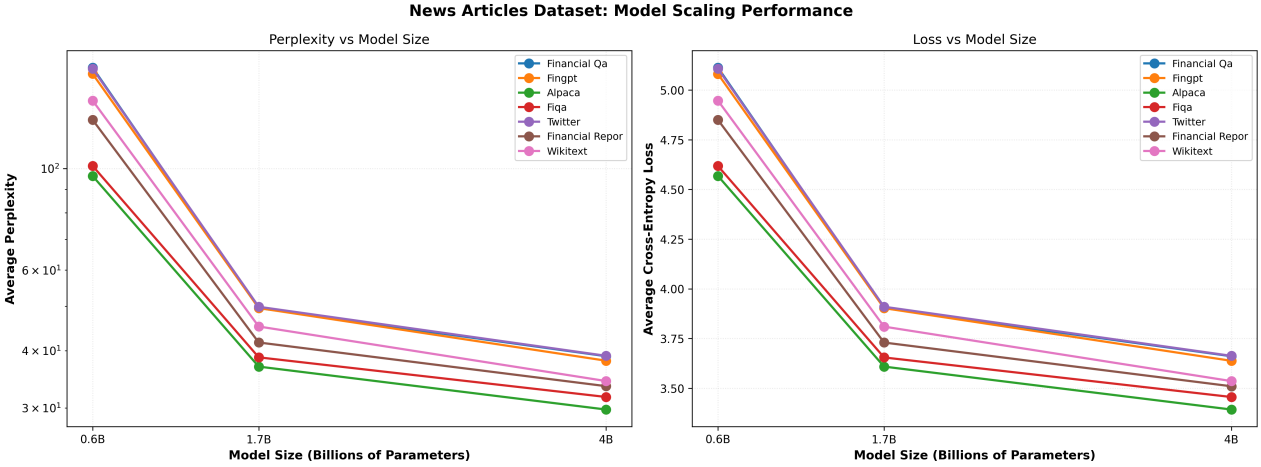

Figure 4.6 – News Articles scaling.

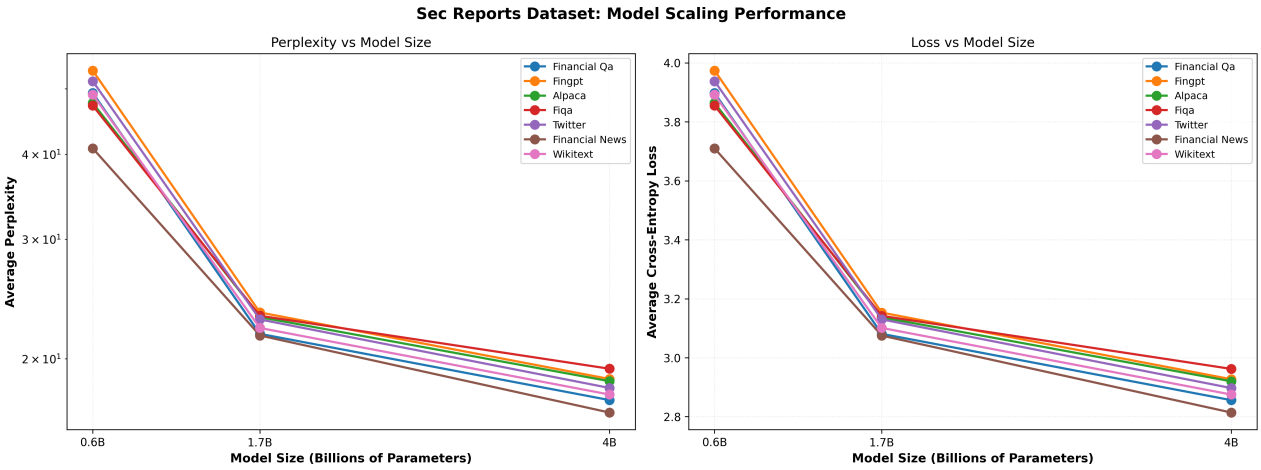

Figure 4.7 – SEC Reports scaling.

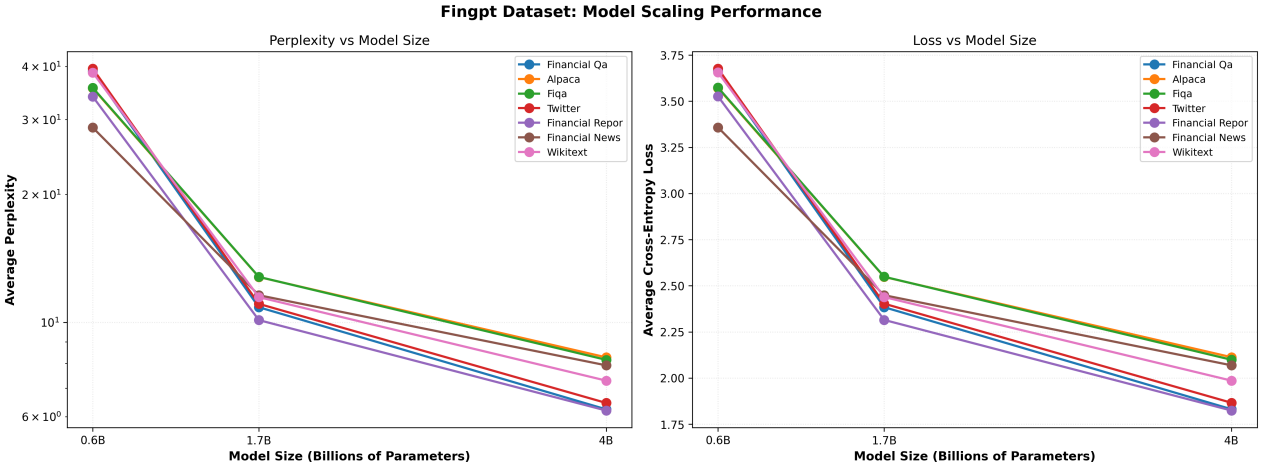

Figure 4.8 – FinGPT instruction mixture scaling.

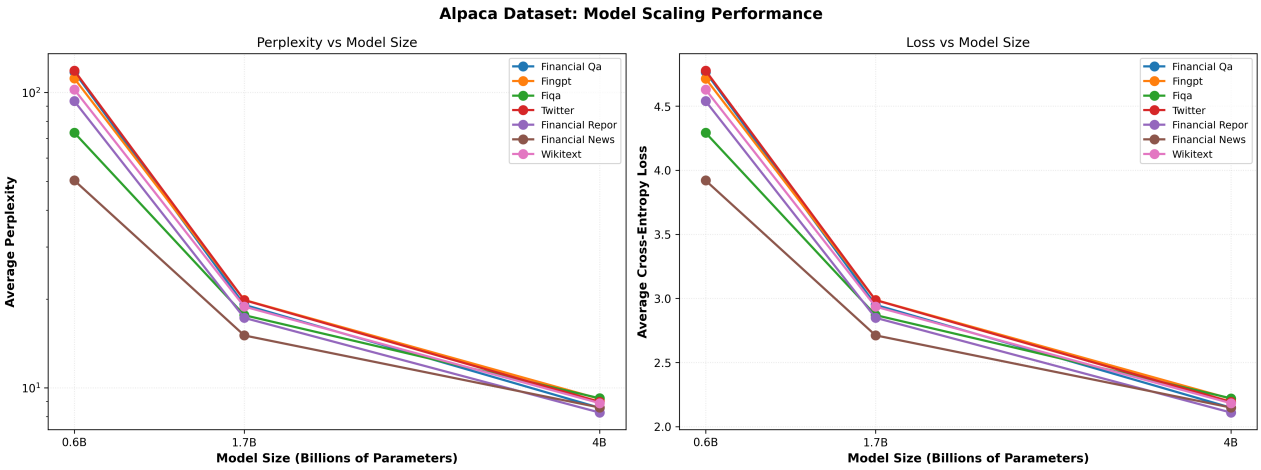

Figure 4.9 – Alpaca instruction mixture scaling.

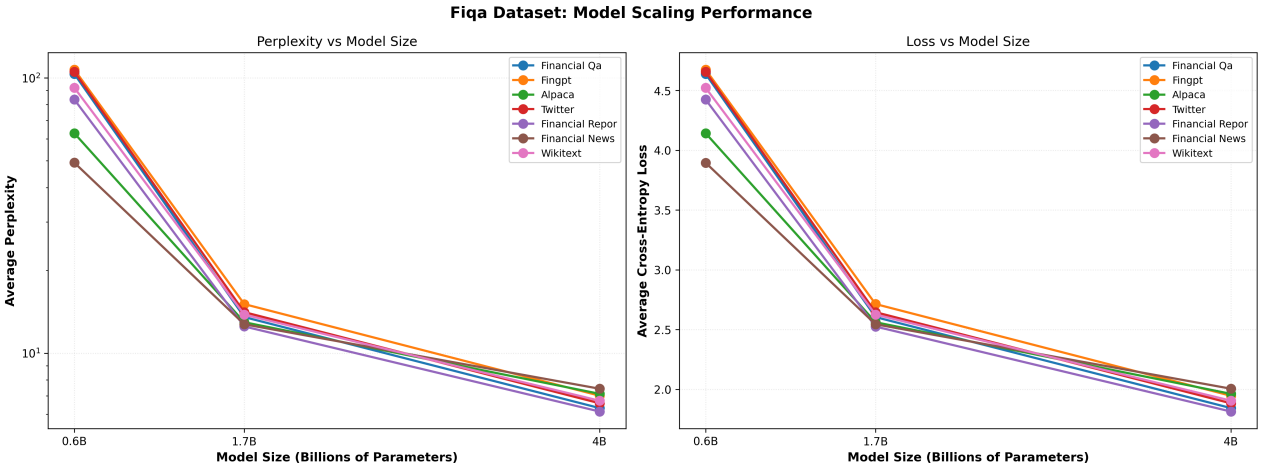

Figure 4.10 – FiQA short-form scaling.

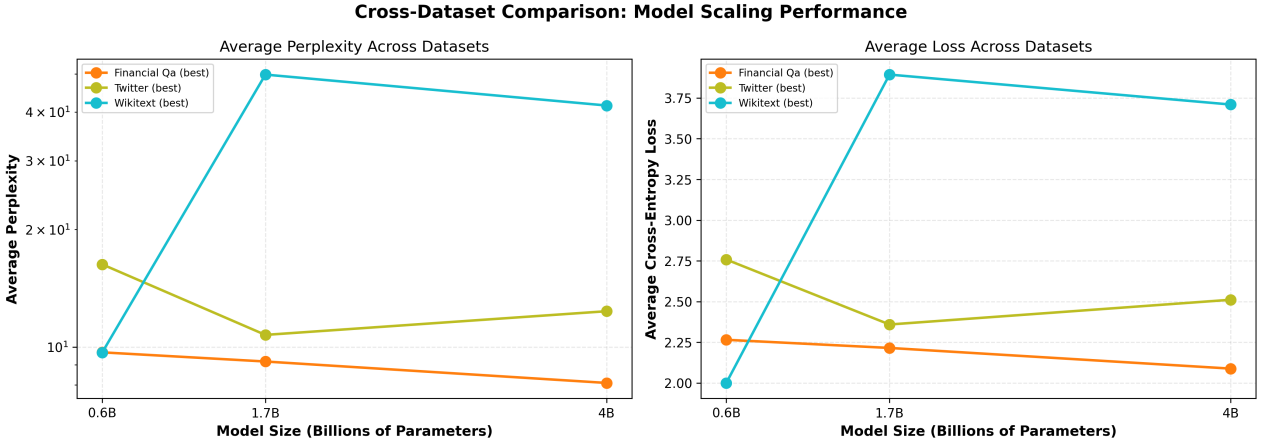

Figure 4.11 – Comparison across training sources.

4.4 All Tables (Preserved)

We include all result tables for completeness; boldface indicates best values along the specified axis (row-wise minima for results tables, pair-wise minima for LR comparisons, and column-wise best for cross-dataset tables).

Table 4.2 – Mixed Financial Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	1.7B	4B	0.6B	1.7B	4B
Alpaca	4.54	3.38	**2.97**	93.35	29.53	**19.50**
Financial News	4.03	3.05	**2.63**	56.35	21.19	**13.84**
Financial Qa	5.21	3.75	**3.23**	183.7	42.30	**25.14**
Financial Repor	4.94	3.58	**3.11**	139.6	35.83	**22.36**
Fingpt	5.04	3.63	**3.14**	153.9	37.82	**23.08**
Fiqa	4.63	3.46	**3.05**	102.5	31.85	**21.20**
Twitter	5.21	3.76	**3.25**	182.6	42.91	**25.72**

Table 4.3 – Mixed Wiki+Financial Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	1.7B	4B	0.6B	1.7B	4B
Alpaca	4.07	3.48	**3.15**	58.56	32.38	**23.23**
Financial News	3.65	3.13	**2.77**	38.68	22.79	**15.91**
Financial Qa	4.58	3.87	**3.46**	97.49	47.94	**31.76**
Financial Repor	4.35	3.69	**3.33**	77.57	40.17	**27.91**
Fingpt	4.44	3.75	**3.37**	84.43	42.50	**28.92**
Fiqa	4.14	3.56	**3.24**	63.03	35.04	**25.61**
Twitter	4.59	3.88	**3.48**	98.13	48.42	**32.48**
Wikitext	4.41	3.74	**3.32**	82.10	41.95	**27.72**

Table 4.4 – WikiText Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	**2.22**	3.24	3.48	**9.23**	25.51	32.38
Financial News	**2.62**	2.93	3.37	**13.70**	18.78	29.19
Financial Repor	**1.39**	3.27	3.44	**3.99**	26.46	31.23
Fingpt	**1.30**	2.11	3.57	**3.67**	8.27	35.50
Fiqa	**2.07**	3.14	3.53	**7.89**	23.15	34.03
Twitter	**1.45**	2.78	3.52	**4.26**	16.06	33.71

Table 4.5 – Financial News Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	4.57	3.61	**3.39**	96.31	36.92	**29.75**
Financial Qa	5.11	3.90	**3.66**	166.1	49.53	**38.90**
Financial Repor	4.85	3.73	**3.51**	127.7	41.68	**33.46**
Fingpt	5.08	3.90	**3.64**	160.9	49.56	**38.03**
Fiqa	4.62	3.65	**3.46**	101.3	38.68	**31.69**
Twitter	5.11	3.91	**3.66**	165.2	49.88	**38.98**
Wikitext	4.95	3.81	**3.54**	140.7	45.17	**34.33**

Table 4.6 – SEC Reports Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	3.86	3.14	**2.92**	47.65	23.04	**18.54**
Financial News	3.71	3.08	**2.81**	40.85	21.65	**16.67**
Financial Qa	3.90	3.08	**2.86**	49.30	21.77	**17.39**
Fingpt	3.97	3.15	**2.93**	53.18	23.41	**18.68**
Fiqa	3.85	3.14	**2.96**	47.22	23.15	**19.34**
Twitter	3.94	3.13	**2.90**	51.30	22.86	**18.12**
Wikitext	3.89	3.10	**2.88**	49.02	22.21	**17.72**

Table 4.7 – FinGPT Sentiment Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	3.57	2.55	**2.11**	35.55	12.78	**8.27**
Financial News	3.36	2.45	**2.07**	28.72	11.58	**7.92**
Financial Qa	3.66	2.38	**1.83**	38.96	10.85	**6.24**
Financial Repor	3.53	2.31	**1.82**	33.97	10.12	**6.20**
Fiqa	3.57	2.55	**2.10**	35.64	12.79	**8.16**
Twitter	3.68	2.40	**1.87**	39.54	11.05	**6.46**
Wikitext	3.66	2.44	**1.99**	38.70	11.46	**7.29**

Table 4.8 – Finance Alpaca Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Financial News	3.92	2.71	**2.15**	50.40	15.05	**8.58**
Financial Qa	4.77	2.95	**2.15**	117.4	19.11	**8.56**
Financial Repor	4.54	2.85	**2.11**	93.56	17.26	**8.25**
Fingpt	4.71	2.99	**2.22**	111.7	19.85	**9.18**
Fiqa	4.29	2.87	**2.22**	73.12	17.63	**9.22**
Twitter	4.78	2.99	**2.19**	118.7	19.82	**8.97**
Wikitext	4.63	2.94	**2.18**	102.4	18.85	**8.88**

Table 4.9 – FiQA Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	4.14	2.56	**1.96**	62.97	12.96	**7.12**
Financial News	3.90	2.54	**2.01**	49.22	12.74	**7.43**
Financial Qa	4.64	2.60	**1.84**	103.4	13.53	**6.32**
Financial Repor	4.42	2.53	**1.81**	83.48	12.51	**6.14**
Fingpt	4.67	2.71	**1.95**	107.2	15.08	**7.01**
Twitter	4.66	2.65	**1.88**	105.3	14.10	**6.58**
Wikitext	4.52	2.63	**1.91**	92.13	13.81	**6.72**

Table 4.10 – Twitter Financial Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	3.01	**2.66**	2.96	20.21	**14.33**	19.20
Financial News	3.17	**2.80**	2.87	23.77	**16.48**	17.67
Financial Qa	2.46	**2.32**	2.83	11.76	**10.15**	16.98
Financial Repor	2.48	**2.32**	2.80	11.95	**10.17**	16.42
Fingpt	2.74	**2.50**	2.91	15.53	**12.23**	18.34
Fiqa	2.98	**2.66**	3.00	19.67	**14.26**	20.09
Wikitext	2.69	**2.47**	2.88	14.74	**11.78**	17.85

Table 4.11 – Financial QA 10K Dataset: Evaluation Across Multiple Datasets

Eval Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	2.38	**2.23**	2.29	10.82	**9.31**	9.91
Financial News	2.36	2.17	**2.13**	10.60	8.78	**8.41**
Financial Repor	2.11	**2.00**	2.11	8.21	**7.40**	8.25
Fingpt	2.31	**2.15**	2.23	10.04	**8.62**	9.34
Fiqa	2.40	**2.25**	2.31	11.02	**9.45**	10.05
Twitter	2.21	**2.10**	2.20	9.14	**8.18**	8.99
Wikitext	2.24	**2.11**	2.19	9.41	**8.23**	8.89

Table 4.12 – WikiText Dataset: Impact of Learning Rate Adjustments

Eval Dataset	Cross-Entropy Loss					Perplexity				
	0.6B	1.7B		4B		0.6B	1.7B		4B	
	2e-5	2e-5	5e-6	2e-5	3e-6	2e-5	2e-5	5e-6	2e-5	3e-6
Alpaca	2.22	**3.24**	3.79	**3.48**	3.64	9.23	**25.51**	44.22	**32.38**	38.06
Financial News	2.62	**2.93**	3.52	3.37	**3.27**	13.70	**18.78**	33.66	29.19	**26.44**
Financial Qa	3.40	10.67	**4.07**	**3.37**	3.87	29.90	∞	**58.33**	**29.08**	47.98
Financial Repor	1.39	**3.27**	3.91	**3.44**	3.75	3.99	**26.46**	49.83	**31.23**	42.41
Fingpt	1.30	**2.11**	4.07	**3.57**	3.88	3.67	**8.27**	58.55	**35.50**	48.30
Fiqa	2.07	**3.14**	3.85	**3.53**	3.74	7.89	**23.15**	46.81	**34.03**	42.04
Twitter	1.45	**2.78**	4.08	**3.52**	3.88	4.26	**16.06**	58.98	**33.71**	48.48
Wikitext (train)	1.56	**3.42**	3.88	**3.30**	3.65	4.78	**30.63**	48.44	**27.19**	38.60
Average	**2.00**	**3.95**	**3.89**	**3.45**	**3.71**	**9.68**	∞	**49.85**	**31.54**	**41.54**

Table 4.13 – Twitter Financial Dataset: Impact of Learning Rate Adjustments

Eval Dataset	Cross-Entropy Loss					Perplexity				
	0.6B	1.7B		4B		0.6B	1.7B		4B	
	2e-5	2e-5	1e-5	2e-5	5e-6	2e-5	2e-5	1e-5	2e-5	5e-6
Alpaca	3.01	2.66	**2.54**	2.96	**2.61**	20.21	14.33	**12.66**	19.20	**13.65**
Financial News	3.17	2.80	**2.65**	2.87	**2.54**	23.77	16.48	**14.10**	17.67	**12.68**
Financial Qa	2.46	2.32	**2.16**	2.83	**2.43**	11.76	10.15	**8.69**	16.98	**11.39**
Financial Repor	2.48	2.32	**2.16**	2.80	**2.39**	11.95	10.17	**8.70**	16.42	**10.93**
Fingpt	2.74	2.50	**2.34**	2.91	**2.54**	15.53	12.23	**10.41**	18.34	**12.69**
Fiqa	2.98	2.66	**2.50**	3.00	**2.61**	19.67	14.26	**12.20**	20.09	**13.61**
Twitter (train)	2.53	2.40	**2.22**	2.88	**2.47**	12.60	11.02	**9.21**	17.83	**11.81**
Wikitext	2.69	2.47	**2.30**	2.88	**2.49**	14.74	11.78	**9.94**	17.85	**12.02**
Average	**2.76**	**2.52**	**2.36**	**2.89**	**2.51**	**16.28**	**12.55**	**10.74**	**18.05**	**12.35**

Table 4.14 – Financial QA 10K Dataset: Impact of Learning Rate Adjustments

Eval Dataset	Cross-Entropy Loss					Perplexity				
	0.6B	1.7B		4B		0.6B	1.7B		4B	
	2e-5	2e-5	1e-5	2e-5	5e-6	2e-5	2e-5	1e-5	2e-5	5e-6
Alpaca	2.38	**2.23**	2.29	2.29	**2.18**	10.82	**9.31**	9.92	9.91	**8.88**
Financial News	2.36	**2.17**	2.23	2.13	**2.04**	10.60	**8.78**	9.25	8.41	**7.71**
Financial Qa (train)	2.12	**2.01**	2.12	2.12	**2.01**	8.29	**7.44**	8.29	8.29	**7.43**
Financial Repor	2.11	**2.00**	2.10	2.11	**2.01**	8.21	**7.40**	8.19	8.25	**7.43**
Fingpt	2.31	**2.15**	2.25	2.23	**2.11**	10.04	**8.62**	9.51	9.34	**8.24**
Fiqa	2.40	**2.25**	2.31	2.31	**2.19**	11.02	**9.45**	10.10	10.05	**8.93**
Twitter	2.21	**2.10**	2.21	2.20	**2.09**	9.14	**8.18**	9.10	8.99	**8.05**
Wikitext	2.24	**2.11**	2.21	2.19	**2.08**	9.41	**8.23**	9.08	8.89	**8.00**
Average	**2.27**	**2.13**	**2.21**	**2.20**	**2.09**	**9.69**	**8.42**	**9.18**	**9.02**	**8.09**

Table 4.15 – Financial News Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	3.92	2.71	2.15	50.40	15.05	8.58
Financial QA (2e-5)	**2.36**	**2.17**	2.13	**10.60**	**8.78**	8.41
Financial QA (1.7B: 1e-5, 4B: 5e-6)	**2.36**	2.23	2.04	**10.60**	9.25	7.71
FinGPT (2e-5)	3.36	2.45	2.07	28.72	11.58	7.92
FiQA (2e-5)	3.90	2.54	**2.01**	49.22	12.74	**7.43**
Mixed Financial (2e-5)	4.03	3.05	2.63	56.35	21.19	13.84
Mixed Wiki+Financial (2e-5)	3.65	3.13	2.77	38.68	22.79	15.91
Financial News (2e-5)	3.96	3.13	2.86	52.25	22.91	17.47
SEC Reports (2e-5)	3.71	3.08	2.81	40.85	21.65	16.67
Twitter Financial (2e-5)	3.17	2.80	2.87	23.77	16.48	17.67
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	3.17	2.65	2.54	23.77	14.10	12.68
WikiText (2e-5)	2.62	2.93	3.37	13.70	18.78	29.19
WikiText (1.7B: 5e-6, 4B: 3e-6)	2.62	3.52	3.27	13.70	33.66	26.44

Table 4.16 – SEC Reports Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.54	2.85	2.11	93.56	17.26	8.25
Financial QA (2e-5)	2.11	**2.00**	2.11	8.21	**7.40**	8.25
Financial QA (1.7B: 1e-5, 4B: 5e-6)	2.11	2.10	2.01	8.21	8.19	7.43
FinGPT (2e-5)	3.53	2.31	1.82	33.97	10.12	6.20
FiQA (2e-5)	4.42	2.53	**1.81**	83.48	12.51	**6.14**
Mixed Financial (2e-5)	4.94	3.58	3.11	139.62	35.83	22.36
Mixed Wiki+Financial (2e-5)	4.35	3.69	3.33	77.57	40.17	27.91
Financial News (2e-5)	4.85	3.73	3.51	127.73	41.68	33.46
SEC Reports (2e-5)	3.72	2.96	2.77	41.12	19.36	15.91
Twitter Financial (2e-5)	2.48	2.32	2.80	11.95	10.17	16.42
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	2.48	2.16	2.39	11.95	8.70	10.93
WikiText (2e-5)	**1.39**	3.27	3.44	**3.99**	26.46	31.23
WikiText (1.7B: 5e-6, 4B: 3e-6)	**1.39**	3.91	3.75	**3.99**	49.83	42.41

Table 4.17 – Alpaca Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.16	2.75	2.11	63.73	15.61	8.22
Financial QA (2e-5)	2.38	**2.23**	2.29	10.82	**9.31**	9.91
Financial QA (1.7B: 1e-5, 4B: 5e-6)	2.38	2.29	2.18	10.82	9.92	8.88
FinGPT (2e-5)	3.57	2.55	2.11	35.55	12.78	8.27
FiQA (2e-5)	4.14	2.56	**1.96**	62.97	12.96	**7.12**
Mixed Financial (2e-5)	4.54	3.38	2.97	93.35	29.53	19.50
Mixed Wiki+Financial (2e-5)	4.07	3.48	3.15	58.56	32.38	23.23
Financial News (2e-5)	4.57	3.61	3.39	96.31	36.92	29.75
SEC Reports (2e-5)	3.86	3.14	2.92	47.65	23.04	18.54
Twitter Financial (2e-5)	3.01	2.66	2.96	20.21	14.33	19.20
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	3.01	2.54	2.61	20.21	12.66	13.65
WikiText (2e-5)	**2.22**	3.24	3.48	**9.23**	25.51	32.38
WikiText (1.7B: 5e-6, 4B: 3e-6)	**2.22**	3.79	3.64	**9.23**	44.22	38.06

Table 4.18 – FinGPT Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.71	2.99	2.22	111.65	19.85	9.18
Financial QA (2e-5)	2.31	2.15	2.23	10.04	8.62	9.34
Financial QA (1.7B: 1e-5, 4B: 5e-6)	2.31	2.25	2.11	10.04	9.51	8.24
FinGPT (2e-5)	3.49	2.26	**1.74**	32.78	9.56	**5.67**
FiQA (2e-5)	4.67	2.71	1.95	107.25	15.08	7.01
Mixed Financial (2e-5)	5.04	3.63	3.14	153.94	37.82	23.08
Mixed Wiki+Financial (2e-5)	4.44	3.75	3.37	84.43	42.50	28.92
Financial News (2e-5)	5.08	3.90	3.64	160.92	49.56	38.03
SEC Reports (2e-5)	3.97	3.15	2.93	53.18	23.41	18.68
Twitter Financial (2e-5)	2.74	2.50	2.91	15.53	12.23	18.34
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	2.74	2.34	2.54	15.53	10.41	12.69
WikiText (2e-5)	**1.30**	**2.11**	3.57	**3.67**	**8.27**	35.50
WikiText (1.7B: 5e-6, 4B: 3e-6)	**1.30**	4.07	3.88	**3.67**	58.55	48.30

Table 4.19 – FiQA Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.29	2.87	2.22	73.12	17.63	9.22
Financial QA (2e-5)	2.40	**2.25**	2.31	11.02	**9.45**	10.05
Financial QA (1.7B: 1e-5, 4B: 5e-6)	2.40	2.31	2.19	11.02	10.10	8.93
FinGPT (2e-5)	3.57	2.55	2.10	35.64	12.79	8.16
FiQA (2e-5)	4.17	2.56	**1.96**	64.75	12.99	**7.08**
Mixed Financial (2e-5)	4.63	3.46	3.05	102.47	31.85	21.20
Mixed Wiki+Financial (2e-5)	4.14	3.56	3.24	63.03	35.04	25.61
Financial News (2e-5)	4.62	3.65	3.46	101.32	38.68	31.69
SEC Reports (2e-5)	3.85	3.14	2.96	47.22	23.15	19.34
Twitter Financial (2e-5)	2.98	2.66	3.00	19.67	14.26	20.09
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	2.98	2.50	2.61	19.67	12.20	13.61
WikiText (2e-5)	**2.07**	3.14	3.53	**7.89**	23.15	34.03
WikiText (1.7B: 5e-6, 4B: 3e-6)	**2.07**	3.85	3.74	**7.89**	46.81	42.04

Table 4.20 – Twitter Financial Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.78	2.99	2.19	118.74	19.82	8.97
Financial QA (2e-5)	2.21	**2.10**	2.20	9.14	**8.18**	8.99
Financial QA (1.7B: 1e-5, 4B: 5e-6)	2.21	2.21	2.09	9.14	9.10	8.05
FinGPT (2e-5)	3.68	2.40	**1.87**	39.54	11.05	**6.46**
FiQA (2e-5)	4.66	2.65	1.88	105.32	14.10	6.58
Mixed Financial (2e-5)	5.21	3.76	3.25	182.63	42.91	25.72
Mixed Wiki+Financial (2e-5)	4.59	3.88	3.48	98.13	48.42	32.48
Financial News (2e-5)	5.11	3.91	3.66	165.22	49.88	38.98
SEC Reports (2e-5)	3.94	3.13	2.90	51.30	22.86	18.12
Twitter Financial (2e-5)	2.53	2.40	2.88	12.60	11.02	17.83
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	2.53	2.22	2.47	12.60	9.21	11.81
WikiText (2e-5)	**1.45**	2.78	3.52	**4.26**	16.06	33.71
WikiText (1.7B: 5e-6, 4B: 3e-6)	**1.45**	4.08	3.88	**4.26**	58.98	48.48

Table 4.21 – Financial QA Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.77	2.95	2.15	117.40	19.11	8.56
Financial QA (2e-5)	**2.12**	**2.01**	2.12	**8.29**	**7.44**	8.29
Financial QA (1.7B: 1e-5, 4B: 5e-6)	**2.12**	2.12	2.01	**8.29**	8.29	7.43
FinGPT (2e-5)	3.66	2.38	**1.83**	38.96	10.85	**6.24**
FiQA (2e-5)	4.64	2.60	1.84	103.40	13.53	6.32
Mixed Financial (2e-5)	5.21	3.75	3.23	183.72	42.30	25.14
Mixed Wiki+Financial (2e-5)	4.58	3.87	3.46	97.49	47.94	31.76
Financial News (2e-5)	5.11	3.90	3.66	166.10	49.53	38.90
SEC Reports (2e-5)	3.90	3.08	2.86	49.30	21.77	17.39
Twitter Financial (2e-5)	2.46	2.32	2.83	11.76	10.15	16.98
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	2.46	2.16	2.43	11.76	8.69	11.39
WikiText (2e-5)	3.40	10.67	3.37	29.90	∞	29.08
WikiText (1.7B: 5e-6, 4B: 3e-6)	3.40	4.07	3.87	29.90	58.33	47.98

Table 4.22 – WikiText Evaluation: Performance Across Training Datasets

Training Dataset	Cross-Entropy Loss			Perplexity		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca (2e-5)	4.63	2.94	2.18	102.41	18.85	8.88
Financial QA (2e-5)	2.24	**2.11**	2.19	9.41	**8.23**	8.89
Financial QA (1.7B: 1e-5, 4B: 5e-6)	2.24	2.21	2.08	9.41	9.08	8.00
FinGPT (2e-5)	3.66	2.44	1.99	38.70	11.46	7.29
FiQA (2e-5)	4.52	2.63	**1.91**	92.13	13.81	**6.72**
Mixed Wiki+Financial (2e-5)	4.41	3.74	3.32	82.10	41.95	27.72
Financial News (2e-5)	4.95	3.81	3.54	140.71	45.17	34.33
SEC Reports (2e-5)	3.89	3.10	2.88	49.02	22.21	17.72
Twitter Financial (2e-5)	2.69	2.47	2.88	14.74	11.78	17.85
Twitter Financial (1.7B: 1e-5, 4B: 5e-6)	2.69	2.30	2.49	14.74	9.94	12.02
WikiText (2e-5)	**1.56**	3.42	3.30	**4.78**	30.63	27.19
WikiText (1.7B: 5e-6, 4B: 3e-6)	**1.56**	3.88	3.65	**4.78**	48.44	38.60

Chapter 5

Discussion

5.1 Key Takeaways

- In-domain diversity beats general corpora for financial pretraining. Mixed Financial achieves lower mean perplexity and lower CV than WikiText and single-dataset alternatives.
- Learning-rate scaling with model size is essential to avoid reverse scaling; proper LR restores expected ordering across 0.6B, 1.7B, 4B.
- Dataset size and format strongly determine transfer. Long-form models transfer across long-form tasks better than across formats; short-form data (Twitter) is highly specialized.

5.2 Practical Guidance

Use Mixed Financial with 50cap when seeking broad financial capabilities; specialize with News/SEC for document analysis; prefer 1.7B for efficiency, 4B for maximum quality (with LR tuning).

Chapter 6

Conclusion

This shortened thesis preserves the core findings: (i) in-domain mixtures deliver the best financial pretraining, (ii) learning-rate scaling resolves reverse scaling, and (iii) dataset size/format drive transfer. We provide complete figures and tables to enable independent evaluation and reuse. Future work should explore dynamic mixtures, larger model scales, and expanded downstream tasks.

Bibliography

Aharoni, Roee and Yoav Goldberg (2020). "Unsupervised Domain Clusters in Pretrained Language Models". In: *arXiv preprint arXiv:2004.02105*. URL: https://arxiv.org/abs/2004.02105.

Araci, Dogu (2019). "Finbert: Financial sentiment analysis with pre-trained language models". In: *arXiv preprint arXiv:1908.10063*.

Arivazhagan, Naveen, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Zhifeng Chen, and Yonghui Wu (2019). "Massively Multilingual Neural Machine Translation in the Wild: Findings and Challenges". In: *arXiv preprint arXiv:1907.05019*. URL: http://arxiv.org/abs/1907.05019.

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. (2020). "Language models are few-shot learners". In: *Advances in neural information processing systems* 33, pp. 1877–1901.

Chen, Zhiyu, Wenhu Chen, Ziyu Fan, Shiyang Chang, and William Yang Wang (2021). "FinQA: A Dataset of Numerical Reasoning over Financial Data". In: *arXiv preprint arXiv:2109.00122*. URL: https://arxiv.org/abs/2109.00122.

French, Robert M (1999). "Catastrophic forgetting in connectionist networks". In: *Trends in Cognitive Sciences* 3.4, pp. 128–135. DOI: 10.1016/S1364-6613(99)01294-2.

Gao, Leo, Stella Biderman, Sidney Black, Laurence Anthony, Xenia Golding, Horace Hoppe, Connor Foster, Jason Phang, Anish He, Aman Thite, Andy Nabeshima, Shawn Presser, and Connor Leahy (2021). "The Pile: An 800GB Dataset of Diverse Text for Language Modeling". In: *arXiv preprint arXiv:2101.00027*. URL: https://arxiv.org/abs/2101.00027.

Gururangan, Suchin, 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". In: *arXiv preprint arXiv:2004.10964*.

Hoffmann, Jordan, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. (2022). "Training compute-optimal large language models". In: *arXiv preprint arXiv:2203.15556*.

Kaplan, Jared, 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". In: *arXiv preprint arXiv:2001.08361*.

Kirkpatrick, James, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis,

Claudia Clopath, Dharshan Kumaran, and Raia Hadsell (2017). "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences* 114.13, pp. 3521–3526. DOI: 10.1073/pnas.1611835114.

Longpre, Shayne, Yao Hou, Aakanksha Deshpande, He He, Thibault Sellam, Alex Tamkin, Slav Petrov, Denny Zhou, Jason Wei, Yi Tay, Quoc V. Le, et al. (2023). "A Pretrainer's Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity". In: *arXiv preprint arXiv:2305.13169*. URL: https://arxiv.org/abs/2305.13169.

McCandlish, Sam, Jared Kaplan, Dario Amodei, and OpenAI Dota Team (2018). "An Empirical Model of Large-Batch Training". In: *arXiv preprint arXiv:1812.06162*. URL: https://arxiv.org/abs/1812.06162.

McCloskey, Michael and Neal J. Cohen (1989). "Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem". In: *Psychology of Learning and Motivation*. Elsevier, pp. 109–165. DOI: 10.1016/S0079-7421(08)60536-8.

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

Narayanan, Deepak, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Anand Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, Amar Phanishayee, and Matei Zaharia (2021). "Efficient Large-Scale Language Model Training on GPU Clusters Using Megatron-LM". In: *arXiv preprint arXiv:2104.04473*. URL: https://arxiv.org/abs/2104.04473.

Quiñonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, eds. (2008). *Dataset Shift in Machine Learning*. MIT Press. DOI: 10.7551/mitpress/9780262170055.001.0001. URL: https://doi.org/10.7551/mitpress/9780262170055.001.0001.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. (2019). "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8, p. 9.

Raffel, Colin, 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". In: *Journal of Machine Learning Research* 21, 140:1–140:67. URL: https://jmlr.org/papers/v21/20-074.html.

Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He (2020). "ZeRO: Memory optimizations Toward Training Trillion Parameter Models". In: *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, pp. 1–16. DOI: 10.1109/SC41405.2020.00024. URL: https://doi.org/10.1109/SC41405.2020.00024.

Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation) (2016). Official Journal of the European Union. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

Sanh, Victor, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. (2022). "Multitask Prompted Training Enables Zero-Shot Task Generalization". In: *arXiv preprint arXiv:2110.08207*. URL: https://arxiv.org/abs/2110.08207.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. (2023). "Llama: Open and efficient foundation language models". In: *arXiv preprint arXiv:2302.13971*.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. URL: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wu, Shijie, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David S. Rosenberg, and Gideon Mann (2023). "BloombergGPT: A Large Language Model for Finance". In: *arXiv preprint arXiv:2303.17564*. URL: https://arxiv.org/abs/2303.17564.

Xia, Mengzhou, Tianyu Gao, Zhiyuan Zeng, and Danqi Chen (2023). "Sheared llama: Accelerating language model pre-training via structured pruning". In: *arXiv preprint arXiv:2310.06694*.

Xie, Sang Michael, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy Liang, Quoc V. Le, Tengyu Ma, and Adams Wei Yu (2023). "DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining". In: *arXiv preprint arXiv:2305.10429*. URL: https://arxiv.org/abs/2305.10429.

Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. (2024). "Qwen2 Technical Report". In: *arXiv preprint arXiv:2407.10671*.

Yang, Hongyang, Xiao-Yang Liu, and Christina Dan Wang (2023). "FinGPT: Open-Source Financial Large Language Models". In: *arXiv preprint arXiv:2306.06031*. URL: https://arxiv.org/abs/2306.06031.

Yang, Yi, Mark Christopher Siy UY, and Allen Huang (2020). "FinBERT: A Pretrained Language Model for Financial Communications". In: *arXiv preprint arXiv:2006.08097*. URL: https://arxiv.org/abs/2006.08097.