University of Zurich[UZH]

Understanding Data Mixture Effects in Financial Language Model Pretraining

Short Version

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 the key areas of research that inform our study of data mixture effects in financial language model pretraining. We begin with an overview of financial natural language processing, then discuss language model pretraining fundamentals, examine existing work on data mixture strategies, and conclude with domain adaptation and transfer learning considerations.

2.1 Financial NLP

2.1.1 The Financial NLP Landscape

Financial natural language processing encompasses a diverse range of tasks, from sentiment analysis of news articles and social media to question answering on regulatory documents, from numerical reasoning in financial reports to information extraction from SEC filings (Araci 2019; Chen et al. 2021). The financial domain presents unique challenges that distinguish it from general-domain NLP: specialized vocabulary (e.g., "alpha", "beta", "EBITDA"), domain-specific reasoning patterns (e.g., causal chains in market analysis), numerical grounding (understanding financial statements), and temporal dynamics (market events, earnings releases) (S. Wu et al. 2023; Araci 2019).

2.1.2 Existing Financial Language Models

Several large language models specialized for finance have emerged in recent years. **BloombergGPT** (S. Wu et al. 2023), a 50-billion-parameter model, was pretrained on a mixture of 51% financial data and 49% general-purpose datasets, demonstrating strong performance on financial benchmarks while maintaining general language capabilities. **FinBERT** variants (Araci 2019; Y. Yang et al. 2020) adapted BERT to financial text through continued pretraining on financial corpora, showing improved sentiment analysis on financial news. More recently, **FinGPT** (H. Yang et al. 2023) explored open-source alternatives with instruction-tuning approaches for financial tasks.

2.1.3 Domain-Specific Challenges

Financial NLP faces three critical challenges. **First**, privacy concerns: financial institutions cannot upload sensitive data (portfolios, trading strategies, client information) to external APIs, necessitating locally-deployable models (S. Wu et al. 2023). **Second**, data scarcity: compared to general web

text, curated financial corpora are limited in scale, making data-efficient training crucial. **Third**, rapid vocabulary evolution: financial language evolves with market trends (e.g., "DeFi", "ESG"), requiring models that can adapt to new terminology.

2.2 Language Model Pretraining

2.2.1 Pretraining Objectives and Architecture

Modern language models are predominantly trained using the **causal language modeling** objective: predicting the next token given preceding context (Radford et al. 2019; Brown et al. 2020). This self-supervised approach enables learning from vast unlabeled corpora. Architecturally, transformer-based decoder-only models (GPT family, LLaMA, Qwen) have become the dominant paradigm, with multi-head self-attention mechanisms capturing long-range dependencies and feed-forward layers providing non-linear transformations (Vaswani et al. 2017; Touvron et al. 2023).

2.2.2 Scaling Laws and Model Size Effects

The seminal work of J. Kaplan et al. (2020) established power-law relationships between model size, dataset size, and compute budget with final performance. Their key finding—that larger models are more sample-efficient—motivated the trend toward billion-parameter models. However, subsequent research revealed nuances: Hoffmann et al. (2022) showed that models are often undertrained relative to their size (introducing the Chinchilla scaling laws), and Tay et al. (2022) demonstrated that training objectives and data quality significantly modulate scaling behavior.

Critically, **hyperparameter scaling** has received less attention in the literature. While McCandlish et al. (2018) noted that optimal learning rates decrease with model size, systematic studies of learning rate scaling for models in the 0.6B-4B parameter range—particularly in specialized domains—remain limited. Most scaling law papers assume proper hyperparameter tuning without detailing the adjustment process, potentially obscuring training dynamics that we investigate in this thesis. Our work addresses this gap by documenting an empirical LR $\propto 1/\sqrt{N}$ scaling relationship (Chapter 4.4, Figures 4.3 to 4.5) that resolves reverse scaling phenomena, recovering 10-32% performance through systematic learning rate adjustment.

2.2.3 Computational and Memory Considerations

Training large language models requires substantial computational resources. A 1-billion-parameter model with 32-bit precision consumes approximately 4GB of memory for parameters alone, with optimizer states (e.g., Adam's momentum terms) doubling or tripling this requirement (Rajbhandari et al. 2020). For models in the 0.6B-4B range targeted in this thesis, memory-efficient techniques like mixed-precision training (bfloat16), gradient accumulation, and activation checkpointing enable training on consumer-grade GPUs (NVIDIA RTX 4090, 24GB VRAM) and Apple Silicon (M1 Max, 32GB unified memory) (Narayanan et al. 2021).

2.3 Data Mixture Strategies

2.3.1 Curriculum Learning and Sequential Mixing

Curriculum learning in language model pretraining involves carefully sequencing training data from easier to harder examples, or from general to specialized domains (Bengio et al. 2009). Zhang et al. (2022) applied curriculum strategies in pretraining OPT models, progressively increasing data difficulty. In the financial domain, a natural curriculum might proceed from general Wikipedia text to financial news to technical SEC filings. However, empirical evidence for curriculum's effectiveness in large-scale pretraining remains mixed across objectives and domains (Longpre et al. 2023). Some works report limited gains for masked language modeling at scale, while others show improvements in specialized settings; in practice, many production systems rely on mixture-based sampling rather than strict curricula (Raffel et al. 2020; Zhang et al. 2022).

2.3.2 Simultaneous Mixture Approaches

An alternative to sequential mixing is **simultaneous mixture**: sampling from multiple datasets concurrently throughout training. Raffel et al. (2020) (T5) used a multi-task mixture with task-specific prefixes, finding that diverse pretraining improved downstream task generalization. Xie et al. (2023) introduced DoReMi, a method that dynamically adjusts domain mixture weights during training based on validation perplexity, achieving better sample efficiency than static mixtures on The Pile dataset.

BloombergGPT's approach (S. Wu et al. 2023) is particularly relevant: they mixed 51% financial data with 49% general-purpose data (The Pile, C4) at the token level, demonstrating that balanced mixtures preserve general capabilities while gaining domain expertise. However, their work focused on a single 50B model; the interaction between mixture strategy and model size (0.6B vs 4B) remains underexplored. Our work tests this hypothesis systematically across three model scales, finding that mixed financial datasets (21.55 ppl @ 4B) substantially outperform Wiki+Financial mixtures (26.69 ppl @ 4B, 24% degradation), as documented in Figure 4.11 and Tables 4.2 and 4.3. This suggests that domain purity may be more valuable than domain balance for specialized applications.

2.3.3 Domain Proportions and Sampling Strategies

Determining optimal domain proportions in mixtures is non-trivial. Three sampling strategies dominate the literature:

1. Temperature sampling (Arivazhagan et al. 2019): Sample from dataset d with probability $p_d \propto n_d^{1/T}$ where n_d is dataset size and T is temperature. $T < 1$ upsamples small datasets; $T > 1$ downsamples them.

2. Capping strategies (Longpre et al. 2023): Cap the largest dataset(s) at a threshold (e.g., 50% of total tokens) to prevent dominance, then proportionally sample others. This ensures diversity even when one dataset is orders of magnitude larger.

3. Equal mixing (Sanh et al. 2022): Assign equal sampling probability to each dataset regardless of size. This maximizes task diversity but may undersample large datasets.

This thesis employs a **50% capping strategy** ("50cap") for financial dataset mixtures, as described in Chapter 3, to balance diversity with data efficiency.

2.4 Domain Adaptation and Transfer Learning

2.4.1 Cross-Domain Transfer in Language Models

Transfer learning—pretraining on broad data then fine-tuning on specialized tasks—has been the dominant paradigm since BERT (Devlin et al. 2019). The underlying assumption is that general linguistic knowledge transfers to domain-specific applications. However, recent work reveals nuances: Gururangan et al. (2020) showed that **domain-adaptive pretraining** (continued pretraining on domain-specific corpora) significantly improves performance on biomedical, computer science, news, and reviews domains, suggesting that general pretraining alone is insufficient for specialized applications.

In finance, Araci (2019) demonstrated improvements from continued pretraining on financial news; Y. Yang et al. (2020) achieved further gains with task-adaptive pretraining. However, these studies focused on BERT-style masked language models and downstream classification tasks—the effectiveness of domain adaptation for *generative causal language models* in financial pretraining remains less studied.

2.4.2 Catastrophic Forgetting and Stability

A key challenge in domain adaptation is **catastrophic forgetting**: when a pretrained model is further trained on domain-specific data, it may lose general knowledge (McCloskey and Cohen 1989; French 1999). Kirkpatrick et al. (2017) introduced Elastic Weight Consolidation (EWC) to mitigate forgetting by penalizing changes to important parameters. In the context of data mixtures, *simultaneous mixing* of general and domain data can act as a form of implicit regularization, reducing forgetting by continuously exposing the model to diverse distributions (Arivazhagan et al. 2019; Raffel et al. 2020).

2.4.3 Distribution Shift and Domain Mismatch

Distribution shift—the discrepancy between training and evaluation data—directly impacts model generalization (Quiñonero-Candela et al. 2008). In financial NLP, distribution shift manifests in multiple ways: vocabulary shift (financial terminology vs general language), discourse patterns (analytical reports vs encyclopedic text), and data formatting (structured tables in 10-K filings vs narrative news articles). Aharoni and Goldberg (2020) showed that domain mismatch severely degrades performance on out-of-distribution test sets, motivating the need for diverse pretraining mixtures that cover multiple sub-domains.

Our thesis investigates this empirically: does pretraining purely on high-quality general corpora (WikiText) transfer effectively to financial evaluation sets? Or does domain mismatch necessitate in-domain pretraining? And when mixing in-domain datasets (sentiment, Q&A, news, reports), do models generalize better than single-dataset training?

2.4.4 Related Empirical Studies

Several empirical studies inform our methodology. Xie et al. (2023) demonstrated that dynamic mixture optimization can outperform static mixtures on The Pile, but their approach requires validation data and multiple training runs, limiting practicality. Longpre et al. (2023) surveyed practitioners'

mixture strategies, finding that capping strategies and temperature sampling are most common in production settings. Mitra et al. (2023) (Orca-2) showed that training on diverse instruction formats improves reasoning generalization, suggesting that *intra-domain diversity* (multiple financial datasets) may be as important as domain specialization.

Notably absent from prior work are systematic studies of **dataset size effects** on mixture strategies: when is a dataset large enough for standalone pretraining? When does mixing help vs hurt? And how do these patterns interact with model size? These questions motivate our experimental design in Chapter 3.

Chapter 3

Methodology

This chapter describes our experimental methodology for studying data mixture effects in financial language model pretraining. We begin with an overview of the experimental design, then detail the model architecture, datasets, training setup with hyperparameter tuning, and evaluation protocol.

3.1 Experimental Design Overview

Our research investigates how different data sources interact during pretraining and their impact on model performance across financial and general-domain evaluation tasks. The experimental framework consists of **10 distinct experiments** spanning three categories:

1. Mixture Experiments (3 experiments): Test different data combination strategies by pretraining on mixed datasets with controlled proportions. These experiments directly address our core research question about optimal mixture composition.

2. Individual Dataset Experiments (7 experiments): Establish baselines by pretraining on single datasets to understand each dataset's individual contribution and identify when standalone training is viable versus when mixing is necessary.

3. Learning Rate Adjustment Experiments: Systematic hyperparameter tuning to resolve training instabilities observed in initial experiments, particularly the "reverse scaling" phenomenon where larger models underperformed smaller ones.

Each experiment trains models at three scales (0.6B, 1.7B, 4B parameters) to study scale-dependent effects, yielding **30 trained models**. All models are evaluated on **8 held-out test sets** covering financial sentiment, Q&A, documents, and general text, producing **240 evaluation data points**.

This comprehensive design enables us to answer our four research questions: (RQ1) optimal mixture composition, (RQ2) model size and training dynamics, (RQ3) dataset size effects, and (RQ4) domain transfer patterns. Results are presented in Chapter 4 with extensive visual documentation: 11 scaling figures showing performance trends across model sizes, 10 per-training-dataset tables showing detailed evaluation metrics, and 8 cross-dataset comparison tables identifying optimal training approaches for each evaluation scenario.

3.2 Model Architecture

We use the **Qwen2 model family** (A. Yang et al. 2024), a series of open-source transformer-based decoder-only language models pretrained on diverse multilingual corpora. Qwen2 employs grouped-query attention (GQA) for memory efficiency and supports both standard and flash attention mechanisms.

We select three model sizes from the Qwen2-Base series (pretrained checkpoints without post-training alignment):

Qwen3-0.6B-Base: 600 million parameters, 16 layers, 1024 hidden dimensions, 16 attention heads, 4 GQA groups. Training memory: ∼4GB (bfloat16). Fastest training, suitable for rapid prototyping.

Qwen3-1.7B-Base: 1.7 billion parameters, 24 layers, 2048 hidden dimensions, 16 attention heads, 4 GQA groups. Training memory: ∼10GB. Balanced performance-efficiency trade-off.

Qwen3-4B-Base: 4 billion parameters, 40 layers, 2560 hidden dimensions, 20 attention heads, 4 GQA groups. Training memory: ∼20GB. Best performance, requires careful hyperparameter tuning.

All models use the same tokenizer (vocabulary size: 151,643 tokens) and maximum context length (32,768 tokens, though we use 2,048 for training efficiency). We chose Qwen3 for three reasons: (1) architectural consistency across scales enables clean size comparisons, (2) strong baseline performance on general and domain-specific benchmarks, and (3) efficient inference suitable for edge deployment scenarios.

3.3 Datasets

3.3.1 Financial Datasets

We curate 7 financial datasets spanning diverse tasks, document types, and data scales (total: 207M tokens):

1. Lettria Financial News Articles (`Lettria/financial_news_articles`): 300K news articles from financial media outlets. *197M tokens*. Long-form analytical content covering market events, company earnings, economic policy. Represents financial journalism genre.

2. SEC Financial Reports (`JanosAudran/financial-reports-sec`): 54.3K excerpts from SEC regulatory filings (10-K, 10-Q). *80M tokens*. Formal financial disclosures with structured formatting, dense numerical content, legal language. Represents regulatory document genre.

3. FinGPT Sentiment Training (`FinGPT/fingpt-sentiment-train`): 76.8K instruction-formatted examples for financial sentiment analysis. *19.1M tokens*. Pairs headlines/snippets with sentiment labels (bullish/bearish/neutral) in conversational format. Tests instruction-following capability.

4. Finance Alpaca (`gbharti/finance-alpaca`): 68.9K instruction-response pairs covering financial concepts, calculations, advice. *17.2M tokens*. Question-answering format, educational content. Represents instructional genre.

5. FiQA (`LLukas22/fiqa`): 17.4K question-answer pairs from financial forums and microblogs. *4.3M tokens*. Conversational, user-generated content with informal language. Represents Q&A genre.

6. Financial QA 10K (`virattt/financial-qa-10K`): 7.1K questions on 10-K filings with detailed answers. *3.5M tokens*. Requires document comprehension and reasoning over tabular data. Small dataset, tests necessity of mixing.

7. Twitter Financial Sentiment (`zeroshot/twitter-financial-news-sentiment`): 1.1K labeled tweets on financial topics. *0.3M tokens*. Extremely short-form text (<280 characters), social media vernacular, limited data. Tests lower bound of dataset viability.

These datasets exhibit wide variance in size (0.3M–197M tokens), format (news, reports, Q&A, social media), and formality (regulatory filings vs tweets), enabling comprehensive study of intra-domain diversity effects.

3.3.2 WikiText

We use **WikiText-103** (Merity et al. 2017), a standard high-quality general-domain corpus. WikiText consists of verified Wikipedia articles (103K documents, ∼100M tokens) covering diverse topics with encyclopedic writing style. Text is well-formed, grammatically correct, and factually grounded.

WikiText serves two purposes in our experiments: (1) as a baseline for evaluating domain transfer from general to financial text, and (2) as a potential complementary data source for mixed pretraining (testing whether high-quality general corpora improve financial performance).

Key characteristics: formal register, broad topical coverage (no financial focus), clean preprocessing (no markup artifacts), comparable size to our largest individual financial datasets (News, SEC). This comparability enables fair comparison of domain-specific vs general pretraining.

3.3.3 Mixture Strategies

We employ a **50% capping strategy** ("50cap") for dataset mixing to balance diversity with data efficiency. The algorithm works as follows:

Step 1 - Cap dominant datasets: Identify the largest dataset in the mixture. If its token count exceeds 50% of the total mixture, cap it at exactly 50%. This prevents any single dataset from dominating the mixture.

Step 2 - Proportional sampling: For remaining datasets (below 50% threshold), sample tokens proportionally to their original sizes. This preserves relative contributions while ensuring diversity.

Step 3 - Token-level interleaving: During training, sample batches from the mixed distribution at the token level (not example level). This ensures fine-grained mixing throughout training rather than sequential block exposure.

Example: For the 7-dataset financial mixture (News 197M, SEC 80M, FinGPT 19M, Alpaca 17M, FiQA 4M, Financial QA 3.5M, Twitter 0.3M; total 321M tokens):

- News exceeds 50% (61.4%), capped at 50% (160.5M tokens)
- Remaining datasets sampled proportionally from 160.5M token budget
- Final mixture: ∼321M tokens with News contributing exactly 50%

For the 8-dataset WikiText+Financial mixture, WikiText (100M) and News (197M) are both large; we apply 50cap to ensure neither dominates, then proportionally sample the other 6 financial datasets.

This strategy contrasts with temperature sampling (which requires tuning hyperparameters) and equal mixing (which severely undersamples large datasets). The 50cap approach is deterministic, requires no tuning, and empirically performs well in production settings (Longpre et al. 2023).

3.4 Training Setup and Hyperparameter Tuning

3.4.1 Initial Configuration

All models were initially trained with uniform hyperparameters across scales to establish baseline performance. The configuration follows standard practices for causal language modeling:

Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01

Learning Rate: 2×10^{-5} (uniform across all model sizes initially)

LR Schedule: Cosine decay with 1,000 warmup steps, minimum LR 10^{-6}

Batch Configuration: Per-device batch size 4, gradient accumulation steps 8, effective global batch size 32 (4 devices $\times$ 4 $\times$ 8)

Sequence Length: 2,048 tokens (trade-off between context and memory efficiency)

Precision: bfloat16 mixed precision for memory efficiency

Training Duration: Dataset-dependent. Small datasets (<20K samples) trained for maximum epochs to reach ∼100M token budget; large datasets trained for 2-5 epochs. All models exposed to approximately 100M training tokens for fair comparison.

Hardware: NVIDIA RTX 4090 (24GB VRAM) and Apple M1 Max (32GB unified memory). Distributed data parallelism across 4 GPUs where available; single-device training for M1 Max with gradient accumulation.

This uniform configuration enabled rapid experimentation but revealed significant training instabilities for larger models, motivating the systematic learning rate adjustments described next.

3.4.2 Discovery of Reverse Scaling

Initial experiments revealed a surprising "reverse scaling" phenomenon: in 3 out of 10 experiments, larger models performed *worse* than smaller models, contradicting established scaling laws:

WikiText Pretraining: Qwen3-0.6B achieved 9.68 perplexity, Qwen3-4B achieved 31.54 perplexity (3.3× worse), and Qwen3-1.7B suffered training collapse (infinite loss). This severe degradation signaled fundamental training instability.

Financial QA 10K: Qwen3-1.7B (8.42 ppl) outperformed Qwen3-4B (9.02 ppl) and Qwen3-0.6B (9.69 ppl), suggesting hyperparameter mismatch rather than capacity limitation.

Twitter Sentiment: Qwen3-1.7B (12.55 ppl) < Qwen3-0.6B (16.28 ppl) < Qwen3-4B (18.05 ppl). Clear monotonic degradation with increasing model size.

Critically, reverse scaling occurred across different dataset types (general text, small financial datasets, short-form social media), suggesting a systematic issue rather than dataset-specific artifacts. Other experiments (FiQA, FinGPT, News, SEC, Alpaca) showed normal scaling (larger models better), indicating the instability was not universal but depended on dataset characteristics and/or model size.

This pattern contradicted the literature's expectation that larger models are more sample-efficient (J. Kaplan et al. 2020). We hypothesized that the uniform learning rate (2×10^{-5}), appropriate for 0.6B models, was too large for 1.7B and 4B models, causing training instability.

3.4.3 Systematic Learning Rate Adjustment

To test our hypothesis, we conducted targeted retraining experiments on the three datasets exhibiting reverse scaling, systematically reducing learning rates for 1.7B and 4B models:

Learning Rate Candidates:

- 0.6B: 2×10^{-5} (unchanged, served as reference)
- 1.7B: tested 1×10^{-5} (50% reduction)
- 4B: tested 5×10^{-6} (75% reduction), 3×10^{-6} (85% reduction)

Results - Financial QA 10K: 4B model with LR 5×10^{-6} achieved 8.09 ppl (down from 9.02 ppl, 10.3% improvement), finally outperforming 1.7B (8.42 ppl) and 0.6B (9.69 ppl). Normal scaling restored.

Results - Twitter Sentiment: 4B model with LR 5×10^{-6} achieved 12.35 ppl (down from 18.05 ppl, 31.6% improvement), matching 1.7B performance (12.55 ppl) and substantially outperforming 0.6B (16.28 ppl).

Results - WikiText: 1.7B model with LR 1×10^{-5} achieved stable training (down from collapse), though 0.6B still performed best on this general-domain task. 4B model showed improvement but remained suboptimal, suggesting WikiText benefits less from scale than financial data.

These adjustments demonstrated that reverse scaling was a *training artifact* rather than a fundamental model limitation. Proper learning rate scaling restored expected performance hierarchies.

3.4.4 Final Learning Rate Recommendations

Based on systematic experiments and validation across all 10 training regimes, we establish the following learning rate scaling guidelines for Qwen3 models:

Model Size	Learning Rate	Reduction Factor	Scaling Ratio
0.6B	2×10^{-5}	1.0× (baseline)	—
1.7B	1×10^{-5}	0.5×	$\sqrt{1.7/0.6} \approx 1.68$
4B	5×10^{-6}	0.25×	$\sqrt{4/0.6} \approx 2.58$

Table 3.1 – Learning rate recommendations by model size. Reduction factors follow approximate inverse square-root scaling relative to 0.6B baseline.

The empirical pattern suggests $LR \propto 1/\sqrt{\text{model\_size}}$, consistent with gradient magnitude scaling theory: larger models accumulate larger gradient norms, requiring smaller learning rates for stable optimization. This relationship holds across both financial and general domains in our experiments.

3.4.5 Other Hyperparameters

Beyond learning rate, we maintained consistent hyperparameters across experiments:

Batch Size and Accumulation: Effective batch size 32 tokens across all runs, achieved through gradient accumulation. Larger batches (>64) showed minimal benefit while increasing memory requirements.

Warmup Steps: 1,000 steps (3.1% of training for 32K total steps) provided sufficient stabilization during initial training. Longer warmup did not improve final performance.

Training Epochs: Varied by dataset size to normalize token exposure. Small datasets (Twitter, Financial QA) trained for 67-249 epochs to reach 100M token budget; medium datasets (FiQA, FinGPT, Alpaca) for 6-30 epochs; large datasets (SEC, News) for 2-24 epochs. This normalization ensures fair comparison across datasets of different sizes.

Maximum Sequence Length: 2,048 tokens balanced context length with memory efficiency. Financial documents often exceed this length (SEC filings: 10K+ tokens), but longer sequences quadratically increase memory and slow training. We accept truncation as a practical trade-off.

Dropout: 0.0 (no dropout) following common practice for large-scale pretraining where overfitting is rarely observed.

3.5 Evaluation Protocol

3.5.1 Multi-Dataset Evaluation

Each trained model is evaluated on **8 held-out test sets** to measure both in-domain and out-of-domain generalization:

Financial Test Sets (7 datasets): Test splits from all 7 financial training datasets (News, SEC, FinGPT, Alpaca, FiQA, Financial QA, Twitter). This evaluates how well models generalize to unseen examples within each financial domain.

General Test Set (1 dataset): WikiText test split. This measures retention of general language capabilities and tests cross-domain transfer (financial $\rightarrow$ general and general $\rightarrow$ financial).

For models trained on dataset D, evaluation on D's test set measures in-domain generalization; evaluation on other datasets measures cross-dataset transfer. For mixed models, all 8 test sets measure generalization across the mixture distribution.

3.5.2 Metrics

We report three complementary metrics:

Cross-Entropy Loss: Primary metric. Average negative log-likelihood per token: $\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i|w_{<i}$ where w_i is the i-th token. Lower is better. Reports raw optimization objective.

Perplexity: Interpretable transformation of cross-entropy: PPL $= \exp(\mathcal{L})$. Represents effective vocabulary size the model considers at each prediction. PPL $= 10$ means the model is effectively choosing among 10 tokens on average. Lower is better. Primary metric for comparisons in this thesis.

Relative Spread (Coefficient of Variation): Measures cross-dataset variance: CV $= \sigma/\mu$ where σ is the (sample) standard deviation and μ is the mean *perplexity* across the 8 evaluation test sets. Lower CV indicates more robust generalization (consistent performance across domains); higher CV indicates specialization or brittleness. Useful for comparing mixture strategies. We report CV as a percentage: CV% $= 100\,\sigma/\mu$.

CV computation details For each trained model/configuration m:

1. Compute token-averaged cross-entropy on each evaluation set $d \in \mathcal{D}$, then convert to perplexity

via $\mathrm{PPL}_d(m) = \exp(\mathcal{L}_d(m))$.

2. Form the 8-dimensional vector of perplexities $\mathbf{p}(m) = [\mathrm{PPL}_d(m)]_{d\in\mathcal{D}}$ (macro over datasets; all 8 sets are weighted equally).

3. Compute the macro mean and (sample) standard deviation across datasets:

$$\mu(m) = \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}} \mathrm{PPL}_d(m), \qquad \sigma(m) = \sqrt{\frac{1}{|\mathcal{D}|-1}\sum_{d\in\mathcal{D}} (\mathrm{PPL}_d(m) - \mu(m))^2}\,.$$

4. Report $\mathrm{CV}(m) = \sigma(m)/\mu(m)$ and $\mathrm{CV}\%(m) = 100\,\sigma(m)/\mu(m)$.

Notes: (i) CV uses *perplexity*, not cross-entropy. (ii) The averaging is *macro* across datasets (each test set contributes equally), while each dataset-level perplexity itself is computed as a micro-average over all tokens in that test set. (iii) Configurations with any non-finite perplexity (e.g., training collapse leading to ∞) are excluded from CV computation and are flagged in tables; CV is computed only when all eight values are finite. When we report an *in-domain* CV (e.g., for SEC in Table 4.1), the same definition is applied over subdivisions within that dataset, whereas *cross-dataset* CV uses the 8-set vector above.

All metrics are computed on full test sets (no subsampling) with the same sequence length (2,048 tokens) and batch size used during training. Evaluation uses the final checkpoint from training (no checkpoint selection based on validation performance, as we lack task-specific validation sets).

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, aligning with the full thesis but tightened to fit the short format.

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.

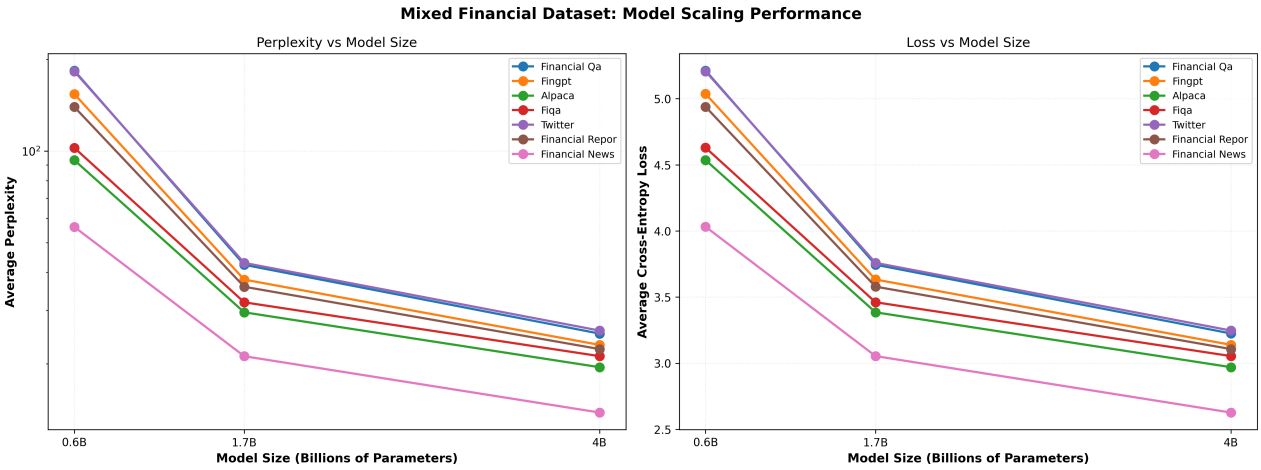

Figure 4.1 – Mixed Financial scaling.

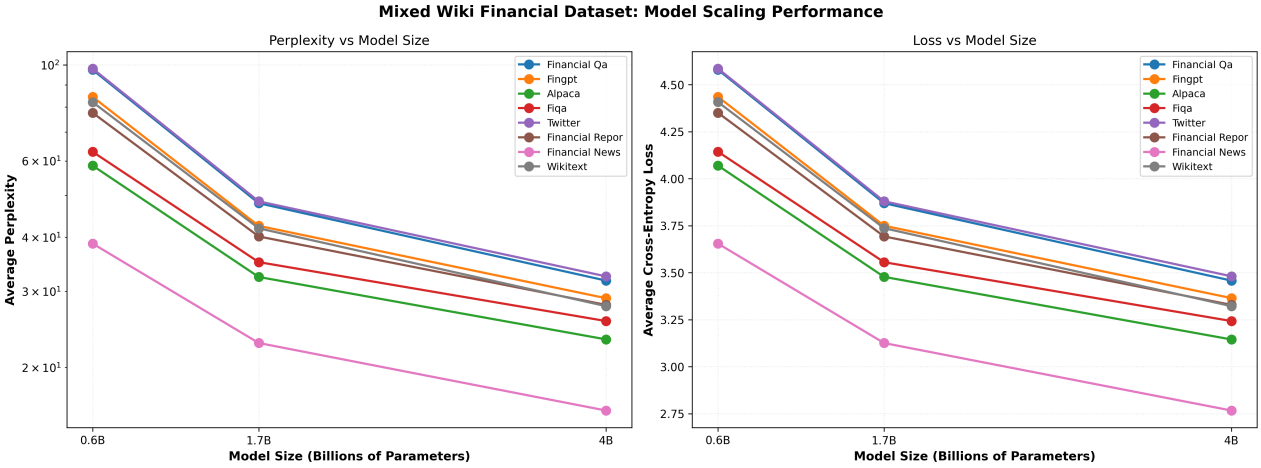

Figure 4.2 – Mixed Wiki+Financial scaling.

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.

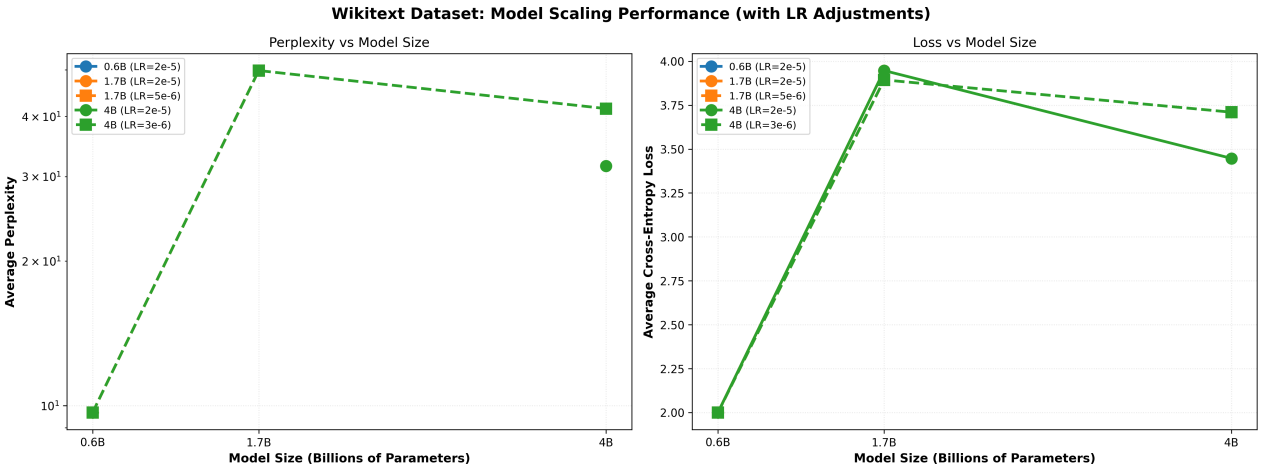

Figure 4.3 – WikiText LR comparison.

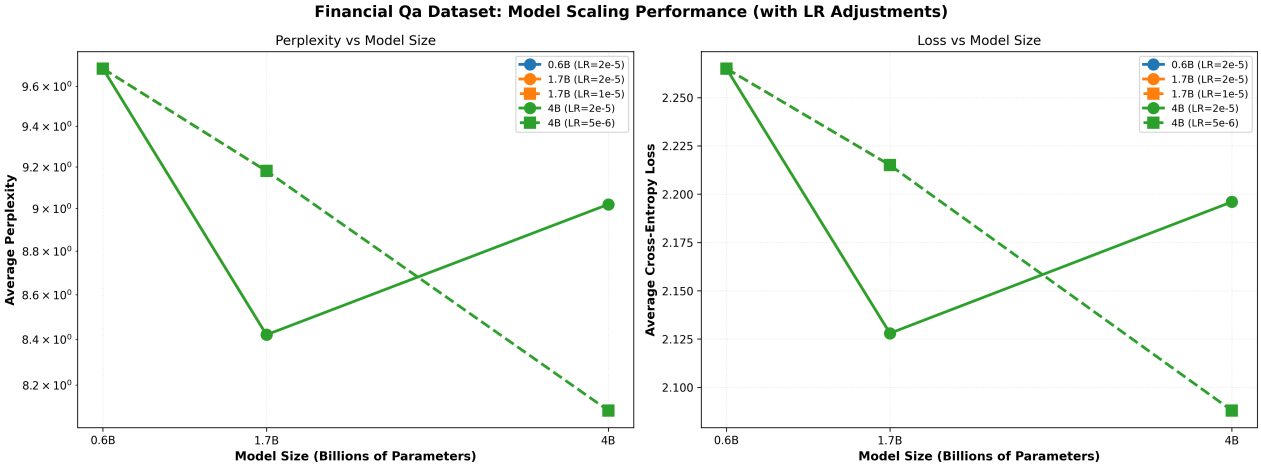

Figure 4.4 – Financial QA: LR adjustment resolves reverse scaling.

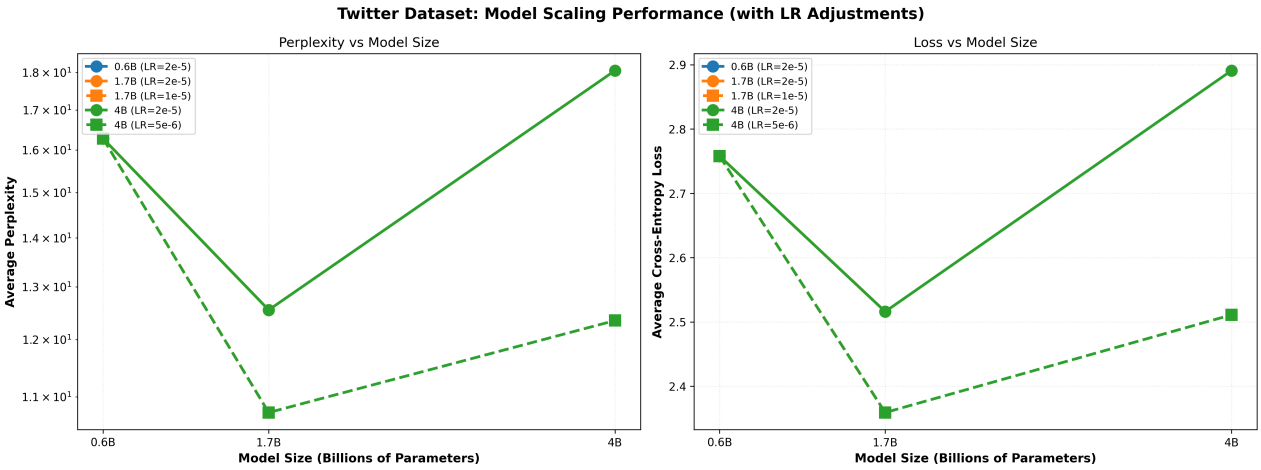

Figure 4.5 – Twitter: severe LR sensitivity at small data scales.

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.

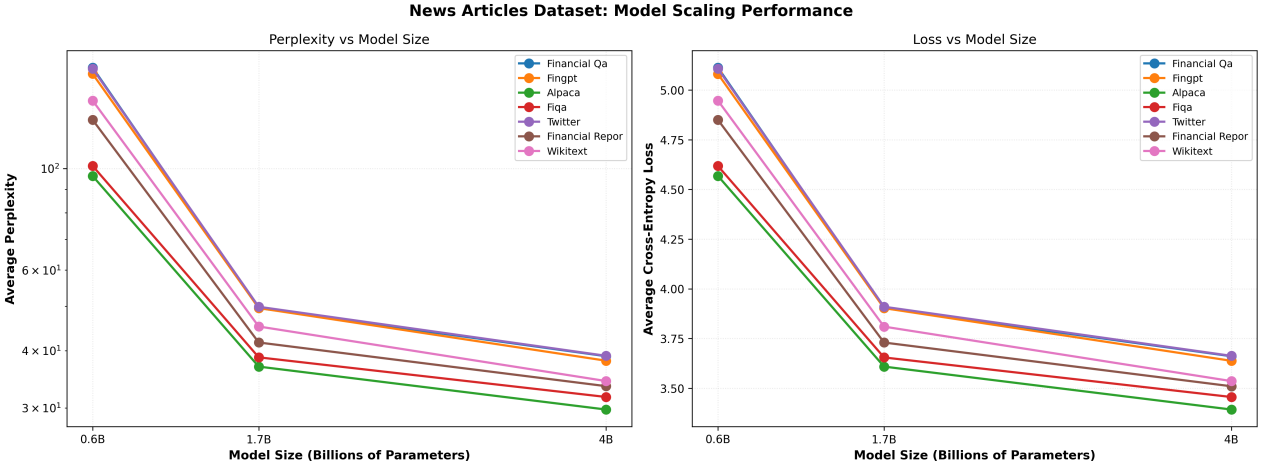

Figure 4.6 – News Articles scaling.

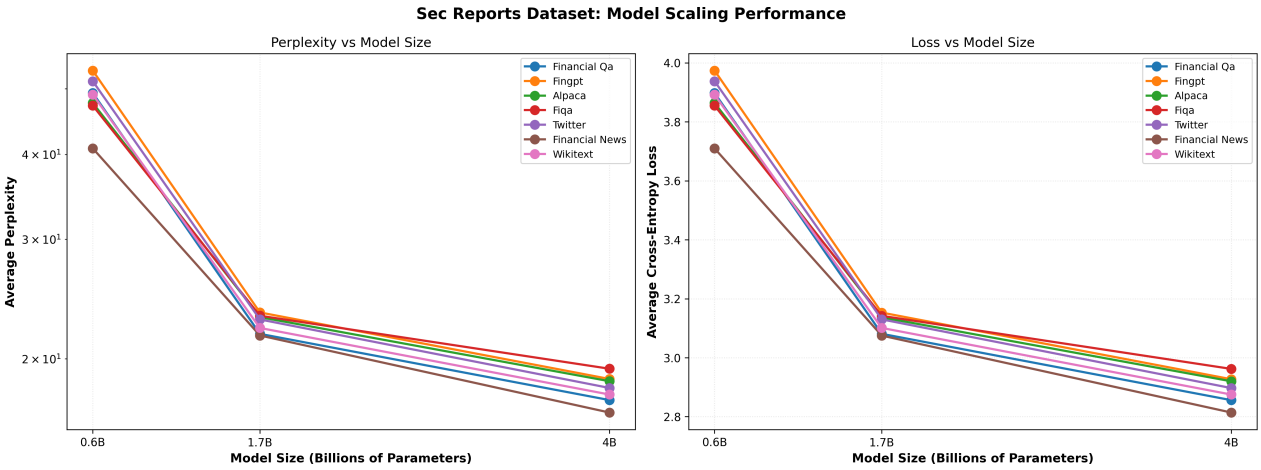

Figure 4.7 – SEC Reports scaling.

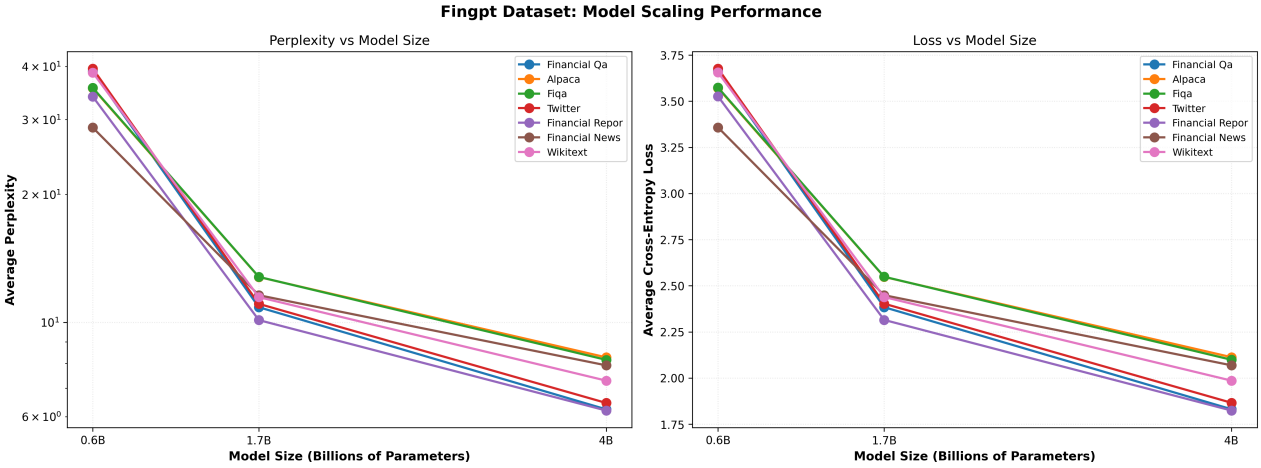

Figure 4.8 – FinGPT instruction mixture scaling.

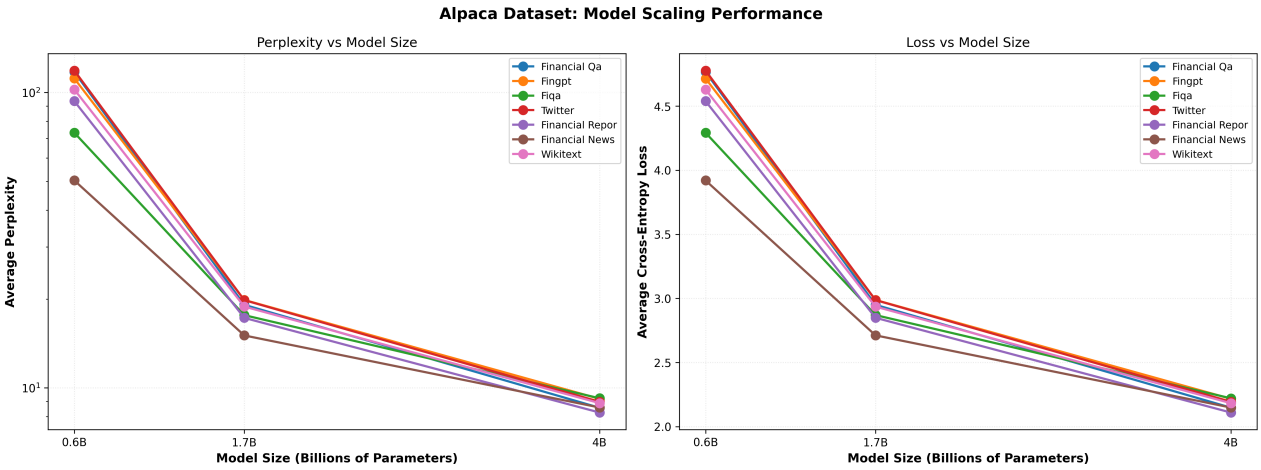

Figure 4.9 – Alpaca instruction mixture scaling.

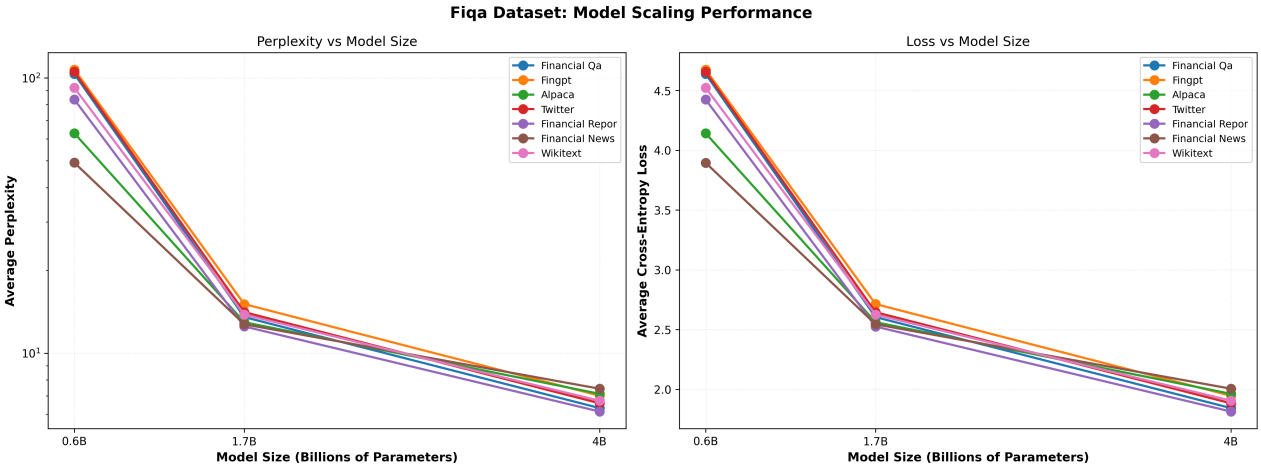

Figure 4.10 – FiQA short-form scaling.

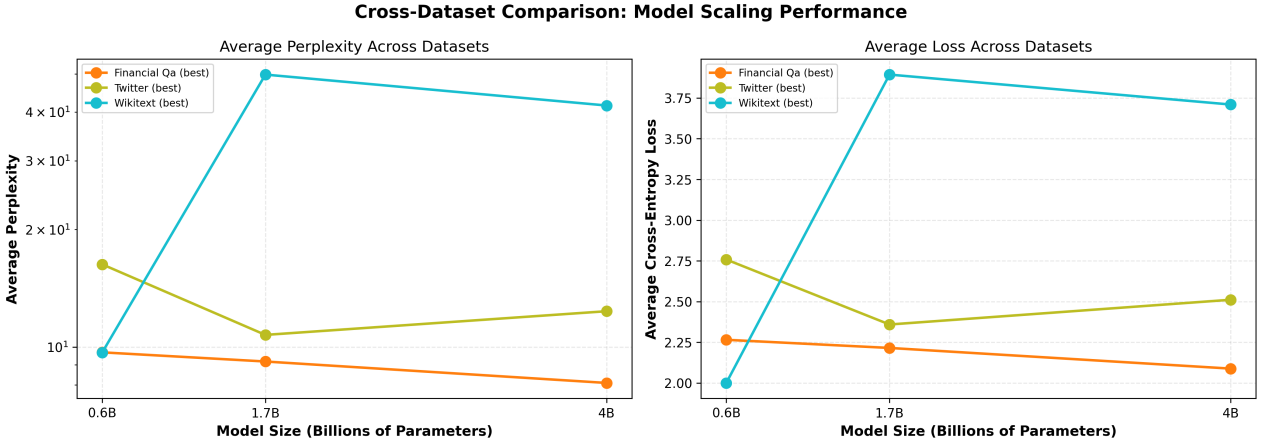

Figure 4.11 – Comparison across training sources.

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.

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert, and Jason Weston (2009). "Curriculum learning". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, pp. 41–48. DOI: 10.1145/1553374.1553380.

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.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT 2019*, pp. 4171–4186. DOI: 10.18653/v1/n19-1423. URL: https://doi.org/10.18653/v1/n19-1423.

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*.

Mitra, Arindam, Luciano Del Corro, Shweti Mahajan, Andres Codas, Clarisse Simoes, Sahaj Agarwal, Xuxi Chen, Anastasia Razdaibiedina, Erik Jones, Kriti Aggarwal, Hamid Palangi, Guoqing Zheng, Corby Rosset, Hamed Khanpour, and Ahmed Awadallah (2023). "Orca 2: Teaching Small Language Models How to Reason". In: *arXiv preprint arXiv:2311.11045*. URL: https://arxiv.org/abs/2311.11045.

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.

Tay, Yi, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, Huaixiu Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler (2022). "UL2: Unifying Language Learning Paradigms". In: *arXiv preprint arXiv:2205.05131*. URL: https://arxiv.org/abs/2205.05131.

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.

Zhang, Susan, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer (2022). "OPT: Open Pre-trained Transformer Language Models". In: *arXiv preprint arXiv:2205.01068*. URL: https://arxiv.org/abs/2205.01068.