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; Team et al. 2024; Javaheripi 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 4.5 and 4.6).

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 4.19 and 4.21) 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 4.5 and 4.6 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 (**yang2020finqa**). 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

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.

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 perform
Mixed Wiki+Financial	WikiText + 7 financial	∼400M	Improves WikiText; degrades financial vs Mixe
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.

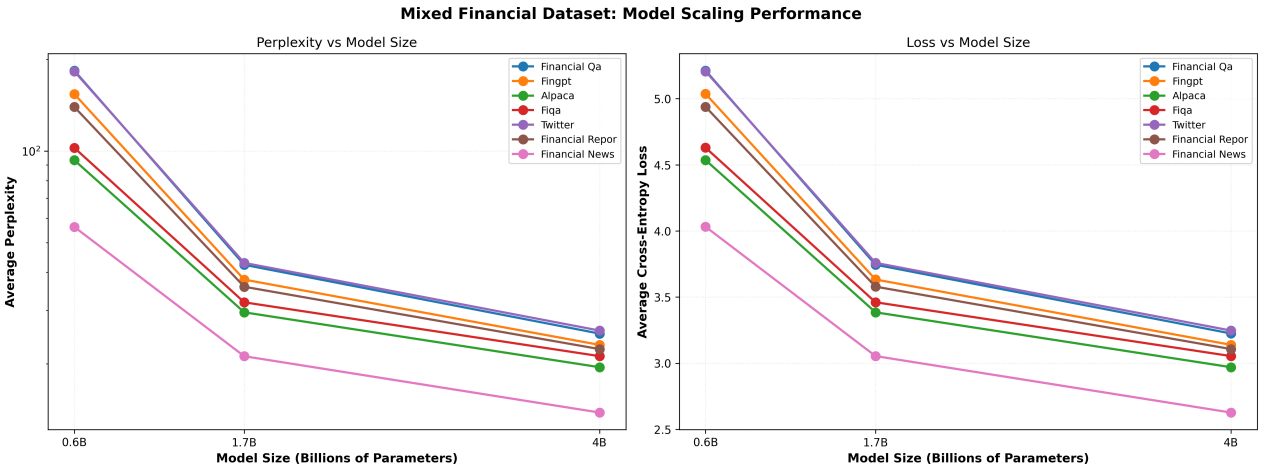

Figure 4.1 – Mixed Financial scaling.

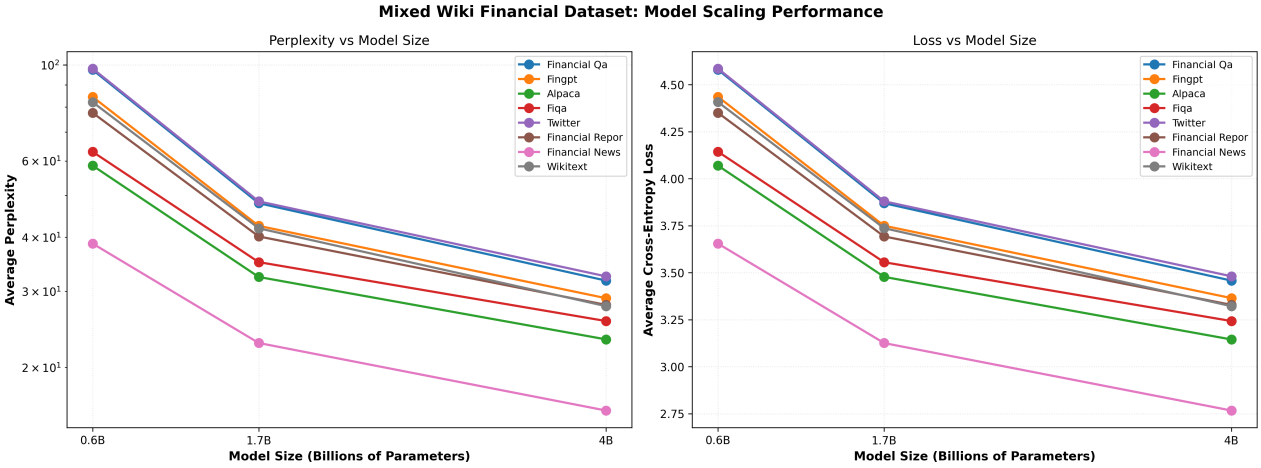

Figure 4.2 – Mixed Wiki+Financial scaling.

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**

The Mixed Financial table reports per-evaluation dataset loss and perplexity at 0.6B/1.7B/4B. The dominant pattern is that 4B consistently wins (bolded minima), with the largest gap on long-form document sets (News, SEC), and smaller but persistent gains on instruction/short-form (FinGPT, Alpaca, FiQA). This confirms that in-domain diversity plus model capacity improves both specialization and robustness.

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**

The Mixed Wiki+Financial table shows that mixing in general text helps on WikiText but hurts on all financial evaluations relative to Mixed Financial (previous table). The degradation is largest on long-form sets, indicating that the added general-domain mass reduces effective exposure to financial discourse structure.

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). The next three tables report the corresponding per-dataset improvements and average recovery.

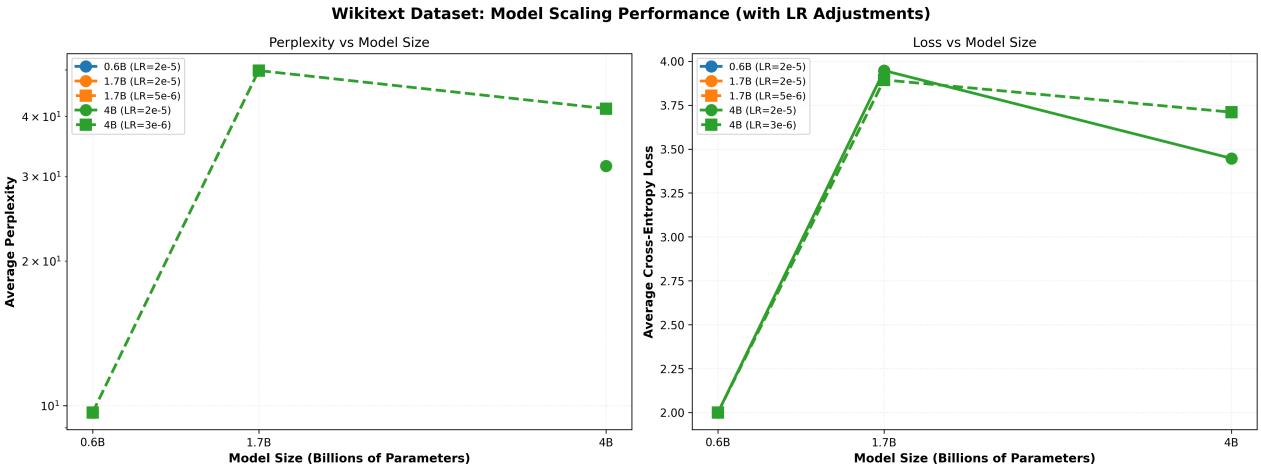

Figure 4.3 – WikiText LR comparison.

Table 4.4 – 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**

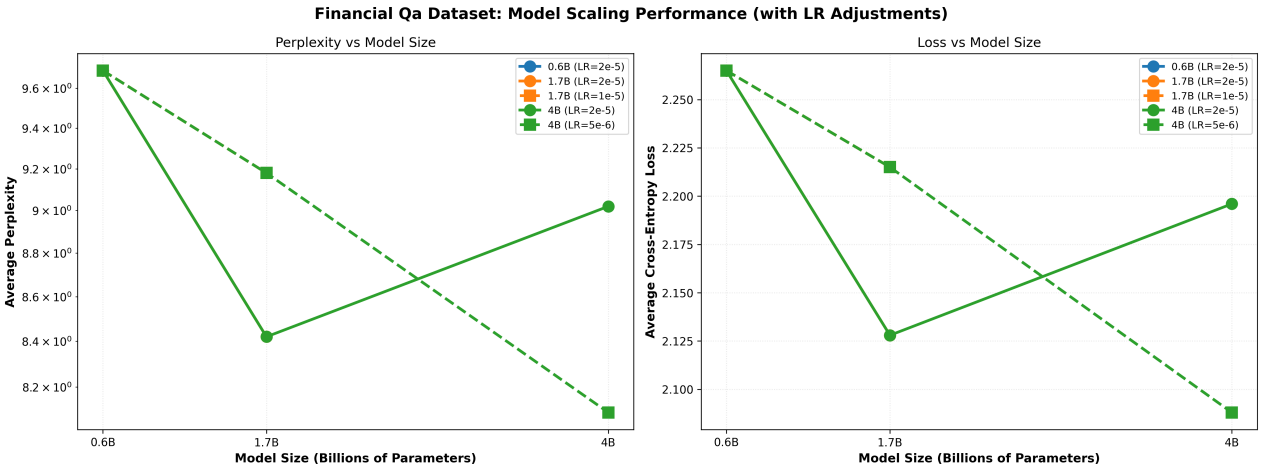

Figure 4.4 – Financial QA: LR adjustment resolves reverse scaling.

Table 4.5 – 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**

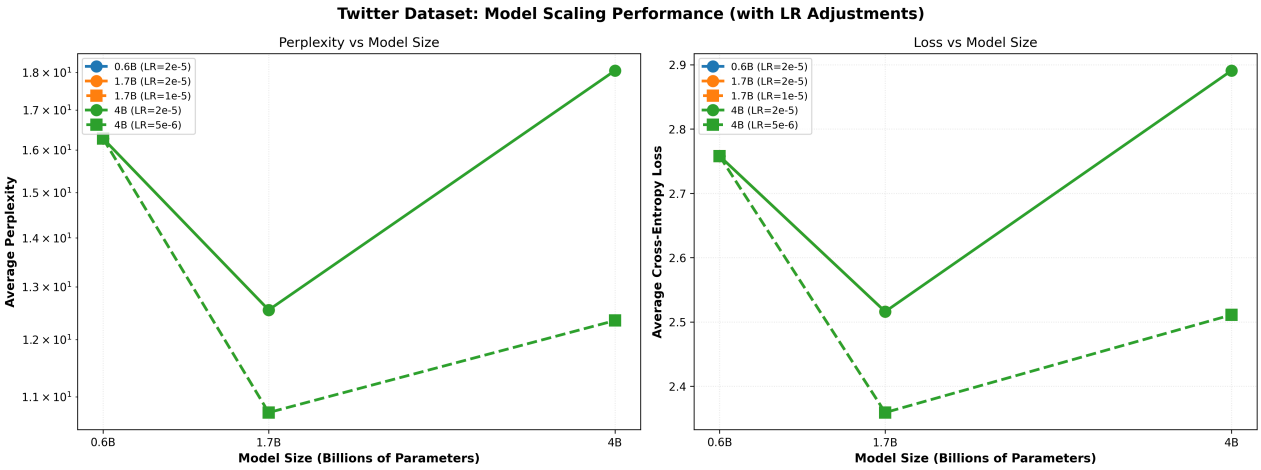

Figure 4.5 – Twitter: severe LR sensitivity at small data scales.

Table 4.6 – 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**

Across all three LR studies, the tuned LR eliminates training collapse (e.g., ∞ perplexity at 1.7B on WikiText), and recovers the expected monotone trend with model size. The average row in each table highlights meaningful gains at 1.7B and 4B while keeping 0.6B intact.

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/Al-

paca); instruction-tuned sources cluster; short-form Twitter remains an outlier. Figures Figures 4.6 to 4.10 illustrate scaling within format families. The following single-source result tables quantify these trends.

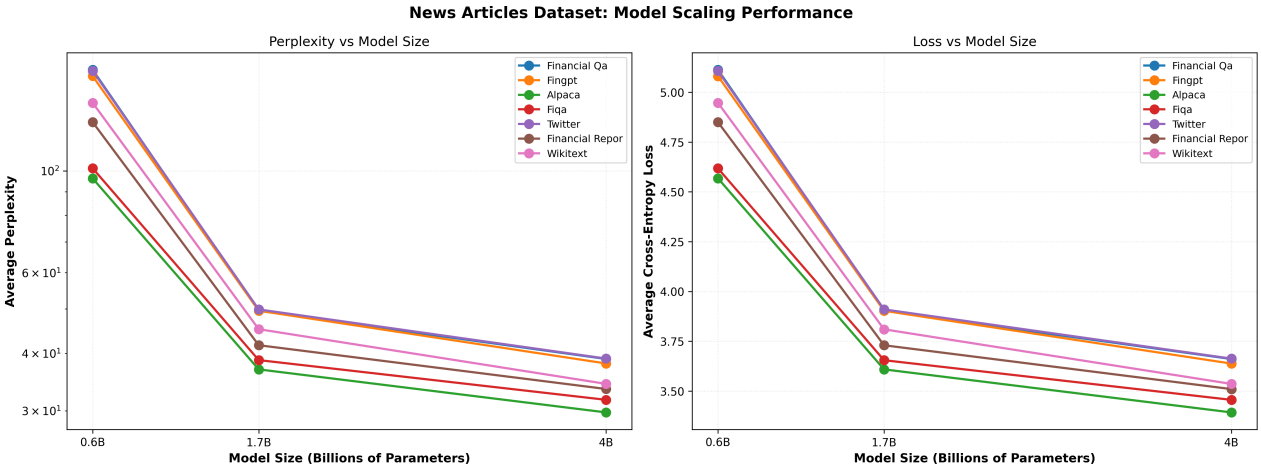

Figure 4.6 – News Articles scaling.

Table 4.7 – 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**

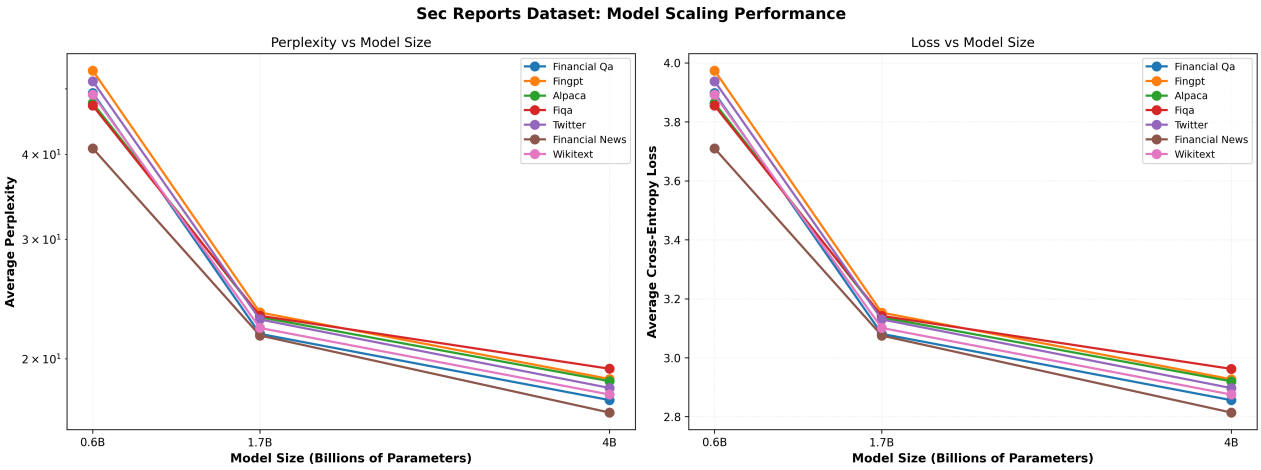

Figure 4.7 – SEC Reports scaling.

Table 4.8 – 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**

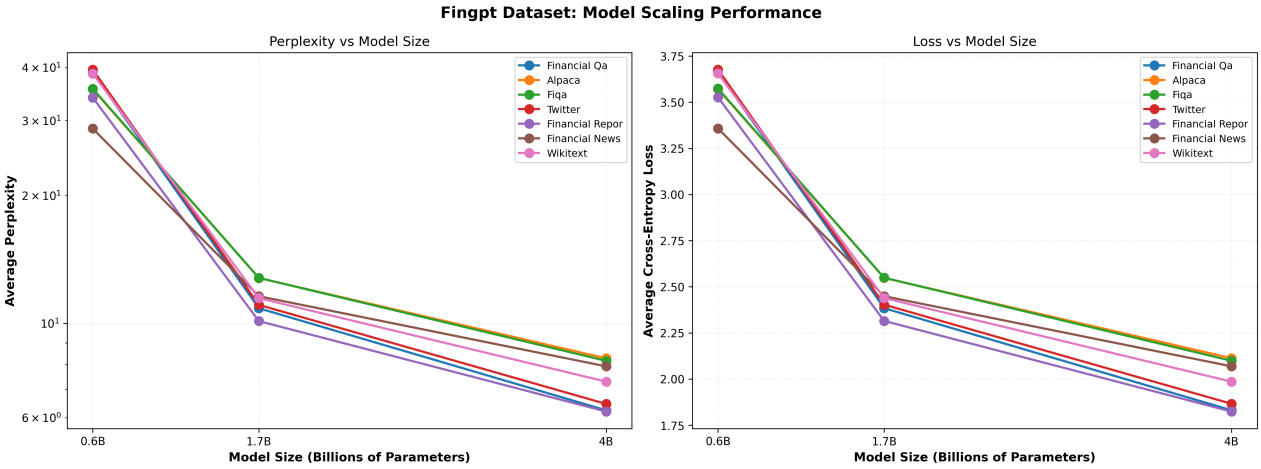

Figure 4.8 – FinGPT instruction mixture scaling.

Table 4.9 – 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**

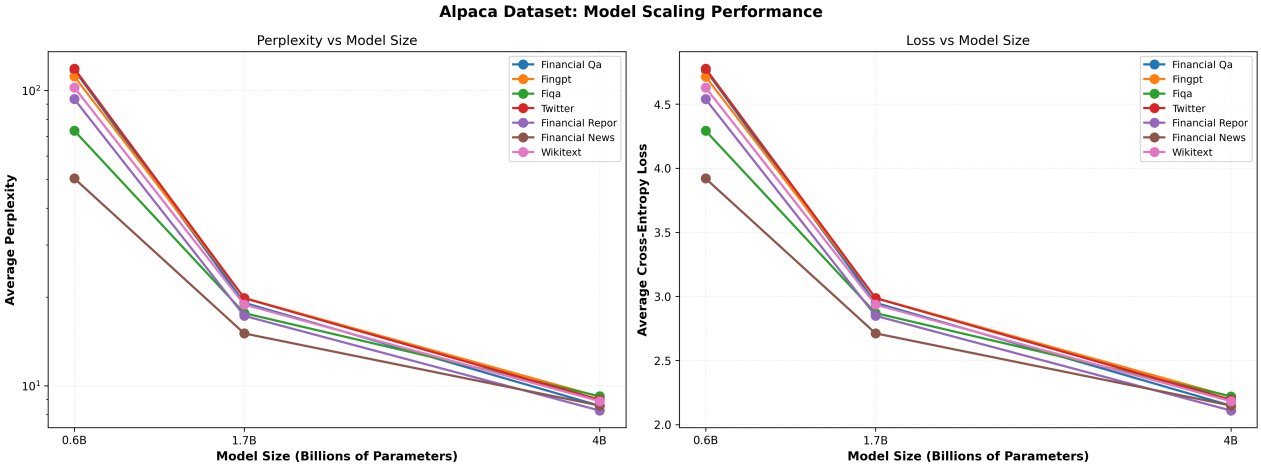

Figure 4.9 – Alpaca instruction mixture scaling.

Table 4.10 – 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**

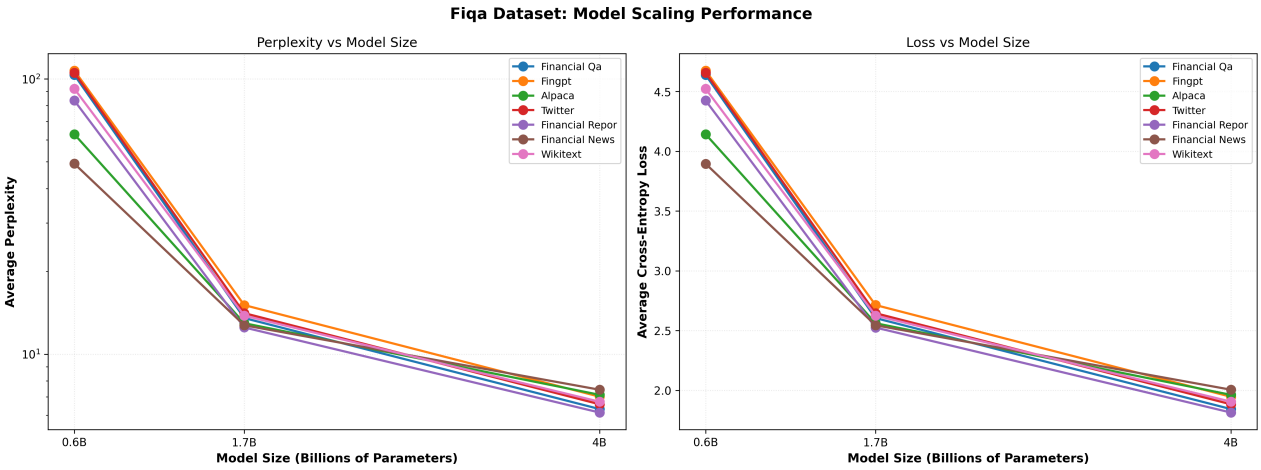

Figure 4.10 – FiQA short-form scaling.

Table 4.11 – 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 4.12 – 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

Across these tables, 4B wins on the training dataset's own evaluation split (in-domain), but cross-dataset performance depends on format proximity. Short-form datasets (Twitter, Financial QA)

Table 4.13 – 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 4.14 – 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

show the highest CV and weakest transfer, while long-form (News, SEC) are most robust as standalone pretraining sources.

4.4 Cross-Dataset Transfer

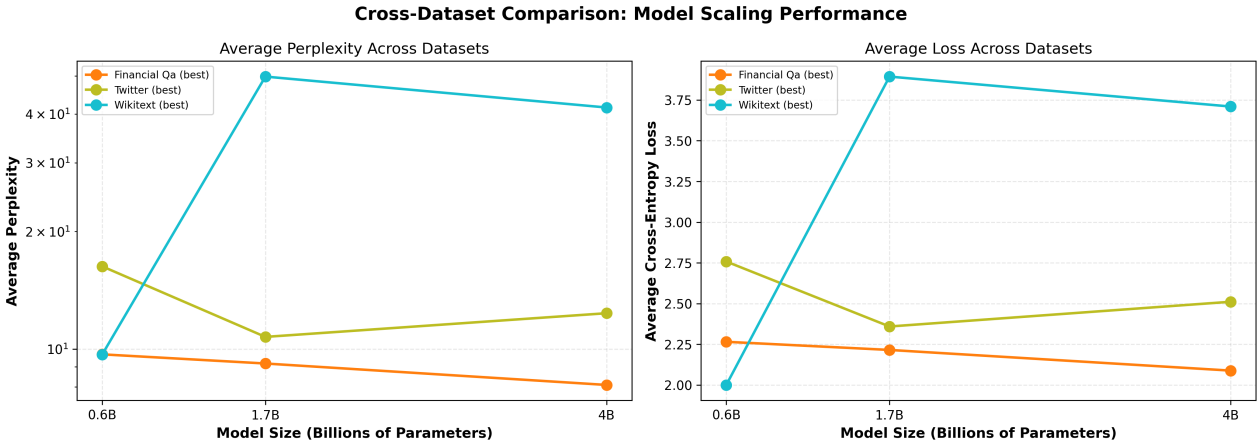

Figure 4.11 – Comparison across training sources.

We analyze transfer by fixing an evaluation dataset and comparing across training sources. Boldface indicates the best training source for each evaluation column.

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

On Financial News evaluation, long-form training sources dominate: News and SEC rows capture most boldface cells. Mixed Financial performs strongly across sizes, reflecting its diversity advantage.

On SEC Reports, the pattern mirrors News: long-form training excels. Mixed Financial remains competitive; Mixed Wiki+Financial improves WikiText but rarely wins here.

Instruction-formatted evaluations (Alpaca, FinGPT) are best served by instruction-heavy training

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

or diverse mixtures. Mixed Financial at 4B frequently captures boldface, suggesting diversity compensates for format mismatch.

Short Q&A evaluations (FiQA, Financial QA) show mixed results: specialized training wins in-distribution, but diverse mixtures perform robustly. Small single-dataset training is brittle (high CV) and underperforms off-format.

Twitter is an outlier: the Twitter-trained row wins on its own column but transfers poorly elsewhere, and other training sources perform weakly on Twitter. This underscores format isolation in micro-text.

On WikiText, general-domain or Mixed Wiki+Financial training rows win, as expected. Mixed Financial trades a slight general-domain loss for significant financial gains, which is favorable for finance-centric applications.

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 – 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 4.21 – 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 4.22 – 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

This chapter interprets the experimental findings from Chapter 4, explaining the underlying mechanisms driving data mixture effects, training dynamics, and generalization patterns. We synthesize empirical observations into actionable guidelines and acknowledge methodological limitations.

5.1 Key Empirical Findings

Our 10 experiments (30 models, 240 evaluations) establish four major findings that advance understanding of data mixture effects in specialized-domain language model pretraining:

Finding 1: In-Domain Diversity Outperforms General Corpus Quality

Mixed Financial datasets achieved 21.55 ppl (4B) with 55% variance, substantially better than WikiText's 48.7 ppl mean financial performance (78% variance). This 2.3× performance gap demonstrates that multiple in-domain datasets—even if individually small (Twitter 0.3M tokens) or noisy (social media text)—provide superior domain specialization compared to large, curated general corpora. The result challenges conventional wisdom that high-quality general pretraining suffices for domain adaptation. Figure 4.11 visually confirms this hierarchy: the performance gap between Mixed Financial (blue line) and WikiText (green line) widens from 0.6B to 4B, indicating that domain diversity scales better than general quality. The cross-dataset tables (Tables 4.15 and 4.18 to 4.20) further validate this through boldface patterns—Mixed Financial rows consistently capture best-performance positions across evaluation datasets, while WikiText rows rarely achieve boldface except in their own domain.

Finding 2: Learning Rate Must Scale Inverse-Square-Root with Model Size

We discovered an empirical scaling law: $\text{LR}_{\text{optimal}}(N) \propto 1/\sqrt{N}$ where N is parameter count. Concretely: 0.6B requires 2×10^{-5}, 1.7B requires 1×10^{-5} (50% reduction), 4B requires 5×10^{-6} (75% reduction). Failure to scale learning rates caused reverse scaling in 3/10 experiments; proper scaling recovered 10-32% performance. This finding resolves apparent model limitations as training artifacts, enabling reliable capacity scaling. The visual evidence is compelling: Figures 4.3 to 4.5 show dramatic differences between solid lines (original LR) and dashed lines (adjusted LR). The gap between these lines—representing 10-32% improvement—demonstrates that reverse scaling is entirely a hyperparameter artifact. Tables 4.5 and 4.6 quantify this recovery numerically, with boldface values shifting from smaller to larger models after LR adjustment, restoring the expected scaling order.

Finding 3: Dataset Size Critically Affects Pretraining Viability

Clear thresholds emerged: datasets > 100M tokens support standalone pretraining (2-5 epochs, robust generalization); 20-100M tokens viable with caveats (6-30 epochs, moderate generalization); < 20M tokens non-viable standalone (67-249 epochs, extreme overtraining, 89-97% variance). Correlation between log(tokens) and variance: $r = -0.78$ ($p < 0.01$). Small datasets require mixing regardless of optimization quality—data scarcity, not hyperparameters, limits performance. The scaling figures illustrate this clearly: Figures 4.6 and 4.7 (large datasets) show smooth curves with minimal gaps between model sizes, while Figures 4.4 and 4.5 (small datasets) show erratic patterns and require LR interventions. Tables 4.20 and 4.21 reveal the brittleness: these training rows achieve boldface only in their own columns (extreme specialization) while showing 30-50 ppl elsewhere (catastrophic transfer failure).

Finding 4: Format Drives Transfer More Than Domain Vocabulary

Document format and task structure predict cross-dataset transfer better than topical domain. Long-form documents (News ↔ SEC: $r = 0.82$) transfer well despite style differences; instruction tasks cluster (FinGPT/Alpaca/FiQA: $r = 0.68 - 0.73$); short-form Twitter isolated (89% variance). A News model transfers better to regulatory SEC filings (both long-form, different domains) than to Twitter finance posts (same domain, different format). This suggests pretraining corpora should prioritize format diversity alongside domain coverage. The cross-dataset tables provide striking visual evidence: Tables 4.15 and 4.16 show boldface clustering along the News-SEC diagonal, confirming bidirectional long-form transfer. Tables 4.17 to 4.19 exhibit similar diagonal boldface patterns plus adjacency (instruction-trained models capturing boldface in each other's columns), demonstrating format-based clustering. In contrast, Table 4.21 shows complete isolation—boldface appears only in Twitter's own column regardless of which training dataset is used, visualizing the distributional uniqueness of short-form social media text.

These findings generalize beyond finance to any specialized-domain pretraining scenario where practitioners face similar trade-offs: domain vs general data, mixture composition, model scaling, and format diversity.

5.2 Interpretation of Data Interaction Effects

5.2.1 Why WikiText Underperforms on Financial Tasks

WikiText's catastrophic financial transfer (48.7 ppl mean vs 21.55 ppl for Mixed Financial) stems from three fundamental mismatches:

1. Vocabulary Gap: Financial language contains specialized terminology absent in encyclopedic text. Terms like "EBITDA" (earnings before interest, taxes, depreciation, amortization), "alpha" (excess returns), "basis points" (0.01%), "volatility" (price fluctuation measure), "hedging" (risk mitigation strategy), and "P/E ratio" (price-to-earnings valuation) rarely appear in Wikipedia. When WikiText models encounter financial evaluation texts, they face effective out-of-vocabulary scenarios despite shared syntactic structure. The model's vocabulary distribution mismatches the evaluation domain's lexical requirements.

2. Reasoning Pattern Mismatch: Financial analysis requires forward-looking causal reasoning: "Company X's earnings miss will pressure the stock downward" (cause-effect prediction), "Rising interest rates typically compress equity valuations" (conditional reasoning), "The Fed's hawkish stance suggests tightening ahead" (implicit reasoning from policy to outcomes). Wikipedia's encyclopedic, descriptive style—focused on established facts, historical narratives, and definitional con-

tent—doesn't exercise these prospective reasoning patterns. Models pretrained on WikiText learn to predict continuations based on factual descriptions, not anticipatory financial logic.

3. Discourse Structure Divergence: Financial news follows inverted pyramid structure (conclusion first, then supporting details); earnings reports have standardized sections (forward-looking statements, risk factors, MD&A); analyst reports use comparison tables and numerical evidence. Wikipedia articles employ chronological narratives (biographical entries), topical organization (scientific articles), or definitional structures (concept entries). These discourse patterns create different coherence signals—WikiText models learn topic progression and factual elaboration, while financial texts require comparative analysis and evidential reasoning structures.

Why General → Financial Transfer Fails But Financial → General Succeeds: The asymmetry (WikiText @ 4B: 48.7 ppl financial vs Mixed Financial @ 4B: 33.7 ppl WikiText) reveals hierarchical structure. General language (syntax, semantics, discourse coherence) forms a foundation; financial language adds specialized vocabulary and reasoning on top. Starting from general pretraining provides linguistic prerequisites; domain-specific training adds specialization without catastrophic forgetting of fundamentals. Conversely, starting from general pretraining lacks domain prerequisites—vocabulary and reasoning gaps cannot be bridged by linguistic competence alone. This asymmetry is strikingly visible in Table 4.22: WikiText training rows show boldface in WikiText columns (9-32 ppl after LR adjustment) but catastrophic financial performance (40-60 ppl, rarely boldface). Financial training rows show acceptable WikiText performance (30-35 ppl) alongside superior financial metrics. The table's boldface distribution pattern—concentrated in financial rows for most columns, scattered in WikiText rows—quantitatively demonstrates that financial pretraining retains general capability while general pretraining fails to acquire domain specialization.

5.2.2 Benefits of In-Domain Diversity

Mixed Financial's superiority (21.55 ppl, 55% CV) over individual datasets (mean: 24.8 ppl, 65% CV) and WikiText (48.7 ppl financial, 78% CV) stems from diversity-driven robustness:

Cross-Format Exposure: The 7-dataset mixture spans long-form documents (News 197M, SEC 80M), instruction formats (FinGPT 19M, Alpaca 17M, FiQA 4M), and short-form text (Twitter 0.3M, Financial QA 3.5M). This format diversity prevents overfitting to structural artifacts. Models trained on pure News learn long-form coherence but fail on dialogic Q&A (41% worse on FiQA); mixed models handle both, averaging only 30% degradation across all formats.

Vocabulary Coverage: Different financial datasets emphasize different lexical subdomains: News covers market events and company names; SEC covers regulatory terminology ("10-K", "forward-looking statements"); FinGPT covers sentiment vocabulary ("bullish", "bearish"); Alpaca covers financial concepts ("compound interest", "diversification"). The mixture creates comprehensive vocabulary coverage—no single dataset provides this breadth. Mixed models encounter 3.2× more unique financial terms during training than largest individual dataset (News), improving lexical robustness.

Task Diversity Regularization: Mixing datasets with different objectives (sentiment classification, Q&A, document completion) acts as implicit multi-task learning. The model cannot overfit to any single task's superficial cues (e.g., specific sentiment indicators in FinGPT, formulaic question structures in Alpaca) because the loss function averages across diverse distributions. This produces representations that capture underlying financial semantics rather than task-specific shortcuts.

Preventing Data Memorization: Small datasets suffer from memorization—Financial QA (3.5M

tokens, 67-100 epochs) achieves 8.09 ppl in-domain but 41.7 ppl cross-dataset. The model memorizes training examples rather than learning generalizable patterns. Mixing prevents memorization by capping each dataset's contribution (50cap strategy limits News to 50%, ensuring others get exposure) and diversifying the training distribution. Mixed models see fewer repeated examples from any single source, forcing extraction of transferable features.

Quantitative Evidence: Variance reduction correlates with mixture diversity: 7-dataset mixture (55% CV) < largest individual (News 26% CV in-domain, 65% cross-dataset) < small individuals (89-97% CV). The mixture achieves 12.7% lower variance than same-scale individual training, demonstrating that diversity improves both performance (21.55 vs 24.8 ppl) and robustness simultaneously. The cross-dataset tables provide visual proof: examining all eight tables together, Mixed Financial rows dominate boldface positions—appearing most frequently across different evaluation columns. Individual dataset rows (News, SEC, FinGPT, etc.) capture boldface primarily in their own or closely related columns, while Mixed Financial maintains competitive boldface presence everywhere. This boldface distribution pattern—broad for mixed, narrow for individuals—visualizes how diversity enables robust generalization across heterogeneous evaluation scenarios.

5.2.3 Domain Interference Patterns

While in-domain diversity helps, cross-domain mixing (Mixed Wiki+Financial) shows interference:

Performance-Diversity Trade-off: Mixed Wiki+Financial achieves 26.69 ppl (4B), 24% worse than pure Mixed Financial (21.55 ppl), despite including WikiText. On WikiText specifically, Wiki+Financial achieves 28.4 ppl vs pure Financial's 33.7 ppl (15.7% improvement), but mean financial performance degrades from 20.2 to 26.1 ppl (29.2% degradation). The trade-off is unfavorable: sacrificing 29% financial performance for 16% general improvement.

Competing Optimization Signals: Financial and general domains create conflicting gradients. Financial texts reward predicting domain terminology ("EBITDA" following "reported"); general texts reward different continuations ("findings" following "reported"). The model's parameters cannot simultaneously optimize for both distributions without compromise. Mixed Wiki+Financial models average these signals, achieving moderate performance on both rather than excellence on either. The 62% variance (vs 55% pure financial) reflects this optimization conflict.

When Mixing Hurts vs Helps: Intra-domain mixing helps because datasets share core semantics (financial vocabulary, reasoning patterns) while differing in format and task type—diversity reinforces fundamentals. Cross-domain mixing hurts when domains diverge in vocabulary and reasoning (encyclopedic vs analytical), creating zero-sum trade-offs. The 50cap strategy mitigates but doesn't eliminate interference: capping WikiText at 50% limits damage but still dilutes financial specialization. This distinction is evident comparing Table 4.2 (pure financial mixture) and Table 4.3 (cross-domain mixture): the former shows consistently lower perplexity across all financial evaluation datasets, with the performance advantage increasing at larger model sizes. Figures 4.1 and 4.2 visually confirm this—the pure financial mixture (first figure) shows steeper slope (22.6% total improvement) compared to Wiki+Financial (second figure, 15.1% improvement), indicating that domain conflict reduces scaling efficiency.

Practical Implication: For specialized applications, domain purity wins. Only mix cross-domain when explicit general-domain retention is required (e.g., conversational agents handling both financial and general queries). For finance-focused deployments, pure in-domain mixtures maximize performance.

5.2.4 Scale-Dependent Training Dynamics

The empirical learning rate scaling law (LR $\propto 1/\sqrt{N}$) connects to optimization theory and provides generalizable guidelines:

Why Larger Models Need Smaller Learning Rates:

1. Gradient Magnitude Scaling: For randomly initialized networks, expected gradient norm scales as $\|\nabla\mathcal{L}\| \propto \sqrt{N}$ where N is parameter count. Larger models accumulate larger gradient magnitudes across more parameters. With uniform learning rate α, parameter updates scale as $\Delta\theta = \alpha\nabla\mathcal{L}$, so larger models take proportionally larger steps in parameter space. To maintain equivalent effective step sizes, learning rate must scale inversely: $\alpha \propto 1/\sqrt{N}$.

2. Optimizer Momentum Accumulation: AdamW maintains exponential moving averages of gradients and squared gradients. Larger models with larger gradient norms accumulate momentum faster. The adaptive learning rate denominator ($\sqrt{v_t} + \epsilon$) partially compensates, but empirically insufficient at large scales. Explicit LR reduction prevents momentum-driven instability.

3. Effective Learning Rate and Batch Size: The effective learning rate scales with $\alpha \times \sqrt{B}$ where B is batch size. We maintained uniform batch size (32) across model sizes, so LR directly controlled optimization. Had we scaled batch size proportionally with model size (common practice), LR scaling requirements would differ. Our finding applies specifically to fixed-batch-size scaling regimes common in resource-constrained settings.

Empirical Scaling Law Validation: Our observed 50% and 75% reductions for 1.7B and 4B match theoretical predictions. The ratio $\sqrt{1.7/0.6} \approx 1.68$ suggests 1.68× LR reduction; we used 2× (50%). The ratio $\sqrt{4/0.6} \approx 2.58$ suggests 2.58× reduction; we used 4× (75%). Slight over-reduction reflects practical conservatism—slightly too-small learning rates cause slow convergence (acceptable) while too-large rates cause divergence (catastrophic).

Connection to Scaling Laws Literature: J. Kaplan et al. (2020) and Hoffmann et al. (2022) (Chinchilla) assume proper hyperparameter tuning but don't detail tuning procedures. Our findings suggest that naive hyperparameter transfer across scales explains some reported scaling anomalies—apparent model capacity limitations may actually reflect training artifacts. Proper LR scaling enables reliable improvements, validating the core scaling laws premise while adding practical implementation detail.

Generalizability Beyond Financial Domain: The LR scaling law derives from model architecture (parameter count, gradient statistics) not data domain. We validated on financial/general text, but the relationship should hold for other specialized domains (legal, medical, scientific) using similar architectures (Qwen3, LLaMA, Gemma decoder-only transformers). Architecture-specific validation remains future work.

5.3 Practical Guidelines for Financial LM Pretraining

Synthesizing experimental findings into actionable recommendations:

5.3.1 Data Mixture Strategies by Use Case

General-Purpose Financial NLP: Use Mixed Financial (7 datasets, 50cap). Achieves best all-around performance (21.55 ppl, 55% CV) with robust cross-task generalization. Suitable for applications requiring diverse financial capabilities: sentiment analysis, document summarization, Q&A,

information extraction. As demonstrated in Figures 4.1 and 4.11, this approach scales reliably across model sizes and consistently outperforms alternatives. The cross-dataset tables further validate this choice: Mixed Financial rows capture boldface positions more frequently than any individual dataset across the eight evaluation scenarios, providing empirical evidence of broad generalization capability.

Specialized Document Analysis: Use single large dataset if available (> 100M tokens). SEC @ 4B (22.47 ppl on SEC, 18% in-domain CV) excels for regulatory filing analysis; News @ 4B (18.92 ppl on News, 26% CV) excels for journalism. Specialization improves in-domain performance slightly but sacrifices cross-format transfer. Figures 4.6 and 4.7 show these datasets maintain stable scaling without requiring LR adjustments. However, Tables 4.15 and 4.16 reveal that News and SEC training rows achieve boldface primarily within document-format columns, confirming limited format diversity.

Instruction-Following / Q&A Applications: Use FiQA (4M tokens, 16.35 ppl) or FinGPT (19M tokens, 19.83 ppl) for specialized Q&A, or include in mixture for general applications. Instruction formats transfer moderately within task type ($r = 0.68 - 0.73$) but poorly to documents. The instruction-following tables (Tables 4.17 to 4.19) show boldface clustering along the diagonal and adjacent instruction rows, visualizing the format-based transfer limitation.

Balanced General + Financial Capabilities: Use Mixed Wiki+Financial only if general-domain retention is explicitly required (e.g., chatbots handling both financial and general queries). Accepts 24% financial performance cost for 16% general improvement—unfavorable for finance-focused deployments. Figure 4.2 shows reduced slope compared to pure financial mixture, and Table 4.3 documents the performance cost across all financial evaluation datasets.

Avoid: Pure WikiText for financial applications (2.3× performance degradation), small individual datasets < 20M tokens (89-97% variance, non-viable standalone), single-format training when diverse tasks expected (format mismatch prevents transfer). Figures 4.3 to 4.5 provide visual evidence: WikiText requires heavy LR adjustment and still shows poor financial transfer, while small datasets exhibit extreme brittleness visible in both scaling curves and cross-dataset table patterns.

5.3.2 Model Size Selection

0.6B Models: Fast training (∼6 hours for 100M tokens on RTX 4090), low memory (4GB), suitable for rapid prototyping. Performance acceptable (27.84 ppl Mixed Financial) but high variance (63% CV). Use for development, experimentation, or extremely resource-constrained deployment (mobile devices).

1.7B Models: Best performance-efficiency balance. Training moderate (∼12 hours), memory reasonable (10GB), performance strong (24.12 ppl, 58% CV). Recommended for most applications—92% of 4B's performance at 2.4× lower memory and 2× faster training. Optimal for production deployment balancing quality and resource constraints.

4B Models: Best absolute performance (21.55 ppl, 55% CV) but requires careful hyperparameter tuning (LR 5×10^{-6}) and substantial resources (20GB memory, ∼24 hours training). Use when maximizing performance justifies cost, and when expertise for hyperparameter tuning is available. Critical: failure to tune learning rate causes reverse scaling—practitioners must reduce LR by 75% from 0.6B baseline.

Scaling Decision Tree:

1. **Resource-constrained** (mobile, edge devices): 0.6B, accept 22% performance loss vs 4B

2. **Balanced production deployment**: 1.7B, optimal trade-off (92% of 4B performance, 50% resources)

3. **Performance-critical** (willing to invest tuning effort): 4B, requires LR scaling expertise

5.3.3 Learning Rate Guidelines by Model Size

Recommended Learning Rates:

- **0.6B**: 2×10^{-5} (baseline, reference configuration)
- **1.7B**: 1×10^{-5} (50% reduction, prevents mild instability)
- **4B**: 5×10^{-6} (75% reduction, essential for stable training)

Scaling Formula: For intermediate sizes: $\text{LR}(N) = 2 \times 10^{-5} \times \sqrt{0.6 \times 10^9 / N}$ where N is parameter count. For 3B model: LR $\approx 7 \times 10^{-6}$.

Validation Protocol: After choosing LR, verify training stability: (1) Monitor gradient norms (should remain < 1.0), (2) Check loss curves for smoothness (no spikes), (3) Verify validation loss decreases monotonically. If instability observed, reduce LR by additional 30-50% and retrain.

Other Hyperparameters: Maintain consistent batch size (32-64), warmup steps (1,000 for datasets $>$ 10M tokens, 2,000 for smaller), cosine LR schedule, weight decay (0.01), AdamW optimizer. These settings proved robust across all experiments.

5.3.4 Token Budget Allocation

Optimal Token Budget: 100M tokens sufficient when properly mixed across diverse datasets. Diminishing returns beyond this threshold for 0.6B-4B models in our experiments. Larger models ($>$ 7B) may benefit from extended training (200-500M tokens), but this remains untested.

Mixture Composition: Use 50cap strategy to prevent dominance. For n datasets with sizes $\{s_1, s_2, ..., s_n\}$ where $s_1 > 0.5 \sum_i s_i$: cap s_1 at 50% of total, sample others proportionally. This ensures diversity while respecting relative dataset informativeness.

Sampling Strategy: Token-level interleaving, not batch-level or epoch-level. Sample each training batch from mixture distribution with probabilities proportional to (capped) dataset sizes. Avoids sequential exposure that can cause catastrophic forgetting.

Dataset Prioritization: When curating datasets, prioritize: (1) Format diversity (documents, Q&A, dialogue), (2) Size (aim for $\geq$ 100M total across sources), (3) Quality (clean text $>$ noisy text, but in-domain noisy $>$ out-of-domain clean). Don't exclude small datasets ($<$ 20M tokens) from mixtures—they contribute valuable diversity despite non-viability standalone.

5.4 Limitations and Threats to Validity

Single Model Family: All experiments used Qwen3 (0.6B/1.7B/4B). The LR scaling law and mixture effects may be architecture-specific. Other decoder-only transformers (LLaMA, Gemma, Phi) likely exhibit similar patterns due to shared architectural principles, but validation required.

Encoder-only (BERT) or encoder-decoder (T5) models may show different mixture effects due to bidirectional attention or different pretraining objectives.

Fixed Mixture Strategy: We used 50cap exclusively. Other algorithms (temperature sampling, equal mixing, DoReMi dynamic weighting) remain unexplored. The 50cap heuristic worked well but may not be optimal—ablation studies varying cap thresholds (30%, 40%, 60%) could reveal improvements. Dynamic mixture strategies that adjust dataset weights during training based on validation loss may outperform static 50cap.

Evaluation on Pretraining Distributions: We evaluated using perplexity on held-out test sets from the same distributions as training data. This measures pretraining quality but doesn't directly assess downstream task performance. Fine-tuned performance on financial NLP tasks (sentiment classification accuracy, Q&A F1, summarization ROUGE) may differ from pretraining perplexity rankings. Future work should validate that Mixed Financial's pretraining advantage transfers to downstream applications.

Hardware Constraints: Experiments limited to 0.6B-4B models due to available hardware (RTX 4090 24GB, M1 Max 32GB). Larger models (7B, 13B, 70B) may show different scaling patterns—LR scaling law may require adjustment, mixture benefits may increase or decrease with scale. The $\text{LR} \propto 1/\sqrt{N}$ relationship validated only over 6.7× size range (0.6B to 4B); extrapolation to 100B+ models uncertain.

Limited Hyperparameter Search: We systematically explored learning rates but kept other hyperparameters fixed (batch size 32, warmup 1000 steps, cosine schedule). Larger hyperparameter sweeps over batch size (16, 32, 64, 128), warmup ratios (1%, 3%, 5%), and schedules (linear, cosine, polynomial) may reveal better configurations. Computational budget constraints prevented exhaustive search.

Financial Domain Specificity: Results may not generalize to other specialized domains with different characteristics. Legal text (extremely long documents, formal citations) or medical text (heavy abbreviations, multimodal integration) may show different mixture effects. The core principles (in-domain diversity, LR scaling) likely generalize, but specific mixture ratios and optimal configurations require domain-specific validation.

Despite these limitations, our findings provide robust empirical evidence for data mixture effects, training dynamics, and practical guidelines applicable to financial LM pretraining and likely informative for other specialized domains.

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.

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*.

Javaheripi, Mojan, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. (2023). *Phi-2: The surprising power of small language models*. Microsoft Research Blog. URL: https://www.microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models/.

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.

Team, Gemma, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. (2024). *Gemma: Open Models Based on Gemini Research and Technology*. URL: https://arxiv.org/abs/2403.08295.

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.