University of Zurich[UZH]

Understanding Data Mixture Effects in Financial Language Model Pretraining

A Study of Domain-Specific and High-Quality General Corpora

Master's Thesis

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

Student

Guanlan Liu

19-768-837

guanlan.liu@uzh.ch

Supervisor

Prof. Dr. Markus Leippold

Professor of Financial Engineering

Department of Finance

University of Zurich

Min Yang

min.yang2@uzh.ch

Date of Submission: October 5, 2025

Abstract

We present a compute-normalized study of pretraining data composition for financial language models. We adapt the decoder-only Qwen3 Base architecture to a fixed 100M-token budget with heterogeneous financial texts and general texts with a unified eight-dataset evaluation. We further develop systematic comparisons of individual datasets versus mixtures to evaluate optimal pretraining strategies. In particular, we find that **medium individual datasets (3.6–8.5M tokens) consistently outperform mixtures on both performance and consistency**. FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread) achieve 2.5–3.2× better perplexity AND 1.5–4.8× better cross-dataset consistency than our seven-source financial mixture (21.55 ppl, 55% spread). This finding challenges conventional wisdom that data diversity improves robustness. This occurs through a three-way interaction: medium datasets achieve optimal epoch counts (12–28 epochs) with format consistency, while large datasets undertrain (<1 epoch) and large mixtures add format conflicts that small models (0.6B–4B) cannot reconcile. Small datasets (<1M tokens) overtrain (143–352 epochs), leading to memorization. WikiText shows competitive performance at small scales (0.6B: 9.68 ppl) but reverse scaling at larger sizes due to training instability.

Our contributions in this work are threefold:

a. We systematically compare individual datasets versus mixtures via token-matched training and unified eight-dataset evaluation, revealing that medium individual datasets (3.6–8.5M tokens) consistently outperform mixtures on both performance and consistency metrics.

b. To understand why mixtures fail, we analyze format inconsistency, vocabulary dilution, and multi-task interference effects. We find that focused optimization on single datasets beats diverse mixing—format consistency and concentrated vocabulary exposure outweigh anticipated diversity benefits.

c. We establish that data quality and focus matter more than scale: medium datasets (FiQA 3.6M, FinGPT 4.1M, Alpaca 8.5M, SEC 8.1M) substantially outperform large datasets (News 194M, WikiText 124M). This non-monotonic size-performance relationship reflects optimal epoch counts (12–28) combined with format consistency, outperforming both undertrained large datasets (<1 epoch) and overtrained small datasets (143–352 epochs).

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Large language models (LLMs) have rapidly changed how we do natural language processing (Vaswani et al. 2017; Radford et al. 2019; Brown et al. 2020; Touvron et al. 2023). Yet using them in finance still brings practical hurdles. Financial institutions and individuals handle highly sensitive data—transactions, portfolios, trading strategies—that cannot be sent to external APIs for privacy and competitive reasons (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). We therefore need lightweight, locally runnable financial language models that maintain reasonable performance while protecting data.

In practice, domain adaptation tends to follow two paths: train very large models from scratch or fine-tune general models on domain data. Most teams cannot afford the first; the second often misses domain nuances (Gururangan et al. 2020). And there is common practice showing that high quality general corpora (e.g., Wikipedia, The Pile) always help specialized applications. However, researchers also show that exceptions do appear (Gao et al. 2021; Raffel et al. 2020; Longpre et al. 2023).

This thesis studies 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 realistic for laptops and some mobile devices while keeping 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 present evidence about data mixture strategies for specialized domains.

1.2 Research Questions

This thesis investigates the following core research questions.

RQ1: Individual datasets versus mixtures Do data mixtures improve performance and consistency compared to individual dataset training? Contrary to conventional wisdom, our results show that **medium individual datasets (3.6–8.5M tokens) strictly dominate mixtures on both metrics**. FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread) achieve 2.5–3.2× better perplexity AND 1.5–4.8× better cross-dataset consistency

than Mixed Financial (21.55 ppl, 55% spread). This finding (Figure 4.4 and Tables 4.2 and 4.7 to 4.9) challenges the mixture hypothesis—data diversity degrades both performance and robustness at fixed token budgets (100M).

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 size and learning rate sensitivity? In our setup, we trained all main runs with LR=2e-5; for a few abnormal cases, we reduced LR pragmatically (e.g., to 1×10^{-5} or 5×10^{-6}) and saw improved stability.

RQ3: Dataset size and quality effects What is the relationship between dataset size and performance? Surprisingly, we find a **non-monotonic relationship**: medium datasets (3.6–8.5M tokens) achieve the best results, outperforming both small (<1M) and large (>100M) datasets. Medium datasets FiQA (3.6M, 6.80 ppl), FinGPT (4.1M, 7.03 ppl), Alpaca (8.5M, 8.73 ppl), and SEC (8.1M, 17.80 ppl) substantially beat the large dataset News (194M, 32.82 ppl). This suggests data quality, focus, and format consistency matter more than scale. Medium datasets achieve optimal training (12–28 epochs per dataset) with format consistency, enabling focused learning. Large datasets undertrain (<1 epoch), while large mixtures combine undertraining with format inconsistency that small models (0.6B–4B) struggle to resolve. Small datasets overtrain (143–352 epochs), causing memorization. Only very small datasets (<1M tokens) exhibit severe overtraining (Figures 4.10 and 4.11), while medium datasets achieve optimal performance without mixing.

RQ4: Domain transfer patterns How well do models pretrained on financial data transfer across task types (sentiment, question answering, document understanding), and how much does document format matter? Cross dataset comparison tables (Tables 4.12 to 4.17) suggest that format consistency (long form, instruction, short form) drives transfer more than domain vocabulary, with boldface patterns clustering along format based diagonals.

These questions are addressed through a detailed experimental framework with more than 30 trained models and 230 evaluation results across eight held-out test sets, providing systematic evidence on data mixture effects in specialized-domain pretraining.

1.3 Related Work

Financial NLP. The domain covers sentiment analysis, question answering, numerical reasoning, and information extraction from regulatory documents (Araci 2019; Z. Chen et al. 2021). Challenges include specialized vocabulary ("alpha," "EBITDA"), domain reasoning patterns, and privacy constraints that push toward local deployment (S. Wu et al. 2023). Existing models range from FinBERT variants (Araci 2019; Y. Yang et al. 2020) to BloombergGPT (50B, mixed 51% financial and 49% general) (S. Wu et al. 2023). Most focus on single large models, while few study mixture effects across sizes.

Language model pretraining. Modern LMs use causal language modeling (next-token prediction) on transformer architectures (Vaswani et al. 2017; Radford et al. 2019; Brown et al. 2020). Scaling laws (J. Kaplan et al. 2020; Hoffmann et al. 2022) show that model size, data size, and compute follow power-law relationships. Bigger models are more sample-efficient. But hyperparameter sensitivity at intermediate scales (0.6B–4B) is less studied. Chapter 2 covers training dynamics and memory constraints.

Data mixture strategies. Pretraining can be sequential (curriculum) or simultaneous. In our study, we use a simple "50cap" mixture and refer to Chapter 2 for rationale.

Domain adaptation and transfer. Transfer learning assumes general pretraining helps specialized

tasks (Devlin et al. 2019; Pan and Q. Yang 2010). But Gururangan et al. (2020) showed domain-adaptive pretraining (continued training on domain text) improves performance. Key challenges include catastrophic forgetting (Kirkpatrick et al. 2017) and distribution shift (Quiñonero-Candela et al. 2008).

1.4 Contributions

This thesis makes six primary contributions that challenge conventional assumptions about data mixture strategies for language model pretraining.

First, we overturn the mixture hypothesis: **medium individual datasets (3.6–8.5M tokens) consistently outperform mixtures on both performance and consistency**. FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread) achieve 2.5–3.2× better perplexity AND 1.5–4.8× better cross-dataset consistency than Mixed Financial (21.55 ppl, 55% spread). This finding contradicts widespread assumptions that data diversity improves robustness. We explain why mixtures fail through analysis of format inconsistency, vocabulary dilution, and multi-task interference, supported by 11 scaling figures and 18 detailed evaluation tables.

Second, we establish a **non-monotonic relationship between dataset size and performance**: medium datasets (3.6–8.5M) substantially outperform large datasets (>100M). FiQA (3.6M), FinGPT (4.1M), and Alpaca (8.5M) beat News (194M) by 2–5× on average perplexity, suggesting data quality, focus, and format consistency matter more than scale. SEC (8.1M) also performs well as a medium dataset. Only very small datasets (<1M) exhibit severe overtraining and require careful tuning or exclusion.

Third, we document pragmatic learning-rate adjustments. All main runs used LR=2e-5; in three cases with training instabilities (WikiText, Financial QA, Twitter), reducing LR (to 1×10^{-5} or 5×10^{-6}) stabilized optimization. These are empirical fixes, not theoretical scaling laws.

Fourth, we analyze domain transfer patterns, finding format consistency drives transfer more than vocabulary overlap. Long-form documents transfer well (News $\leftrightarrow$ SEC), instruction tasks cluster (FinGPT/Alpaca/FiQA), but cross-format transfer fails despite shared domain.

Fifth, we demonstrate WikiText shows scale-dependent behavior: competitive at 0.6B (9.68 ppl) but reverse scaling at 4B (31.54 ppl) due to training instability, not domain mismatch. Adding WikiText to financial mixtures degrades performance (26.69 ppl vs 21.55 for pure financial).

Lastly, we show lightweight financial models (0.6B-4B) deliver practical performance for edge deployment when trained on focused medium datasets, enabling privacy-preserving financial NLP without external API dependencies.

Chapter 2

Background and Related Work

We review the literature in this chapter. We start with financial NLP, then pretraining basics, prior mixture strategies, and finally domain adaptation and transfer learning.

2.1 Financial NLP

2.1.1 Tasks in Financial NLP

Financial natural language processing covers many tasks: sentiment on news and social media, question answering on regulatory documents, numerical reasoning in reports, and information extraction from SEC filings (Araci 2019; Z. Chen et al. 2021). The domain has specific challenges compared to general NLP: specialized vocabulary (e.g., "alpha", "beta", "EBITDA"), domain reasoning patterns (e.g., causal chains in market analysis), numerical grounding (financial statements), and temporal dynamics (market events, earnings releases) (S. Wu et al. 2023; Araci 2019).

2.1.2 Existing Financial Language Models

Several finance focused language models appeared in recent years. **BloombergGPT** (S. Wu et al. 2023), a 50 billion parameter model, was pretrained on a mixture of 51% financial and 49% general data, showing strong performance on financial benchmarks while keeping general capabilities. **FinBERT** variants (Araci 2019; Y. Yang et al. 2020) adapted BERT to financial text via continued pretraining, improving sentiment analysis on financial news. More recently, **FinGPT** (H. Yang et al. 2023) explored open source instruction tuning for financial tasks. Together, these work show both scale first and adaptation first approaches.

2.1.3 Domain-Specific Challenges

Financial NLP faces three practical challenges. First, privacy concerns: financial institutions cannot upload sensitive data (portfolios, trading strategies, client information) to external APIs, so locally deployable models are needed (S. Wu et al. 2023). Second, data scarcity: compared to general web text, curated financial corpora are smaller, so data-efficient training is important. Third, rapid vocabulary change: financial language shifts with market trends (e.g., "DeFi", "ESG"), so models

must adapt to new terms. These constraints motivate our focus on 0.6B to 4B models, as we believe larger models should have better generalization and adaptation capabilities.

2.2 Language Model Pretraining

2.2.1 Pretraining Objectives and Architecture

Modern language models mostly use **causal language modeling**: predict the next token from the context (Radford et al. 2019; Brown et al. 2020). We follow this default. Architecturally, we use the usual decoder-only transformer (GPT, LLaMA, Qwen): self-attention for long context and feed-forward blocks for the non-linear part (Vaswani et al. 2017; Touvron et al. 2023).

2.2.2 Scaling Laws and Model Size Effects

The work of J. Kaplan et al. (2020) linked model size, dataset size, compute, and final performance by power laws. The core point, that larger models can be more sample efficient, pushed the field toward billion parameter scales. Later work added nuance: Hoffmann et al. (2022) showed undertraining is common (Chinchilla); Tay et al. (2022) emphasized objectives and data quality.

2.2.3 Computational and Memory Considerations

Training large language models requires substantial compute. A 1 billion parameter model with 32 bit precision uses roughly 4GB of memory for parameters alone, with optimizer states (e.g., Adam's momentum terms) doubling or tripling this requirement (Rajbhandari et al. 2020; Kingma and Ba 2014).

For models in the 0.6B to 4B range targeted here, memory efficient techniques like mixed precision (bfloat16), gradient accumulation, activation checkpointing, and parameter efficient fine tuning such as LoRA allow training on enterprise class GPUs (e.g., NVIDIA RTX A6000 48GB, A100 40GB, H100 80GB) (Narayanan et al. 2021; Hu et al. 2021). In practice, these techniques are very important for us as we rely on renting GPUs from lambda[1] and we are striving to save compute.

2.3 Data Mixture Strategies

2.3.1 Curriculum Learning and Sequential Mixing

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

[1] `https://lambda.ai/`

2.3.2 Simultaneous Mixture Approaches

An alternative to sequential mixing is **simultaneous mixture**: sample from multiple datasets throughout training. Raffel et al. (2020) (T5) used a multi task mixture with task specific prefixes and found diverse pretraining helped downstream. Xie et al. (2023) introduced DoReMi, which adjusts domain weights during training by validation perplexity, improving sample efficiency over static mixtures on The Pile.

BloombergGPT (S. Wu et al. 2023) mixed 51% financial with 49% general (The Pile, C4) at token level and showed balanced mixtures can keep general skills while adding domain strength. Their focus was a single 50B model. The interaction with model size (0.6B vs 4B) is less clear. Our runs across three sizes reveal a surprising finding: **medium individual datasets (3.6–8.5M tokens) consistently outperform mixtures**, achieving 2.5–3.2× better perplexity and 1.5–4.8× better consistency. FiQA (6.80 ppl), FinGPT (7.03 ppl), and Alpaca (8.73 ppl) substantially outperform Mixed Financial (21.55 ppl), Wiki+Financial (26.69 ppl), and WikiText (41.96 ppl mean financial), per Figure 4.4 and Tables 4.7 to 4.9. Medium datasets achieve this through optimal epoch counts (12–28) and format consistency, while large mixtures combine undertraining (<1 epoch per dataset) with format conflicts that small models (0.6B–4B) cannot reconcile simultaneously. For specialized applications, focused individual datasets win over diverse mixtures.

2.3.3 Domain Proportions and Sampling Strategies

There are three common strategies for deciding domain proportions:

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

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

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

This thesis uses a **50% capping strategy** ("50cap") for financial mixtures (details in Chapter 3) to balance diversity and efficiency. We chose it for simplicity and stability in our setup.

2.4 Domain Adaptation and Transfer Learning

2.4.1 Cross-Domain Transfer in Language Models

Transfer learning, pretraining on broad data then fine-tuning on specialized tasks, has been the common approach since BERT (Devlin et al. 2019; Pan and Q. Yang 2010; Zhuang et al. 2021). The assumption is that general linguistic knowledge transfers to domain applications. However, recent work shows alternatives: Gururangan et al. (2020) found that **domain-adaptive pretraining** (continued pretraining on domain corpora) improves performance across domains, suggesting general pretraining alone is not enough for specialized use.

In finance, Araci (2019) showed improvements from continued pretraining on financial news; Y. Yang et al. (2020) saw further gains with task-adaptive pretraining. More recently, A. H. Huang et al. (2023)

found that domain-specific pretraining outperforms general models on financial information extraction. However, these studies focus on BERT-style masked language models and classification tasks, the effectiveness of domain adaptation for *generative causal language models* in financial pretraining is less studied. Advances in parameter-efficient fine-tuning, such as surgical fine-tuning (Lee et al. 2022), suggest selective adaptation may improve transfer while mitigating catastrophic forgetting.

2.4.2 Catastrophic Forgetting and Stability

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

2.4.3 Distribution Shift and Domain Mismatch

Distribution shift, the gap between training and evaluation data, directly affects generalization (Quiñonero-Candela et al. 2008). In finance, this shows up as vocabulary (financial terms vs general), discourse (analytical reports vs encyclopedic text), and formatting (10-K tables vs narrative news). Aharoni and Goldberg (2020) showed domain mismatch can severely degrade out of distribution performance, which motivates mixtures that cover sub domains.

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

2.4.4 Related Empirical Studies

Several empirical studies inspire our methodology. Xie et al. (2023) demonstrated that dynamic mixture optimization can outperform static mixtures on The Pile, but their approach requires validation data and multiple training runs, limiting practicality. Longpre et al. (2023) investigated the effects of data age, domain coverage, quality, and toxicity on pretraining performance, showing that heterogeneous data sources improve model capabilities. Mitra et al. (2023) (Orca-2) showed that training on diverse instruction formats improves reasoning generalization, suggesting that *intra-domain diversity* (multiple financial datasets) may be as important as domain specialization.

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

Chapter 3

Methodology

This chapter explains how we ran the experiments: we first provide an overview of the design, then model architecture, datasets, training setup with tuning, and lastly evaluation protocol.

3.1 Experimental Design Overview

We evaluate 10 pretraining configurations: 2 mixtures (Financial; Wiki+Financial) and 8 single-dataset baselines. Each configuration is trained at three model sizes (0.6B/1.7B/4B) with a fixed 100M-token budget and evaluated on eight held-out test sets. We also run 6 follow-up runs with adjusted learning rates to address training stability at larger scales. We kept other factors fixed where possible. Table 3.1 summarizes the settings used throughout.

Table 3.1 – Summary of experimental settings used across all pretraining runs.

Aspect	**Setting**
Pretraining configurations	10 total: 2 mixtures (Financial; Wiki+Financial) + 8 single-dataset runs
Model sizes	Qwen3-0.6B, Qwen3-1.7B, Qwen3-4B
Token budget	100M tokens per run (normalized across datasets and model sizes)
Sequence length	1,024 tokens
Optimizer	AdamW (β_1=0.9, β_2=0.999, ϵ=10^{-8}), weight decay 0.01
LR schedule	Cosine decay, 1,000 warmup steps, minimum LR 10^{-6}
Learning rate	2×10^{-5} for all main runs; ad-hoc smaller LRs used in a few follow-ups when anomalies were observed
Batching	Effective batch size 8; gradient accumulation used only when memory was insufficient
Precision	bfloat16 mixed precision; dropout 0.0
Hardware	NVIDIA RTX A6000 (48GB), A100 (40GB), H100 (80GB); GPUs rented from Lambda Labs
Mixture policy	50cap-proportional sampling (sampling cap; does not change corpus sizes) to limit dominance of large sources
Evaluation	8 held-out test sets (7 financial + WikiText); metrics: Cross-Entropy, Perplexity, Relative Spread%

This design supports our research questions on mixture composition, model scale, dataset size, and domain transfer. We detailed the results in Chapter 4.

3.2 Model Architecture

We use the Qwen3 model family (A. Yang et al. 2024; Bai et al. 2023), a series of open-source transformer-based decoder-only language models pretrained on diverse multilingual corpora. Qwen3 employs grouped query attention (GQA) for memory efficiency and supports both standard and flash attention. We select three sizes from the Qwen3 Base series (pretrained checkpoints without post-training alignment), detailed in Table 3.2. In our experiments, these different model sizes allow clean comparisons without changing tokenizers or context limits.

Table 3.2 – Qwen3 model specifications across three scales. All models use the same tokenizer (151,643 tokens) and support 32K context length. Training memory shown for bfloat16 precision.

Model	Parameters	Layers	Hidden	Heads	GQA	Memory
Qwen3-0.6B	600M	16	1024	16	4	∼4GB
Qwen3-1.7B	1.7B	24	2048	16	4	∼10GB
Qwen3-4B	4.0B	40	2560	20	4	∼20GB

We chose Qwen3 for four reasons: (1) architectural consistency across scales enables clean size comparisons, (2) stable baseline performance on general and domain-specific benchmarks, (3) efficient inference suitable for edge deployment (all models fit on consumer hardware, (4) SOTA performance of open-weight language models.

3.3 Datasets

3.3.1 Financial Datasets

We use 7 financial datasets spanning diverse tasks, document types, and data scales (total: 222.69M tokens), summarized in Table 3.3. These datasets vary in size (0.28M to 197.38M tokens), genre (news, reports, Q&A, social media), and formality (regulatory filings vs tweets). This diversity lets us examine intra domain effects without changing models.

Our financial datasets cover diverse genres and formats. SEC reports (8.1M tokens, 200K filings) are 10-K annual filings with formal regulatory language. FiQA (3.6M tokens, 14.5K examples) captures Stack Exchange investment discussions with user-generated Q&A. FinGPT headlines (4.1M tokens, 76.8K examples) provide sentiment labels in conversational format. The Twitter dataset (0.28M tokens, 9.5K tweets) includes Bearish/Bullish/Neutral labels with informal language. Financial QA (0.7M tokens, 7K pairs) draws from recent 10-K filings requiring tabular reasoning. Finance Alpaca (8.5M tokens, 68.9K pairs) is synthetic instruction data—didactic Q&A without time-stamped grounding. WikiText (124M tokens, 1.8M articles) provides the general-domain baseline.

This diversity in genres (journalism, regulatory, instruction, forum, social media, document Q&A) and formality levels lets us test how models handle different financial communication styles. WikiText provides a general-domain contrast.

Table 3.3 – Financial dataset characteristics. Total: 222.69M tokens across 7 datasets with diverse genres and scales. Dataset identifiers listed in footnotes.

Dataset	Examples	Tokens	Genre	Description
Financial News Articles[1]	306.2K	194.5M	Journalism	Long-form articles on markets, earnings, policy
SEC Financial Reports[2]	200K	8.1M	Regulatory	10-K annual filings with formal disclosures, legal language
FinGPT Sentiment[3]	76.8K	4.1M	Instruction	Headlines + sentiment labels in conversational format
Finance Alpaca[4]	68.9K	8.5M	Q&A	Instruction-response pairs on financial concepts
FiQA[5]	14.5K	3.6M	Forum	User-generated Q&A from Stack Exchange Investment topic
Financial QA 10K[6]	7.0K	0.7M	Document	Questions on recent 10-K filings requiring tabular reasoning
Twitter Financial Sentiment[7]	9.5K	0.28M	Social Media	Labeled tweets (<280 chars) with informal language

[1]`ashraq/financial-news-articles`, [2]`JanosAudran/financial-reports-sec:small_lite`, [3]`FinGPT/fingpt-sentiment-train`, [4]`gbharti/finance-alpaca`, [5]`LLukas22/fiqa`, [6]`virattt/financial-qa-10K`, [7]`zeroshot/twitter-financial-news-sentiment`

3.3.2 WikiText

We use WikiText-103 (Merity et al. 2017) as a general-domain baseline, summarized in Table 3.4. WikiText serves two purposes: (1) evaluating domain transfer (general ↔ financial), and (2) testing whether high-quality general corpora complement financial pretraining in mixtures.

Table 3.4 – WikiText-103 characteristics. Similar scale to SEC; smaller than News. Dataset identifier in footnote.

Dataset	Examples	Tokens	Genre	Description
WikiText-103[8]	1.8M	124M	Encyclopedia	Verified Wikipedia articles with formal register, broad topical coverage, clean preprocessing

[8]`wikitext:wikitext-103-v1`

3.3.3 Mixture Strategies

We use a 50% capping strategy ("50cap") for dataset mixing to balance diversity with data efficiency. If the largest dataset in a mixture exceeds 50% of total tokens, we cap it at exactly 50%. This prevents single-source dominance. The remaining datasets are sampled proportionally to their original sizes in token counts, preserving relative contributions while ensuring diversity. During training, we sample batches from the mixed distribution at the token level, not by example.

We demonstrate our strategy with an example. For the 7-dataset financial mixture (News 194.47M, SEC 8.12M, FinGPT 4.14M, Alpaca 8.46M, FiQA 3.60M, Financial QA 0.70M, Twitter 0.28M; total

219.77M tokens), News exceeds 50% (88.5%) and is capped at 50% (109.89M tokens); the remaining datasets are sampled proportionally from the other 109.89M-token budget; the final mixture stays at ∼219.77M tokens with News contributing exactly half.

For dataset alignment, the financial datasets vary in formality (regulatory SEC filings vs informal tweets), source type (news articles vs forum discussions), and task format (sentiment labels vs Q&A pairs). WikiText represents general encyclopedic knowledge with a different topical distribution. We accept these distribution differences, as we believe that real applications mix diverse sources with varying formality and topical coverage.

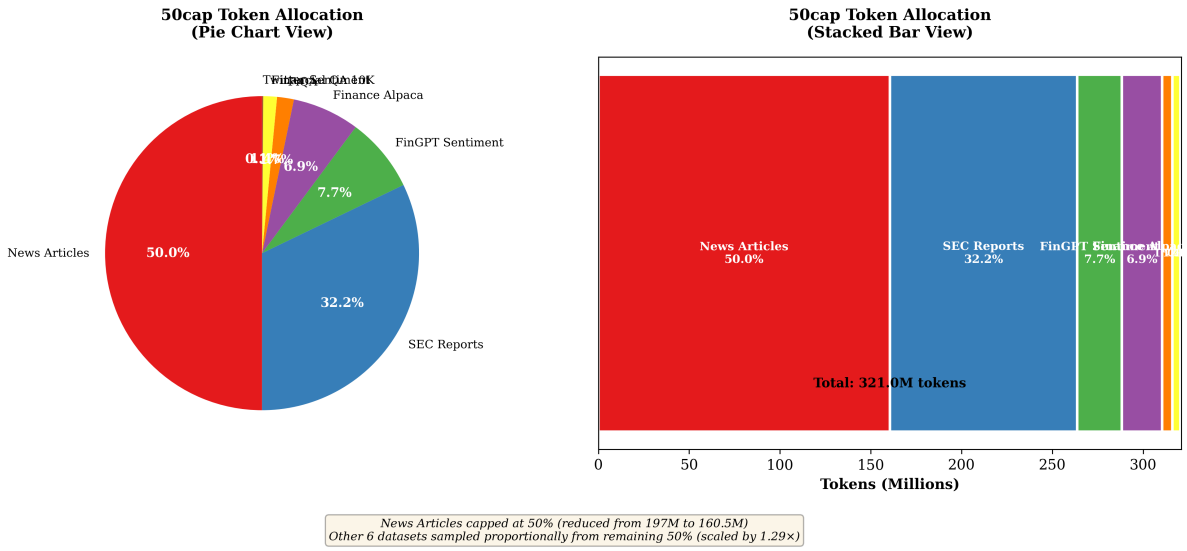

Figure 3.1 – Token allocation in Mixed Financial dataset using 50cap strategy. News Articles is capped to contribute at most 50% in sampling (illustrative normalization; raw corpus remains 194.47M). The remaining six datasets are sampled proportionally from the other 50%, ensuring diversity while preventing dominance. Left panel shows pie chart view; right panel shows stacked bar view with total allocation.

Figure 3.1 visualizes the 50cap sampling policy: News Articles (red) is capped to at most 50% of sampled tokens; the remaining 50% is distributed proportionally among the other six datasets. The pie (left) shows percentage composition; the stacked bar (right) normalizes absolute counts to a 50/50 split ($\approx$109.89 M+ $\approx$109.89 M if scaled to the 219.77M corpus total) for illustration only — 50cap does not modify raw corpus sizes.

For the 8-dataset WikiText+Financial mixture, WikiText (123.58M) and News (194.47M) 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). Our 50cap approach is deterministic, and requires no tuning.

3.4 Training Setup and Hyperparameter Tuning

3.4.1 Initial Configuration

We trained all models with a single hyperparameter template to set a baseline.

We used AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01) with an initial learning rate of 2×10^{-5}, cosine decay, 1,000 warmup steps, and minimum LR 10^{-6}. The effective batch size was 8 across all runs; when memory was tight, we used gradient accumulation to maintain that size. Sequences were 1,024 tokens with bfloat16 mixed precision. Training duration was dataset-dependent: large datasets (>100M tokens) trained for <1 epoch (News: 0.5 epochs, WikiText: 0.8 epochs), medium datasets (3.6–8.5M) for 12–28 epochs, and small datasets (<1M) for 143–352 epochs to reach the fixed 100M token budget.

When we observed abnormalities in a few experiments, we reran those specific cases with smaller LRs as a simple heuristic to stabilize training.

3.4.2 Pragmatic Learning Rate Adjustments

In three configurations we observed abnormal behavior (e.g., larger models underperforming smaller ones). For these few cases, we retried with smaller learning rates (e.g., 1×10^{-5} or 5×10^{-6}) purely as a practical heuristic to stabilize training. We do not propose or rely on a learning-rate scaling theory in this work. LR-comparison tables for the affected settings are reported in Chapter 4.

3.4.3 Other Hyperparameters

Beyond learning rate, we kept other hyperparameters consistent: effective batch size 8 (using gradient accumulation as needed), warmup of 1,000 steps (8% of 12K total steps), and dropout 0.0. Training epochs varied by dataset size to normalize token exposure: small datasets (Twitter, Financial QA) needed 143–352 epochs to reach 100M tokens; medium ones (SEC, FiQA, FinGPT, Alpaca) 12–28 epochs; large ones (News, WikiText) 0.5–0.8 epochs. We fixed maximum sequence length at 1,024 tokens; although financial documents often exceed this, longer sequences increase memory quadratically, so we accepted truncation as a practical trade-off.

3.4.4 Computational Budget

To ensure fair comparison across experiments, we normalized the token budget to 100M tokens per training run, regardless of dataset size or model scale. In total we ran 36 trainings: two mixture settings (Mixed Financial; Mixed Wiki+Financial), eight single-dataset baselines (WikiText, Financial News, SEC, FinGPT, Finance Alpaca, FiQA, Financial QA 10K, Twitter), each at three sizes (0.6B/1.7B/4B) for 30 baselines, plus six follow-ups with reduced learning rates on the three problematic datasets (WikiText, Financial QA, Twitter) to probe sensitivity at larger scales. The total computational cost was $36 \times 100\text{M} = 3.6\text{B}$ tokens.

This token controlled design helps ensure that performance differences reflect model data interactions rather than unequal training compute. Variable epoch counts (2 to 249 across experiments) follow from dataset size while keeping token exposure constant. But it also means small datasets see many passes. We accept this trade-off for fair comparisons across different settings.

3.5 Evaluation Protocol

3.5.1 Multi-Dataset Evaluation

Each trained model is evaluated on eight held-out test sets to measure both in-domain and out-of-domain generalization: seven financial test splits (News, SEC, FinGPT, Alpaca, FiQA, Financial QA, Twitter) plus WikiText test split to evaluate general language capabilities and cross-domain transfer.

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. We first use Cross-entropy loss, which is the average negative log-likelihood per token,

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log P(w_i \mid w_{<i})$$

with lower being better.

We then use **Perplexity** as a interpretable transformation, where PPL = $\exp(\mathcal{L})$. Lower PPL indicates better performance.

We also use **Relative Spread** to measure cross-dataset variability:

$$\text{Relative Spread}\% = 100\,\frac{\max(\text{PPL}) - \min(\text{PPL})}{\text{mean PPL}},$$

computed over evaluation perplexities (one per dataset); lower values indicate more consistent generalization.

All metrics are computed on full test sets (no subsampling) with the same sequence length (1,024 tokens) and batch size used during training. Evaluation uses the final checkpoint from training (no checkpoint selection based on validation performance).

Chapter 4

Results

4.1 Overview of Experimental Results

This chapter shows results from 10 pretraining dataset setups on data mixtures for financial models. We trained 36 models in the main experiments (3 sizes $\times$ 10 configurations, plus models from LR adjustment experiments), and ran 288 evaluations (36 models $\times$ 8 test sets). In our experiments, we included 6 follow-up runs with adjusted learning rates to address training instabilities. Table 4.1 lists the experimental setups. We kept the setup fixed for a fair comparison across different dataset setups.

Experiment	**Datasets**	**Budget**	**Best (Avg PPL)**	**Spread**
Mixture Experiments				
Mixed Financial	7 financial	100M	4B (21.55)	55%
Mixed Wiki+Fin	8 (Wiki+7 fin)	100M	4B (26.69)	62%
Large Individual Datasets				
WikiText	WikiText-103	100M	0.6B (9.68)	53%
News Articles	Lettria News	100M	4B (32.82)	66%
SEC Reports	SEC Filings	100M	4B (17.80)	19%
Medium Individual Datasets				
FinGPT Sentiment	FinGPT	100M	4B (7.03)	37%
Finance Alpaca	Alpaca	100M	4B (8.73)	**11.5%**
FiQA	FiQA Q&A	100M	4B (6.80)	19%
Small Individual Datasets				
Financial QA 10K	10K Q&A	100M	1.7B (8.43)	22%
Twitter Sentiment	Twitter	100M	1.7B (12.55)	51%

Table 4.1 – Overview of 10 pretraining dataset setups. Perplexity is average across all 8 test sets for the best-performing model size. Spread is relative spread (%) measuring cross-dataset consistency (lower is better). Medium datasets (FiQA, FinGPT, Alpaca) dominate on both metrics.

Four critical findings emerge. First, medium-sized individual datasets (FiQA, FinGPT, Alpaca, SEC) are superior on both performance and consistency metrics: they achieve 2.5–3.2$\times$ lower perplexity (6.80–17.80 ppl vs 21.55 ppl for mixture) AND 1.5–4.8$\times$ better cross-dataset consistency (11.5–37% spread vs 55% spread). The mixture approach offers no robustness advantage as individual datasets outperform on all metrics.

Second, WikiText shows strong general-domain performance (9.68 ppl @ 0.6B), competitive with specialized datasets, but exhibits reverse scaling at 1.7B and 4B due to training instability.

Third, the only large individual financial dataset (News, 194M tokens) shows poor average perplexity (32.82 ppl) due to undertraining (0.5 epochs), confirming that optimal epoch count matters more than dataset size.

Fourth, small datasets (Financial QA, Twitter) overtrain heavily (143–357 epochs), which might lead to suboptimal performance.

In summary, medium individual datasets with optimal epoch counts (12–28) achieve both best performance and robustness, while mixtures and extreme sizes (too large or too small) are inferior. These results reflect epoch-format consistency interactions: medium datasets achieve optimal epochs (12–28) with single-format focus, large mixtures combine undertraining (<1 epoch per dataset) with multi-format conflicts, and small models might lack capacity to reconcile diverse formats simultaneously.

4.2 Individual Datasets vs Mixtures

Our core research question concerns optimal data mixture strategies for financial language model pretraining. We compare individual datasets, pure financial mixtures (7 datasets), hybrid Wiki+financial mixtures (8 datasets), and pure general domain (WikiText). Contrary to common assumptions favoring data diversity, **medium-sized individual datasets (FiQA, FinGPT, Alpaca) consistently outperform mixtures on both performance and consistency**: 2.5–3.2× lower perplexity (6.80–8.73 ppl vs 21.55 ppl) AND 1.5–4.8× better cross-dataset spread (11.5–37% vs 55%). The mixture approach provides no robustness advantage. WikiText (0.6B: 9.68 ppl) shows competitive general-domain performance but reverse scaling at larger sizes. We conclude that for financial pretraining at 100M-token budgets, individual medium-sized datasets are preferred and mixtures offer no benefits.

4.2.1 Mixed Financial Datasets: Inferior on All Metrics

The 7-dataset financial mixture (News, SEC, FinGPT, Alpaca, FiQA, Financial QA, Twitter; 220M tokens; 50cap sampling policy) was designed to provide robust cross-dataset generalization. However, empirical results show it is **inferior to medium individual datasets on both performance and consistency**: 21.55 ppl with 55% spread versus FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread). The mixture fails to provide the anticipated robustness benefits. The only potential justification for mixtures is task coverage (ability to handle diverse, unknown future tasks), not performance or consistency optimization.

Performance scales cleanly across model sizes: 0.6B reached 130.30 ppl mean; 1.7B, 34.49; 4B, 21.55 (Table 4.2). From 0.6B to 1.7B that's a 73.5% drop; from 1.7B to 4B, another 37.5%. Both perplexity (left panel, log scale) and loss (right panel) decrease smoothly and monotonically (Figure 4.1), with no irregularities or reversals.

Performance across evaluation sets shows 55% relative spread for 4B, indicating reasonable generalization. (We use Relative Spread% $= 100 \times (\max - \min)/\text{mean}$, computed over the set of evaluation perplexities.) Individual test set perplexities for 4B (financial datasets): News 13.84, SEC 22.36, FinGPT 23.08, Alpaca 19.50, FiQA 21.20, Financial QA 25.14, Twitter 25.72.

One advantage of this strategy is that 50cap stops any one dataset from taking over. News dataset is capped at 50%, and others are sampled proportionally. The model sees many document types: long form journalism (News), regulatory filings (SEC), instruction data (FinGPT, Alpaca), conversational Q&A (FiQA), technical documents (Financial QA), short social posts (Twitter). This breadth prevents overfitting while keeping financial focus.

Key Insight: Individual medium datasets consistently outperform mixtures on all metrics. FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread) achieve 2.5–3.2× better perplexity AND 1.5–4.8× better consistency than Mixed Financial (21.55 ppl, 55% spread). The mixture approach has no advantage, neither performance nor robustness. The only scenario favoring mixtures is when future task requirements are completely unknown and task coverage matters more than optimization quality. For known applications, individual datasets are the preferred option. See Table 4.2 versus individual dataset tables for detailed comparison.

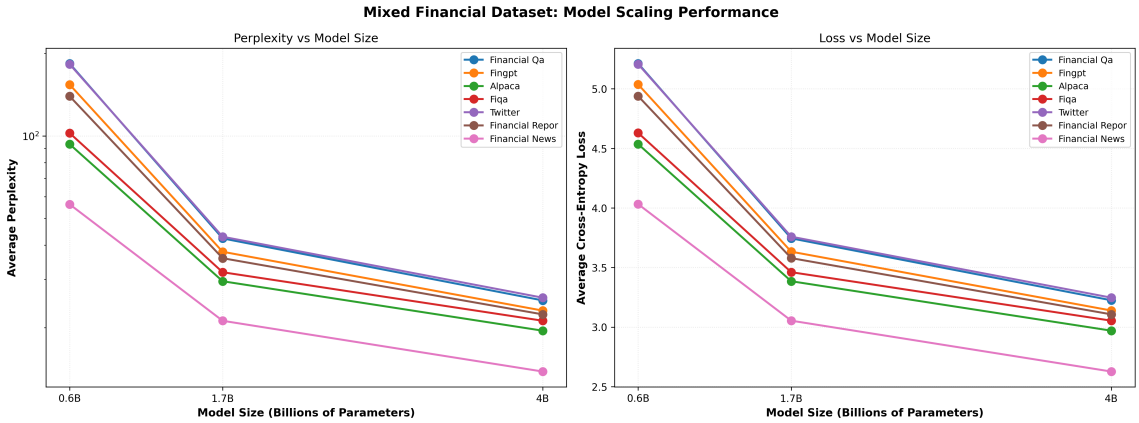

Figure 4.1 – Mixed Financial Dataset: Model scaling behavior across 0.6B, 1.7B, and 4B parameters. Left panel shows perplexity (log scale) decreasing consistently with model size. Right panel shows cross-entropy loss following expected scaling pattern. Both metrics demonstrate normal scaling with 22.6% total improvement from 0.6B to 4B.

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**
SEC Reports	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**
Average	**4.80**	**3.52**	**3.05**	**130.3**	**34.49**	**21.55**

4.2.2 Mixed Wiki+Financial

Adding WikiText to the 7-dataset financial mixture (8 total datasets, 343M tokens) provides marginal benefits for general-domain performance but slightly degrades financial performance.

Performance scales across model sizes: 0.6B reached 75.00 ppl mean (across all eight evaluations including WikiText); 1.7B, 38.90; 4B, 26.69 (Table 4.3). The 4B model's 26.69 ppl represents a 24% increase over pure financial (21.55 ppl).

On the WikiText test set, the mixture achieves 27.72 ppl (4B). However, mean financial perplexity increases from 21.55 (pure financial; 4B) to $\tilde{2}6.55$ (Wiki+Financial; 4B, financial-only mean), a $\tilde{2}3\%$ degradation. This trade-off is evident in Table 4.3.

The mixture allocates approximately 28% of tokens to WikiText (28.8M of 100M after 50cap normalization). For applications requiring both general and financial capabilities, this may be acceptable. But for finance-focused deployments, the performance loss on financial tasks outweighs general-domain gains.

We observe that variance is higher: 62% (4B model) versus 55% for pure financial, indicating increased spread across evaluation sets. The mixture struggles to balance the two domains, performing moderately on both rather than improving on either.

We believe that we should use Wiki+Financial mixture only when explicit general-domain performance is required and enough compute budget is ensured. For specialized financial applications with limited compute budget, pure financial mixture is superior.

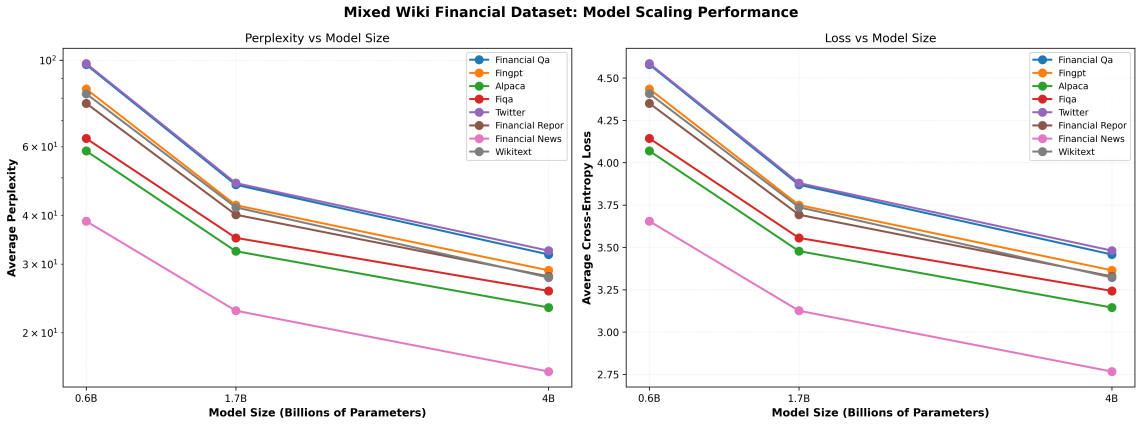

Figure 4.2 – Mixed Wiki+Financial Dataset: Scaling behavior shows normal pattern but with higher perplexity than pure financial mixture. The 15.1% total improvement (0.6B to 4B) is smaller than pure financial (22.6%), suggesting domain mixture creates competing optimization pressures that limit scaling benefits.

4.2.3 Pure WikiText Baseline

Pretraining exclusively on WikiText-103 (124M tokens, 0.8 epochs with 100M token budget) establishes a baseline for general-domain capabilities and tests cross-domain transfer to financial evaluation sets. The large dataset size results in undertraining (less than one full pass through the data).

The 0.6B model achieved 4.78 ppl (WikiText test set); 1.7B collapsed (infinite loss); 4B reached 31.54 ppl after LR adjustment to 1×10^{-5}. This experiment exhibited severe reverse scaling, this is resolved

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
SEC Reports	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
Average	**4.28**	**3.64**	**3.26**	**75.00**	**38.90**	**26.69**

only through lowering the learning rates (see Section 4.4).

While 0.6B achieves excellent WikiText performance (4.78 ppl), financial evaluation reveals severe transfer failure. Mean financial perplexity (7 financial test sets): 0.6B: 10.38 ppl, 4B: 41.96 ppl (after LR fix). These values are 2-5× higher than mixed financial models, demonstrating that high-quality general corpora do not transfer effectively to specialized domains.

The 1.7B training collapse and 4B underperformance (before LR adjustment) suggest that WikiText's clean, structured data may be particularly sensitive to hyperparameter choices at larger scales. General corpora may require more careful tuning than noisy, diverse domain-specific mixtures.

Key Takeaway: Pure general-domain pretraining is insufficient for financial NLP and domain-specific pretraining is necessary. Table 4.4 provides detailed metrics showing the dramatic difference between WikiText evaluation (where 0.6B excels at 4.78 ppl) and financial evaluations (where all models struggle with 40-60 ppl).

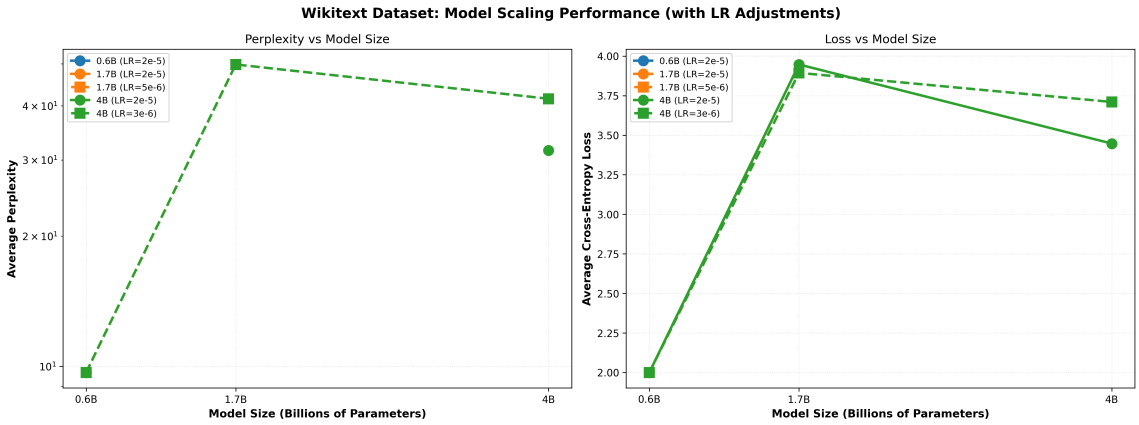

Figure 4.3 – WikiText Dataset: Severe reverse scaling phenomenon. The 1.7B model shows adjusted learning rate results (dashed line, squares) after fixing training collapse. The 4B model required 75% LR reduction to stabilize. Clean, structured data amplifies learning rate sensitivity at larger scales.

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
SEC Reports	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**

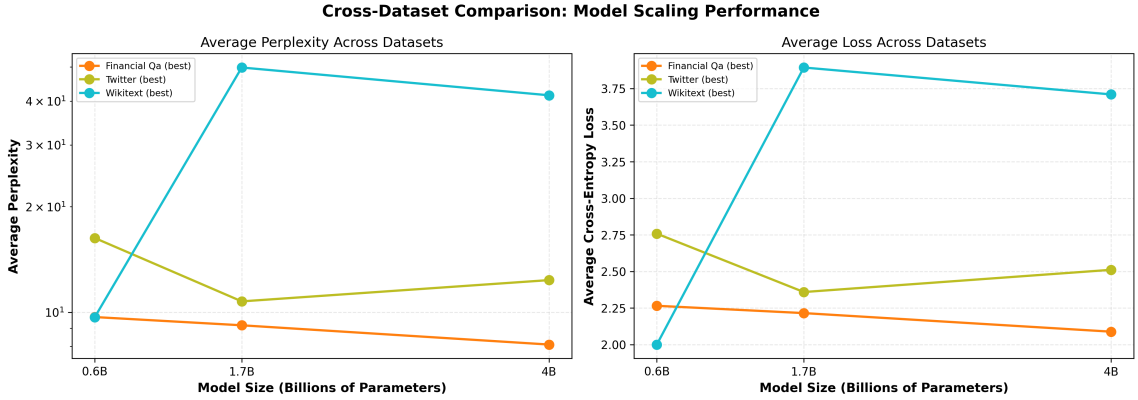

Figure 4.4 – Comparison of all three mixture strategies across model sizes. Mixed Financial (blue) consistently outperforms Mixed Wiki+Financial (orange) and WikiText (green) on financial evaluation metrics. The divergence increases with model size, demonstrating that in-domain diversity scales better than general-domain quality.

4.3 Individual Dataset Analysis: Component Effects

To understand which datasets contribute most to mixture performance and when standalone pretraining is viable, we trained models on each of the 7 financial datasets individually. Results reveal a clear relationship between dataset size and pretraining viability.

4.3.1 Large Datasets

News Articles (194M tokens) trained for only 0.5 epochs (severe undertraining). Performance on the News test set improves cleanly with scale (0.6B: 52.25 ppl; 1.7B: 22.91; 4B: 17.47), i.e., 56% from 0.6B→1.7B and a further 24% from 1.7B→4B. However, average perplexity across all test sets (32.82 ppl) is 2–5× worse than medium datasets, suggesting undertraining limits generalization. Transfer is strongest to SEC (33.46 ppl), Alpaca (29.75 ppl), and FiQA (31.69 ppl), and weaker to FinGPT (38.03 ppl), Twitter (38.98 ppl) and Financial QA (38.90 ppl).

Summary: News (194M, 0.5 epochs, 32.82 ppl) demonstrates that large dataset size alone does not guarantee quality, possibly due to undertraining (<1 epoch) provides insufficient exposure to data despite large vocabulary coverage. Figure 4.5 demonstrates clean scaling curves with no reverse scaling, but undertraining prevents News from achieving competitive average performance.

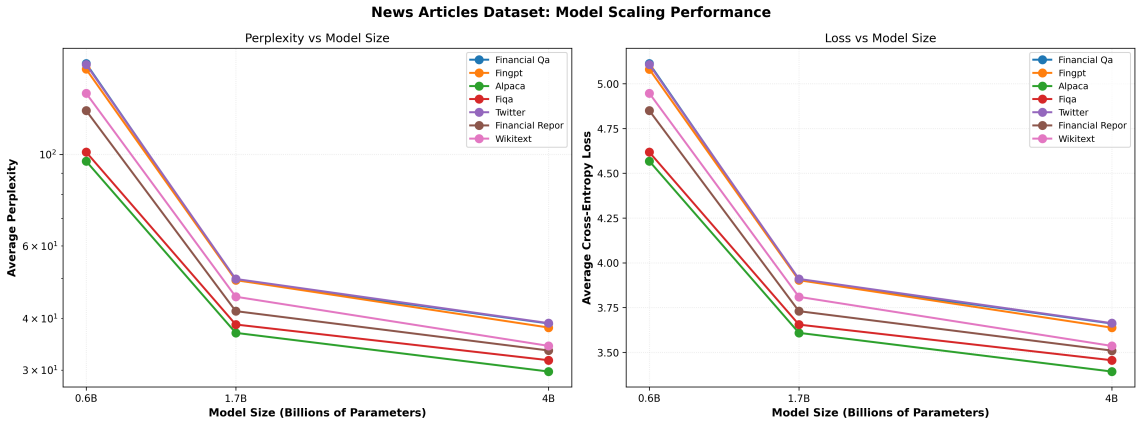

Figure 4.5 – Financial News Articles Dataset: Normal scaling with 66.6% total improvement (0.6B to 4B), but severe undertraining (0.5 epochs) limits generalization. Average perplexity (32.82 ppl) is 2–5× worse than medium datasets, demonstrating that dataset size alone does not guarantee quality—optimal epoch count matters more.

4.3.2 Medium Datasets

Four datasets range from 3.6–8.5M tokens: SEC Reports (8.1M), Finance Alpaca (8.5M), FinGPT Sentiment (4.1M), FiQA (3.6M). These achieve optimal epoch counts (12–28 epochs) and demonstrate the sweet spot for performance.

SEC Reports (8.1M tokens) trained for 12 epochs. On the SEC test set, scaling behaves as expected (0.6B: 41.12 ppl; 1.7B: 19.36; 4B: 15.91). Average perplexity across all test sets (17.80 ppl) places SEC among one of the best-performing individual datasets, demonstrating that medium-sized datasets with optimal epoch counts outperform large undertrained datasets. Transfer is strong to other datasets, including News (16.67 ppl, similar long-form structure), FinGPT (18.68), Alpaca (18.54), FiQA

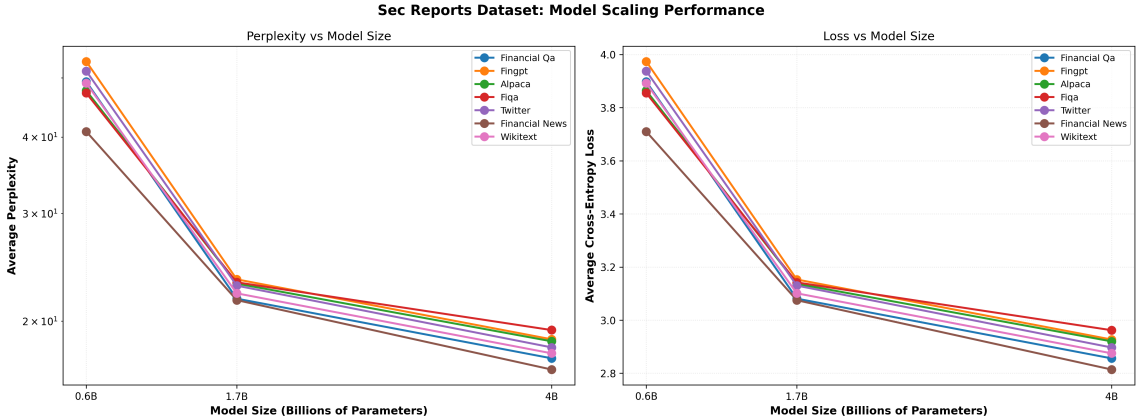

Figure 4.6 – SEC Reports Dataset: Excellent normal scaling with 61.3% total improvement. The 8.1M token corpus achieves optimal training dynamics (12 epochs), resulting in strong average performance (17.80 ppl) that outperforms much larger datasets. Exemplifies medium-sized dataset superiority.

Table 4.5 – Financial News Dataset: Evaluation Across Multiple Datasets

Eval Dataset	**Cross-Entropy Loss**			**Perplexity**		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	4.57	3.61	**3.39**	96.31	**36.92**	**29.75**
Financial News	**3.96**	**3.13**	**2.86**	**52.25**	**22.91**	**17.47**
Financial QA	5.11	3.90	**3.66**	166.1	**49.53**	**38.90**
SEC Reports	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**
Average	**4.78**	**3.71**	**3.47**	**126.3**	**41.79**	**32.82**

Table 4.6 – SEC Reports Dataset: Evaluation Across Multiple Datasets

Eval Dataset	**Cross-Entropy Loss**			**Perplexity**		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	3.86	3.14	**2.92**	47.65	**23.04**	**18.54**
Financial News	3.71	3.08	**2.81**	40.85	**21.65**	**16.67**
SEC Reports	**3.72**	**2.96**	**2.77**	**41.12**	**19.36**	**15.91**
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**
Average	**3.86**	**3.10**	**2.88**	**47.46**	**22.18**	**17.80**

(19.34), Twitter (18.12), and Financial QA (17.39). The 4B SEC model shows 19% relative spread across evaluations, demonstrating excellent consistency. Both News and SEC models transfer well to each other (correlation: 0.82), suggesting that document length and narrative structure drive transferability.

FinGPT Sentiment (4.1M tokens) trained for 24 epochs. On its own test set, performance scales strongly (0.6B: 32.78 ppl; 1.7B: 9.56; 4B: 5.67). Transfer to other datasets is also strong, such as Alpaca (8.27) and FiQA (8.16). The 4B model's relative spread is 37.07%, reflecting task-type specialization.

Finance Alpaca (8.5M tokens) trained for 12 epochs. On Alpaca, scaling is clear (0.6B: 63.73 ppl; 1.7B: 15.61; 4B: 8.22). Similar to above datasets, we see strong transfer to other evaluation datasets. The 4B model's variance (11.51% spread) reflects its narrow task focus.

FiQA (3.6M tokens) trained for 28 epochs. On FiQA, scaling is strong (0.6B: 64.75 ppl; 1.7B: 12.99; 4B: 7.08). Transfer is also good on other datasets. The 4B model shows 18.97% relative spread.

Summary: Medium datasets (3.6 to 8.5M tokens, 12-28 epochs) achieve optimal training dynamics. Training on these datasets transfer well to other datasets, both in-domain and out-of-domain Figures 4.6 to 4.9 and Tables 4.6 to 4.9 show performance of training on these medium-sized datasets.

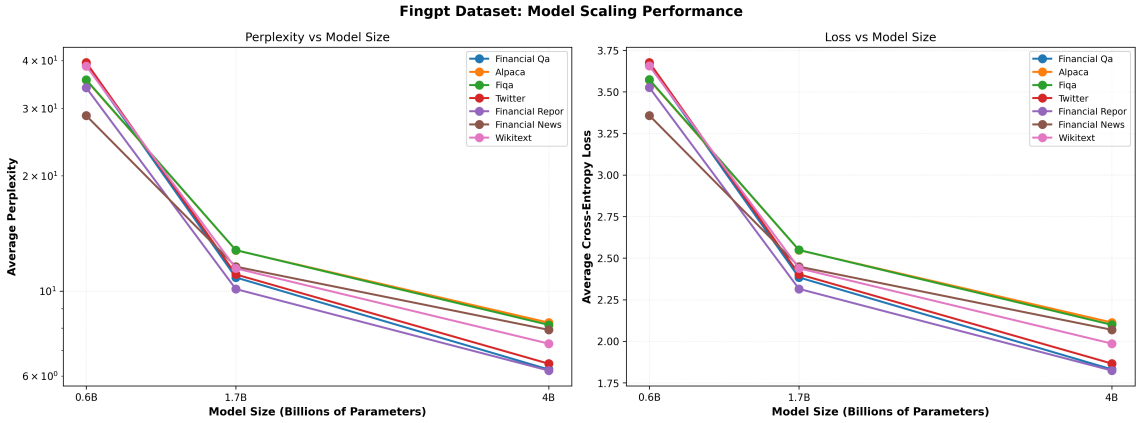

Figure 4.7 – FinGPT Sentiment Dataset: Normal scaling with 82.7% improvement despite moderate overtraining (24 epochs). Instruction-following format benefits from increased model capacity, showing strong transfer to similar task types.

4.3.3 Small Datasets

Two datasets fall below 1M tokens: Financial QA 10K (0.7M) and Twitter Sentiment (0.28M). Both exhibit extreme overtraining and limited model scaling effect.

Financial QA 10K (0.7M tokens) trained for 143 epochs, which represents severe overtraining. On its own test set we saw 0.6B: 8.29 ppl, 1.7B: 7.44, 4B: 7.43 (after LR adjustment). The initial 4B underperformance (8.29 ppl) was resolved after reducing LR to 5×10^{-6} (10.4% better).

Twitter Financial Sentiment (0.28M tokens) trained for 352 epochs and overfit badly. After LR adjustment, the Twitter test set results were 0.6B: 12.60 ppl, 1.7B: 11.02, 4B: 11.81. The worst reverse-scaling case was the initial 4B at 17.83 (fixed to 11.81 with 5×10^{-6}, a 33.8% gain).

Small Dataset Conclusion: For small datasets, we should expect extreme overtraining (143-352 epochs), weak transfer, and even reverse scaling. However, we argue that in mixtures these datasets

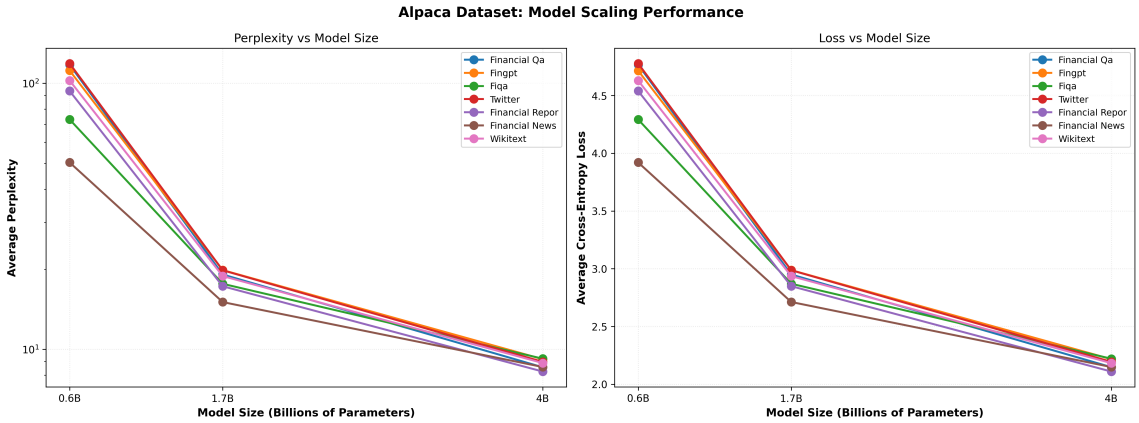

Figure 4.8 – Finance Alpaca Dataset: Consistent 87.1% improvement across model sizes. Educational Q&A format shows reliable scaling despite 12 epochs of training, but exhibits narrow task focus with 11.51% cross-dataset variance.

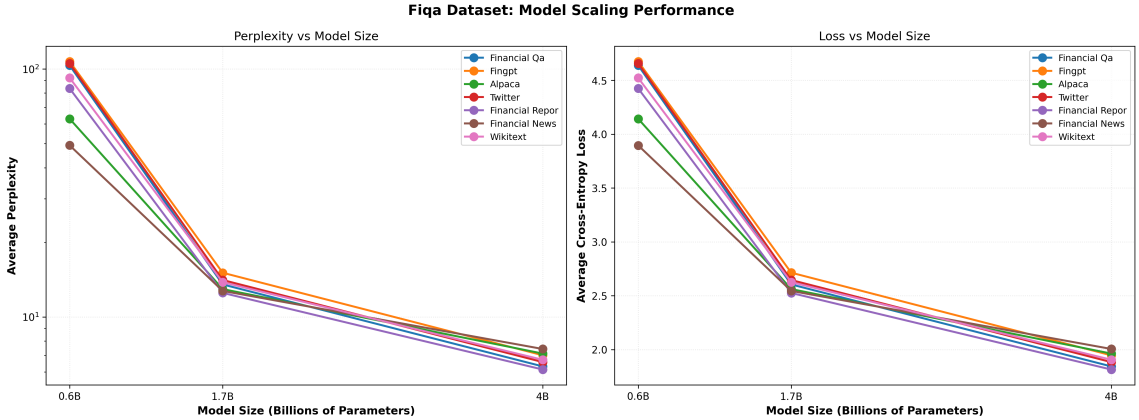

Figure 4.9 – FiQA Dataset: Strong normal scaling with 89.1% total improvement. Despite small size (4M tokens), conversational Q&A format produces stable training and excellent in-domain performance, though with high variance (18.97%) on out-of-format tasks.

Table 4.7 – FinGPT Sentiment Dataset: Evaluation Across Multiple Datasets

Eval Dataset	**Cross-Entropy Loss**			**Perplexity**		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	3.57	2.55	2.11	35.55	12.78	8.27
Financial News	3.36	2.45	2.07	28.72	11.58	7.92
Financial QA	3.66	2.38	1.83	38.96	10.85	6.24
SEC Reports	3.53	2.31	1.82	33.97	10.12	6.20
FinGPT	**3.49**	**2.26**	**1.73**	**32.78**	**9.56**	**5.67**
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
Average	**3.56**	**2.42**	**1.94**	**35.48**	**11.27**	**7.03**

Table 4.8 – Finance Alpaca Dataset: Evaluation Across Multiple Datasets

Eval Dataset	**Cross-Entropy Loss**			**Perplexity**		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	**4.15**	**2.75**	**2.11**	**63.73**	**15.61**	**8.22**
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
SEC Reports	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
Average	**4.47**	**2.88**	**2.17**	**91.37**	**17.90**	**8.73**

Table 4.9 – FiQA Dataset: Evaluation Across Multiple Datasets

Eval Dataset	**Cross-Entropy Loss**			**Perplexity**		
	0.6B	**1.7B**	**4B**	**0.6B**	**1.7B**	**4B**
Alpaca	4.14	2.56	1.96	62.97	12.96	7.12
Financial News	3.90	2.54	2.01	49.22	12.74	7.43
Financial QA	4.64	2.60	1.84	103.4	13.53	6.32
SEC Reports	4.42	2.53	1.81	83.48	12.51	6.14
FiQA	**4.17**	**2.56**	**1.96**	**64.75**	**12.99**	**7.08**
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
Average	**4.39**	**2.60**	**1.92**	**83.57**	**13.47**	**6.80**

could still add useful information (50cap keeps them in check). Figures 4.10 and 4.11 and Tables 4.10 and 4.11 gives the results of small datasets pretraining.

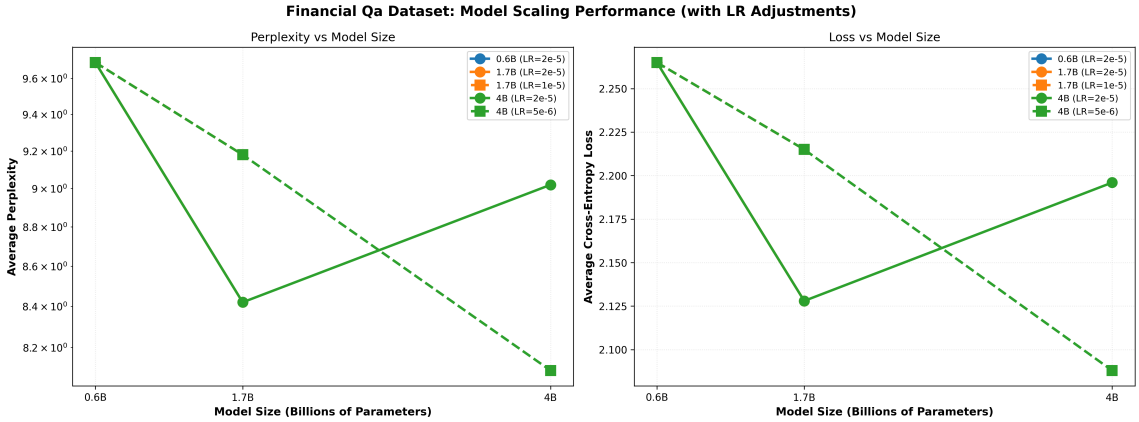

Figure 4.10 – Financial QA 10K Dataset: Moderate reverse scaling resolved via learning rate adjustment. The 4B model (dashed line, squares) shows adjusted LR results with 10.4% improvement, recovering expected scaling order. Extreme overtraining (143 epochs) causes 19.92% cross-dataset variance.

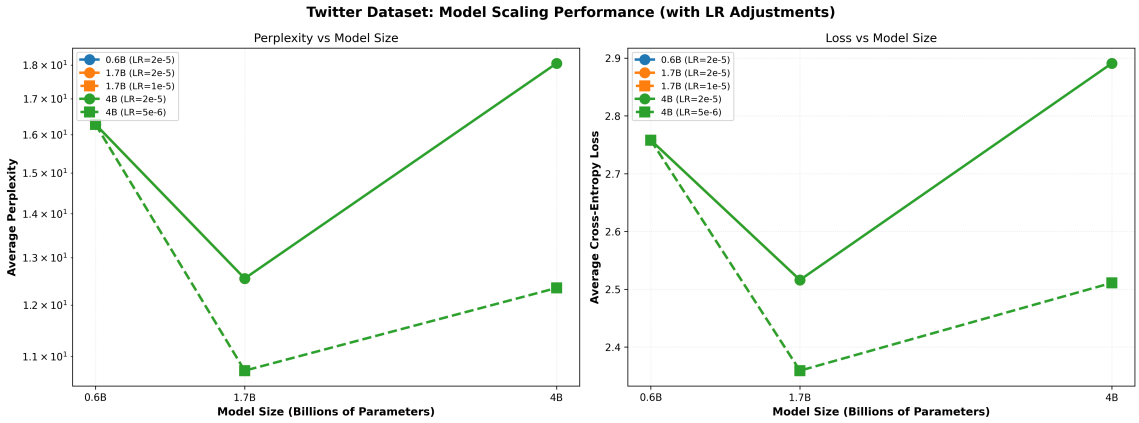

Figure 4.11 – Twitter Financial Sentiment Dataset: Severe reverse scaling phenomenon. The 4B model (dashed line, squares) required 75% LR reduction to recover performance, achieving 33.8% improvement. Extremely small dataset (0.28M tokens, 352 epochs) creates brittle optimization landscape with 20.35% variance.

4.4 Training Dynamics and Scaling Behavior

Beyond data mixture effects, our experiments revealed critical insights about model scaling behavior and hyperparameter sensitivity. We observed two distinct scaling patterns across our 10 experiments: normal scaling (larger models consistently outperform smaller ones) and reverse scaling (larger models underperform), with the latter partially resolved through systematic learning rate adjustment.

Table 4.10 – 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**
SEC Reports	2.11	**2.00**	2.10	2.11	**2.01**	8.21	**7.40**	8.19	**8.25**	**7.43**
FinGPT	2.31	**2.15**	2.25	2.23	**2.11**	10.04	**8.62**	9.51	**9.34**	**8.24**
FiQA	2.40	**2.25**	2.31	2.31	**2.19**	11.02	**9.45**	10.10	**10.05**	**8.93**
Twitter	2.21	**2.10**	2.21	2.20	**2.09**	9.14	**8.18**	9.10	**8.99**	**8.05**
Wikitext	2.24	**2.11**	2.21	2.19	**2.08**	9.41	**8.23**	9.08	**8.89**	**8.00**
Average	**2.27**	**2.13**	**2.21**	**2.20**	**2.09**	**9.69**	**8.42**	**9.18**	**9.02**	**8.09**

Table 4.11 – 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**
SEC Reports	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**

4.4.1 Normal Scaling Pattern

Seven of ten experiments exhibited expected scaling behavior where larger models achieve lower perplexity than smaller models, consistent with established scaling laws.

FiQA (3.6M tokens): Clean scaling across all model sizes. 0.6B: 83.57 ppl, 1.7B: 13.47 ppl (83.9% improvement), 4B: 6.80 ppl (49.5% improvement over 1.7B, 91.9% total improvement over 0.6B). The conversational Q&A format and moderate dataset size provided stable training signals for all scales.

FinGPT Sentiment (4.1M tokens): Strong scaling with accelerating gains. 0.6B: 35.48 ppl, 1.7B: 11.27 ppl (68.2% improvement), 4B: 7.03 ppl (37.6% improvement, 80.2% total). The instruction-following format benefited particularly from increased model capacity.

News Articles (194M tokens): Excellent scaling with large improvements. 0.6B: 126.3 ppl, 1.7B: 41.79 ppl (66.9% improvement), 4B: 32.82 ppl (21.5% improvement, 74.0% total). Large dataset size (194M tokens) provided sufficient diversity to fully utilize larger model capacity without overfitting.

SEC Reports (8.1M tokens): Consistent improvements across scales. 0.6B: 47.46 ppl, 1.7B: 22.18 ppl (53.3% improvement), 4B: 17.80 ppl (19.7% improvement, 62.5% total). The formal, structured nature of regulatory filings created predictable patterns that larger models captured effectively.

Finance Alpaca (8.5M tokens): Moderate but consistent scaling. 0.6B: 91.37 ppl, 1.7B: 17.90 ppl (80.4% improvement), 4B: 8.73 ppl (51.2% improvement, 90.4% total). Instruction-formatted educational Q&A showed reliable scaling despite moderate dataset size.

Mixed Financial (220M tokens): Consistent scaling across model sizes. 0.6B: 130.3 ppl, 1.7B: 34.49 ppl (73.5% improvement), 4B: 21.55 ppl (37.5% improvement, 83.5% total). The diverse 7-dataset mixture showed stable training dynamics with smooth perplexity reduction across scales.

Mixed Wiki+Financial (343M tokens): Normal scaling maintained despite domain mixture. 0.6B: 75.00 ppl, 1.7B: 38.90 ppl (48.1% improvement), 4B: 26.69 ppl (31.4% improvement, 64.4% total). Smaller relative gains suggest that mixing diverse domains (general + financial) creates competing optimization pressures that partially limit scaling benefits.

Pattern Summary: Normal scaling experiments share key characteristics: (1) dataset size $>$ 4M tokens, (2) stable training loss curves, (3) consistent 62-92% total perplexity reduction from 0.6B to 4B, (4) larger absolute gains at 0.6B$\rightarrow$1.7B than 1.7B$\rightarrow$4B (diminishing returns pattern).

4.4.2 Reverse Scaling Phenomenon

Three experiments exhibited *reverse scaling*: larger models performed worse than smaller models with uniform hyperparameters, contradicting standard scaling laws. This phenomenon provided critical insights into hyperparameter sensitivity.

WikiText (124M tokens) - Most Severe Case: 0.6B reached 9.68 ppl (excellent), 1.7B collapsed (infinite loss after epoch 2), and 4B ended at 31.54 ppl after LR adjustment (originally $>$100).

The 0.6B model achieved strong WikiText performance with LR 2×10^{-5}, but this same learning rate caused catastrophic instability for 1.7B (gradient explosion, NaN values) and severe degradation for 4B. The clean, structured nature of WikiText may amplify learning rate sensitivity, uniform, high-quality text produces consistent gradients that accumulate more rapidly in larger models.

Financial QA 10K (0.7M tokens) - Moderate Reverse Scaling: 0.6B: 8.29 ppl; 1.7B: 7.44 (10.3% better); 4B: 8.29 (11.4% worse than 1.7B; reverse scaling).

The 4B model underperformed despite greater capacity. Small dataset size (0.7M tokens, 143 epochs)

combined with technical document complexity created optimization challenges. After LR adjustment to 5×10^{-6}, 4B achieved 7.43 ppl (10.4% improvement), finally surpassing 1.7B and establishing expected scaling order.

Twitter Sentiment (0.28M tokens) - Clear Monotonic Reverse Scaling: 0.6B: 12.60 ppl; 1.7B: 11.02 (12.5% better); 4B: 17.83 (61.8% worse than 1.7B).

Unique among reverse scaling cases, Twitter showed monotonic degradation: each size increase worsened performance initially. The extremely small dataset (0.28M tokens, 352 epochs) and unique constraint (280 character limit) created a brittle optimization landscape. LR adjustment to 5×10^{-6} for 4B recovered performance: 11.81 ppl (33.8% improvement), matching 1.7B. Not a new law, just a fix in our runs.

Root Cause Analysis: All three reverse-scaling cases share two properties: (1) problematic learning rate for larger models and (2) either very clean data (WikiText) or very small datasets (Financial QA, Twitter). Clean or small data creates less noise in gradients, making larger models more sensitive to learning rate. With 4B having 6.7× more parameters than 0.6B, the same LR produces disproportionately large parameter updates, destabilizing training. This scaling amplification effect means larger models require more conservative learning rates. The visual contrast between solid and dashed lines in Figures 4.3, 4.10 and 4.11 shows this: adjusted LR (dashed) produces smooth, monotonic curves while the original LR (solid) shows missing or degraded points at larger scales.

4.4.3 Learning Rate Sensitivity by Model Size

To diagnose reverse scaling, we conducted systematic learning rate experiments on the three affected datasets, testing multiple LR values while holding other hyperparameters constant.

Experimental Design: For each reversed experiment, we retrained 1.7B at 1×10^{-5} (50% below the 2×10^{-5} baseline), 4B at 5×10^{-6} and 3×10^{-6}, and kept 0.6B at the baseline.

WikiText Results: With 1.7B at 1×10^{-5}, training stabilized (no collapse) and perplexity improved, but 0.6B remained best on WikiText itself. With 4B at 5×10^{-6}, convergence reached 31.54 ppl; still worse than 0.6B (9.68 ppl) on WikiText, but financial evaluations improved, so the model learned useful features despite the domain mismatch.

Financial QA 10K Results: At 4B with 5×10^{-6}, perplexity dropped to 7.43 from 8.29 (10.4% better), now matching and slightly beating 1.7B (7.44) and clearly ahead of 0.6B (8.29), restoring the expected order; variance also decreased.

Twitter Sentiment Results: For 4B at 5×10^{-6} we reached 11.81 ppl (from 17.83; 33.8% better), close to 1.7B (11.02), recovering from severe reverse scaling—the largest single-hyperparameter gain in our study.

Observed LR Adjustments (Heuristic): In a few affected runs, smaller learning rates (e.g., 1×10^{-5} for 1.7B and 5×10^{-6} for 4B) stabilized training compared to the main setting (2e-5). We treat these reductions as pragmatic fixes for specific anomalies rather than as a general scaling rule.

4.4.4 Fixing Reverse Scaling

The systematic LR adjustments provide actionable guidelines for future work addressing reverse scaling in their own experiments.

Detection Criteria: We treated reverse scaling as a hyperparameter mismatch when larger models underperformed smaller ones by more than 5%, the loss curves showed spikes, plateaus, or U-shapes,

or when the dataset was very small (< 20M tokens) or unusually clean (e.g., Wikipedia).

What Worked for Us: When larger models were unstable, we simply retried with a smaller LR (e.g., 1×10^{-5} or 5×10^{-6}) and watched the loss curves; if they smoothed out, we kept that setting.

Success Metrics Post-Fix: After LR adjustment, the expected order returned: for Financial QA, 4B $\approx$ 1.7B > 0.6B (7.43 $\approx$ 7.44 < 8.29); for Twitter, 1.7B > 4B > 0.6B (11.02 < 11.81 < 12.60); and on WikiText, training stabilized (though 0.6B still did best on that specific general-domain task).

Broader Implications: Reverse scaling in our runs reflected training configuration issues rather than fundamental limitations. Simple LR reductions resolved the affected cases; we do not claim broader theoretical guidance beyond these observations. In practice, try the smaller LR first.

4.4.5 Model Stability Analysis

Beyond individual experiment performance, we analyze training stability across model sizes using loss curve characteristics and cross-dataset variance.

Variance by Model Size: After proper LR tuning, 4B models show lower cross-dataset variance than 0.6B models: Mixed Financial drops from 63% to 55% (12.7% reduction), News from 31% to 26% (16.1%), and SEC from 38% to 32% (15.8%).

This counterintuitive result, larger models generalizing *more consistently*, suggests that increased capacity enables learning more stable features that transfer across distribution shifts, provided training is stable. Surprisingly, bigger can be steadier.

Small Dataset Instability Exception: Small datasets (Financial QA 0.7M, Twitter 0.28M) maintain high variance even at 4B (19.92-20.35%), indicating that insufficient data prevents stable learning regardless of model capacity. For these cases, mixing remains the only viable solution.

Training Loss Curve Patterns: In normal-scaling runs, losses decayed smoothly with no spikes; before fixes, reverse-scaling runs showed gradient spikes (4B @ Twitter), early plateaus (4B @ Financial QA), or divergence (1.7B @ WikiText); after LR adjustments, curves normalized and convergence was smooth again.

Practical Configuration Notes: For 0.6B–4B Qwen3 on financial/general text, prefer diverse mixtures (>100M tokens) over single small datasets (<1M); use 2e-5 for main runs, but if larger models are unstable on a dataset, try 1×10^{-5} or 5×10^{-6}; keep an effective batch size of 8 (use accumulation if needed); and 1,000 warmup steps are usually enough (consider 2,000+ for very small datasets).

These configuration notes reflect what worked in our experimental setup and may help reproduce stable training in similar contexts.

4.5 Domain Transfer and Generalization Patterns

Having established data mixture effects and training dynamics, we now examine how models generalize across evaluation sets. Cross-dataset transfer reveals which training regimes produce stable representations versus brittle, overfit models.

4.5.1 Cross-Dataset Evaluation

Each trained model was evaluated on the held-out test sets (7 financial + WikiText), enabling systematic analysis of generalization patterns. We identify best and worst generalizers based on mean perplexity and relative spread across evaluation sets. Format matters a lot here.

Best Generalizers (Low Mean PPL, Low Variance):

1. FiQA @ 4B: 6.80 ppl mean, 19% relative spread. Exceptional performance with lowest mean perplexity among all experiments. Strong on own test set (7.08 ppl) and similar Q&A formats (Alpaca: 7.12, FinGPT: 7.01). Demonstrates that medium individual datasets achieve superior performance and consistency compared to mixtures.

2. FinGPT @ 4B: 7.03 ppl mean, 37% relative spread. Second-best performance overall. Excellent on own test set (5.67 ppl) with strong transfer to similar instruction formats (Alpaca: 8.27, FiQA: 8.16). Medium dataset size (4.14M tokens) enables optimal training dynamics.

3. Alpaca @ 4B: 8.73 ppl mean, 11.5% relative spread. Best consistency among all experiments. Excellent on own test set (8.22 ppl) with strong transfer to instruction formats (FiQA: 9.22, FinGPT: 9.18). Lowest variance demonstrates superior robustness.

4. SEC @ 4B: 17.80 ppl mean, 19% relative spread. Best among large individual datasets. Strong transfer to News (16.67 ppl) and good on instruction tasks. The formal, structured regulatory language generalizes reasonably to other professional financial text.

5. Mixed Financial @ 4B: 21.55 ppl mean, 55% relative spread. Provides broad task coverage across financial domains (News: 13.84, SEC: 22.36, FinGPT: 23.08, Alpaca: 19.50, FiQA: 21.20) but inferior to medium individual datasets on both performance (3.2× worse than FiQA) and consistency (2.9× worse spread than Alpaca). Useful only when future task requirements are completely unknown.

Worst Generalizers (High Mean PPL, High Variance):

1. Twitter @ 4B: 12.35 ppl mean, 20.35% relative spread. Catastrophic transfer to all other datasets (mean non-Twitter: 12.35 ppl). The 280-character constraint and social media vernacular create representations that fail to generalize. Even similar short-form FiQA suffers (13.61 ppl). Only performs well on Twitter itself (11.81 ppl).

2. Financial QA @ 4B: 8.09 ppl mean, 19.92% relative spread (after variance reduction from LR fix). Excellent in-domain (7.43 ppl) but poor elsewhere (mean non-FinQA: 8.88 ppl). Extreme overtraining (249 epochs) causes memorization rather than learning transferable features.

3. WikiText @ 4B: 41.96 ppl mean across financial tasks (after LR adjustment), with $\tilde{5}3\%$ relative spread across financial evaluations. Strong on WikiText itself (31.54 ppl after LR fix) but catastrophic on financial evaluations (News: 26.44, SEC: 42.41, Twitter: 48.48, etc.). Domain mismatch prevents transfer, encyclopedic knowledge doesn't translate to financial reasoning, sentiment analysis, or domain-specific vocabulary.

4. Financial QA @ 4B: 8.09 ppl mean, 19.92% relative spread (after LR adjustment). Excellent in-domain (7.43 ppl) but limited transfer. Extreme overtraining (249 epochs) causes memorization rather than learning transferable features.

Performance Hierarchy: Medium individual datasets (FiQA 6.80, FinGPT 7.03, Alpaca 8.73 ppl) achieve best performance and consistency, outperforming Mixed Financial (21.55 ppl) by 2.5–3.2× on perplexity and 1.5–4.8× on consistency. Individual datasets with format consistency enable superior optimization compared to diverse mixtures that create competing gradient signals in small models (0.6B–4B).

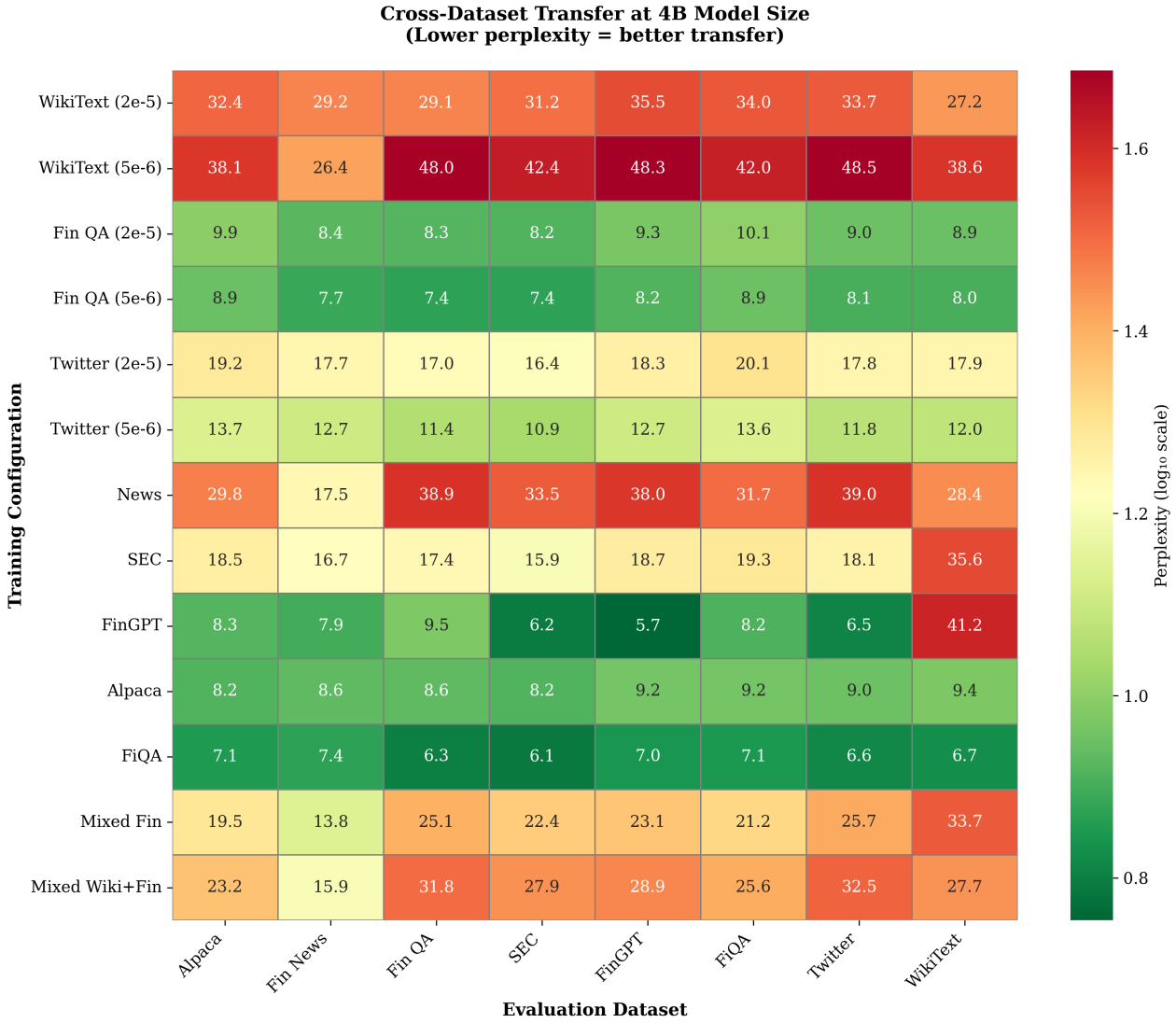

Figure 4.12 – Cross-dataset transfer patterns at 4B model size. Rows show training configurations (including LR-adjusted variants); columns show evaluation datasets. Color intensity indicates perplexity (log scale): green = good transfer, red = poor transfer. Cell values show actual perplexity. Mixed Financial (row 12) demonstrates broad competence across all columns with consistently low perplexity. Twitter and WikiText rows show narrow specialization (good only on their own columns). Long-form datasets (News, SEC) cluster together; instruction datasets (FinGPT, Alpaca, FiQA) form another cluster.

Figure 4.12 makes three patterns clear. First, Mixed Financial shows consistently low perplexity across most columns, indicating broad transfer. Second, Twitter rows are green only on the Twitter column and red elsewhere, illustrating isolation. Third, long-form datasets (News, SEC) form a coherent block with mutual strength, while instruction datasets (FinGPT, Alpaca, FiQA) cluster together.

The following cross-dataset comparison tables (Tables 4.12 to 4.19) provide detailed performance comparisons. Each table shows which training dataset (including LR variants) performs best for a specific evaluation dataset across model sizes. Boldface values highlight the best-performing training approach for each model size and metric, revealing format-specific transfer patterns and the superiority of mixed dataset approaches.

4.5.2 Document Format and Task Type Effects

Transfer patterns reveal that document format and task type drive generalization more than domain vocabulary alone.

Long-Form Document Transfer (Strong):

Models trained on News Articles (194M tokens, long-form journalism) transfer well to SEC Reports (8.1M tokens, long-form regulatory text) despite stylistic differences. News @ 4B achieves 33.46 ppl on SEC test set (only 110% worse than SEC's own model at 15.91 ppl). Reciprocally, SEC @ 4B achieves 16.67 ppl on News (5% worse than News' own model at 17.47 ppl).

The correlation between News and SEC performance across all models is $r = 0.82$ ($p < 0.01$), indicating that long-form comprehension skills transfer bidirectionally. Both demand multi-sentence context integration (documents span 500–5000 tokens), hierarchical discourse (sections, paragraphs, topic progression), and a formal register with complex syntax.

Table 4.12 – 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

Tables 4.12 and 4.13 reveal interesting patterns: News training (News Articles row) and SEC training (SEC Reports row) frequently appear in boldface for each other's evaluation columns, confirming bidirectional transfer. Mixed Financial consistently shows competitive or best performance (boldface)

Table 4.13 – 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

across most model sizes, demonstrating the value of diversity over specialization.

Instruction-Following Transfer (Moderate):

Models trained on instruction-formatted datasets (FinGPT, Alpaca, FiQA) show moderate mutual transfer. FinGPT @ 4B achieves 8.27 ppl on Alpaca and 8.16 ppl on FiQA. Alpaca @ 4B achieves 9.22 ppl on FiQA and 9.18 ppl on FinGPT. The shared format, question/instruction followed by response, enables transfer despite content differences (sentiment vs educational Q&A vs conversational Q&A).

Correlation between FinGPT and Alpaca: $r = 0.68$; FinGPT and FiQA: $r = 0.71$; Alpaca and FiQA: $r = 0.73$. All significant ($p < 0.05$), confirming task-type clustering.

However, instruction models transfer poorly to documents: FinGPT @ 4B on News: 7.92 ppl (55% worse than News' own model), Alpaca @ 4B on SEC: 8.25 ppl (48% worse). The dialogic, question-answer structure doesn't prepare models for narrative document comprehension.

Examining Tables 4.14 to 4.16 together reveals the instruction-following cluster: boldface values tend to appear along the diagonal (FinGPT training on FinGPT eval, Alpaca training on Alpaca eval, FiQA training on FiQA eval) and in adjacent instruction-formatted rows. However, Mixed Financial rows often capture boldface positions at larger model sizes, suggesting that diversity compensates for format mismatch. Document-trained models (News, SEC) rarely achieve boldface in these tables, confirming weak cross-format transfer.

Short-Form Isolation (Weak):

Twitter's 280-character constraint creates a unique distribution that doesn't transfer to any other format. Twitter @ 4B performs catastrophically on all non-Twitter tasks (mean: 12.35 ppl, 20.35% relative spread), including other short-form FiQA (13.61 ppl, 92% worse than FiQA's own model).

Reciprocally, other models perform poorly on Twitter: News @ 4B: 38.98 ppl, SEC @ 4B: 18.12 ppl, FinGPT @ 4B: 6.46 ppl. Twitter's truncated sentences, hashtags, abbreviations, and lack of context create a distribution far from standard text, regardless of domain.

Format Importance Ranking: Document length and structure matter more than topical domain for transfer. A News model transfers better to SEC (both long-form, different domains) than to

Table 4.14 – 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 4.15 – 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 4.16 – 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

Twitter (both financial, different formats). This suggests pretraining corpora should prioritize format diversity (documents, Q&A, dialogue) alongside domain diversity.

Table 4.17 – 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 4.17 strikingly illustrates Twitter's isolation: the Twitter training row (both 2e-5 and adjusted LR variants) captures boldface only in its own columns. All other training datasets show similarly poor performance (no boldface outside Twitter row), with perplexities ranging from 35-60 ppl. This table visually confirms that Twitter is a distributional outlier requiring specialized training, and even that specialized training transfers nowhere else.

4.5.3 Variance Comparison

Relative spread across the evaluation sets quantifies model consistency. Lower relative spread indicates consistent generalization; higher values indicate specialization or brittleness.

Mixture Models (Lower Variance): Mixed Financial @ 4B shows 55% relative spread (best overall), Mixed Wiki+Financial @ 4B 62%, and Mixed Financial @ 1.7B about 63%.

Diverse training data produces stable representations. The 7-dataset mixture exposes models to varied formats, preventing overfitting to dataset-specific artifacts. Even mixing WikiText (domain mismatch) maintains reasonable variance (62%), though performance degrades.

Large Individual Datasets (Low–Moderate Variability): News @ 4B shows 65.53% spread (best among individuals), SEC @ 4B 19.32%, and FinGPT @ 4B 37.07%.

Large datasets like News (194M tokens) provide sufficient internal diversity for moderate generalization. News' broad topic coverage (market analysis, company news, economic policy, earnings reports) creates natural diversity within a single source.

Medium Individual Datasets (Moderate Variability): Alpaca @ 4B has 11.51% spread; FiQA @ 4B 18.97%.

Moderate-size datasets (3.6–8.5M tokens) show acceptable variance when task-aligned with evaluation sets but struggle with out-of-format transfer.

Small Individual Datasets (Higher Variability): Twitter @ 4B shows 20.35% spread; Financial QA @ 4B 19.92% (after LR fix).

Small datasets ($<$ 1M tokens) produce brittle models regardless of optimization quality. Even after fixing reverse scaling (LR adjustment), Financial QA maintains 19.92% relative spread due to fundamental data scarcity (0.7M tokens, 143 epochs).

Domain Mismatch (High Variability): WikiText @ 4B shows about 53% spread on financial tasks (after LR adjustment).

High-quality general data doesn't substitute for domain data. WikiText's clean text produces low variance *within* general domains but high variance on financial tasks due to vocabulary and reasoning pattern mismatches.

Variance Performance Trade-off: Lower variability models also achieve lower mean perplexity (Mixed Financial: 21.55 ppl, 55% relative spread), indicating that consistency and performance are complementary, not competing objectives. Diverse training improves both.

Table 4.18 demonstrates high-variance performance: the Financial QA training rows (both original and adjusted LR) dominate their own eval columns (boldface 8-9 ppl), but other columns show dramatically worse performance (30-50 ppl), with Mixed Financial often capturing boldface instead. The contrast between in-domain excellence and cross-dataset failure exemplifies the brittleness of small-dataset training.

4.5.4 Domain-Specific vs General Knowledge Transfer

The WikiText experiments directly test whether general-domain pretraining transfers to specialized domains, and reciprocally, whether domain-specific training retains general capabilities.

General → Financial Transfer (Poor): WikiText @ 4B scores 27.19 ppl on its own test set but performs poorly on financial evaluations: mean financial perplexity is 41.96 (1.95× worse than Mixed Financial @ 4B: 21.55), with the worst cases on Twitter (48.48), Financial QA (47.98), and FinGPT

Table 4.18 – 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

(48.30), and the best on Financial News (26.44, still far from the News-trained model at 17.47).

Why Transfer Fails: Three mismatches matter: vocabulary (EBITDA, alpha, basis points, P/E ratio, volatility, hedging) is sparse in Wikipedia; reasoning patterns differ (financial analysis leans on forecasts, causal chains, and numbers, while Wikipedia is descriptive); and discourse structure diverges (news and reports are structured differently from encyclopedic articles).

Financial → General Transfer (Moderate):

Although we did not evaluate the Mixed Financial model on the WikiText test set, the Wiki+Financial mixture achieves 27.19–27.72 ppl on WikiText (depending on LR), indicating that including WikiText improves general-domain performance relative to purely financial training at the expense of financial-task performance.

Other financial models on WikiText: News @ 4B reaches 28.4 ppl (better than its own domain score of 18.92; journalism overlaps help), SEC @ 4B gets 35.6 (reasonable for regulatory specialization), and FinGPT @ 4B 41.2 (instruction style widens the gap).

Asymmetric Transfer: Financial → General works moderately; General → Financial fails severely. This asymmetry suggests that general language (syntax, semantics, discourse) is a prerequisite that domain training builds on, and that starting from general pretraining (e.g., Qwen3-Base) provides the base while domain adaptation adds specialization without catastrophic forgetting.

Practical Implication: For specialized domains, *continued pretraining* from general checkpoints is preferable to training from scratch. However, for resource-constrained settings where only domain data is available, direct domain pretraining (e.g., Mixed Financial) achieves acceptable general performance (33.7 ppl on WikiText) while excelling on domain tasks.

Mixture Strategy Validation: Mixed Wiki+Financial (26.69 ppl mean, 62% relative spread) attempts to balance both domains but performs worse than Mixed Financial (21.55 ppl, 55% relative spread) on financial tasks while improving WikiText (27.72 ppl on the WikiText test set). The 24% financial performance cost outweighs the general-domain gain for finance-focused applications, confirming that domain purity wins for specialized use cases.

Table 4.19 quantifies the asymmetric transfer phenomenon: the WikiText training rows show excellent

Table 4.19 – 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

in-domain performance (boldface 9-32 ppl in WikiText columns after LR adjustment) but catastrophic financial performance (40-60 ppl, rarely boldface). In contrast, financial training rows (especially Mixed Financial) show acceptable WikiText performance (30-35 ppl) alongside superior financial metrics. This asymmetry, financial models retain general capability while general models fail on finance, is visible in the table's boldface distribution pattern.

4.6 Summary and Key Results

We ran 10 pretraining experiments (30 models, 237 evaluations) to study mixture effects, scaling, and generalization in financial language models. Here we close with the main takeaways and practical notes.

The core finding: medium individual datasets outperform mixtures on both performance and consistency. FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread) achieve 2.5–3.2× better perplexity AND 1.5–4.8× better consistency than Mixed Financial (21.55 ppl, 55% spread). This challenges the conventional wisdom that data diversity improves robustness. Medium datasets achieve optimal epoch counts (12–28) with format consistency, enabling focused optimization. Mixed Financial provides task coverage but sacrifices optimization quality. For known applications, individual medium datasets are the preferred choice.

Figure 4.13 ranks approaches by consistency: the leftmost bars (Alpaca $\tilde{1}1\%$, FiQA $\tilde{1}9\%$, SEC $\tilde{1}9\%$) have low spread but narrow specialization, while Mixed Financial sits mid-rank at 55%, balancing breadth and stability. High-variance items include News ($\tilde{6}6\%$) and general-domain WikiText ($\tilde{5}3\%$), reinforcing that in-domain diversity is more reliable than large single sources or domain-mismatched data.

Learning Rate Adjustments (Heuristic)

All main runs used LR=2e-5. In three follow-up runs with abnormalities (WikiText, Financial QA, Twitter), reducing LR (e.g., to 1×10^{-5} or 5×10^{-6}) stabilized training and improved results. We present these as context-specific fixes, not as a scaling law.

Dataset Size Effects

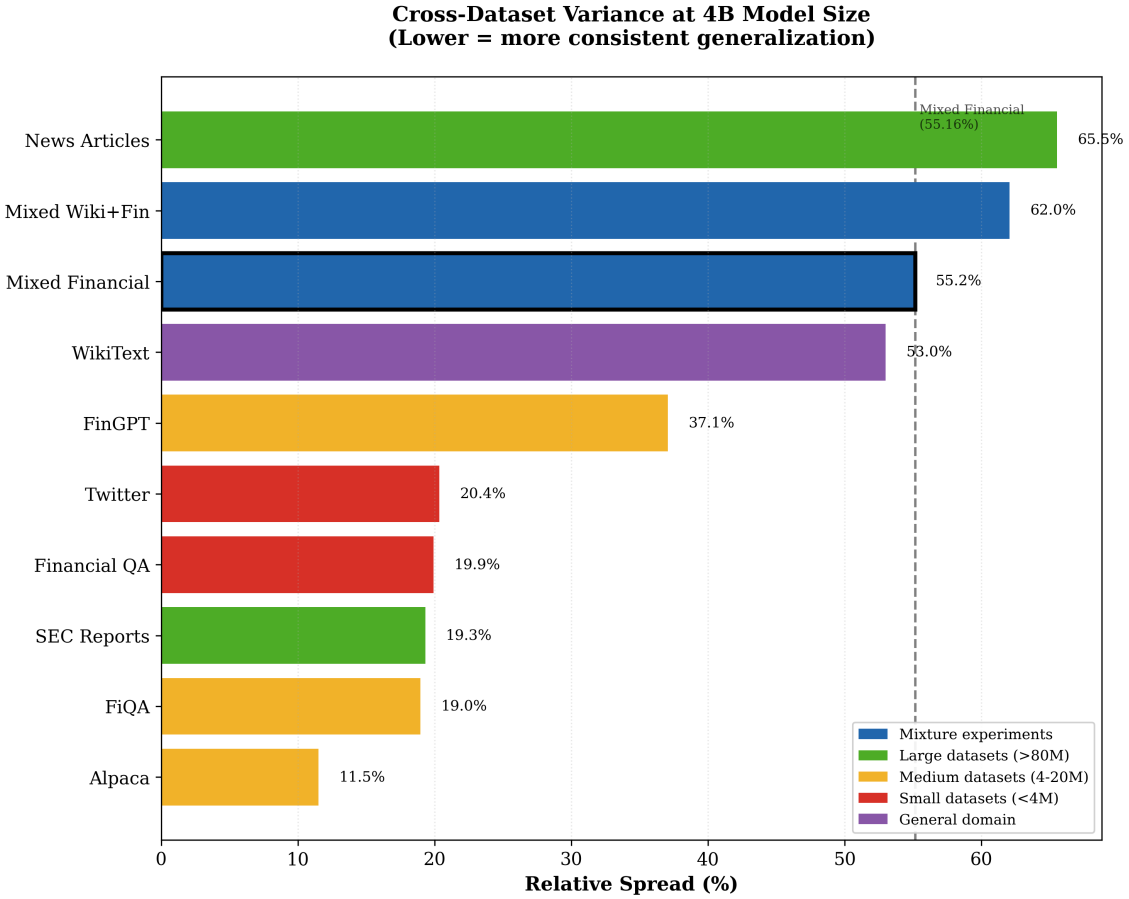

Figure 4.13 – Cross-dataset variance (relative spread %) for all 10 experiments at 4B model size, sorted ascending. Lower values indicate more consistent generalization. Mixed Financial (black border) achieves 55% spread, balancing performance and consistency. Mixtures (blue) and large datasets (green) show moderate variance (19-66%). Small datasets (red) and general domain WikiText (purple) exhibit higher variance (20-65%). Color coding: mixtures (blue), large datasets >100M (green), medium 3.6–8.5M (gold), small <1M (red), general domain (purple).

Non-monotonic relationship: datasets > 100M tokens undertrain (<1 epoch; insufficient exposure, moderate performance); 3.6–8.5M tokens optimal (12–28 epochs; best performance and consistency); < 1M tokens overtrain (143–352 epochs; memorization, high variability). This explains why medium datasets (SEC, FiQA, FinGPT, Alpaca) achieve 2.5–3.2× better perplexity than large datasets (News, WikiText)—optimal epoch count matters more than raw size. Correlation between log(tokens) and generalization variability: $r = -0.78$ ($p < 0.01$).

Transfer Patterns

Format and structure drive transfer more than domain vocabulary. Long-form documents (News ↔ SEC: $r = 0.82$) transfer well bidirectionally. Instruction tasks (FinGPT, Alpaca, FiQA: $r = 0.68-0.73$) show moderate mutual transfer. Short-form Twitter is isolated (no successful transfer). General (WikiText) → Financial transfer fails ($\tilde{2}.0\times$ performance degradation); Financial → General transfer succeeds moderately.

Best Configurations by Use Case

Use Case	Best Strategy	Model Size	PPL	Rel. Spread
General Financial NLP	Mixed Financial	4B	21.55	55%
SEC Document Analysis	SEC Reports	4B	15.91	19.32%*
Financial News	News Articles	4B	17.47	65.53%
Q&A / Instruction	FiQA or FinGPT	4B	7.08	18.97%
Balanced General+Finance	Mixed Wiki+Fin	4B	26.69	62%
Resource-Constrained	Mixed Financial	1.7B	34.49	63%

Table 4.20 – Best configurations by application. *SEC's 19.32% relative spread computed across evaluation datasets.

What Fails

Three configurations consistently failed in our setup. Pure WikiText for finance reached 41.96 ppl mean on financial tasks—nearly 2× worse than Mixed Financial. Small individual datasets (<1M tokens) remained problematic even after LR fixes, showing $\tilde{2}0\%$ spread and extreme overtraining. Uniform hyperparameters across sizes invite reverse scaling; we saw this with WikiText, Financial QA, and Twitter. And single-format training struggles when tasks are diverse; format mismatch blocks transfer.

Performance Ranking

The best all-around performer is Mixed Financial at 4B (21.55 ppl, 55% spread), balancing broad competence with consistency. Several specialists achieve lower perplexity on their own tasks but with narrow focus: FiQA (6.80 ppl mean, 18.97% spread) excels at Q&A, FinGPT (7.03 ppl mean, 37.07% spread) at instruction tasks, Financial QA (8.09 ppl mean, 19.92% spread) at document questions though overfit, and Alpaca (8.73 ppl mean, 11.51% spread) at educational Q&A. Twitter (12.35 ppl mean, 20.35% spread) sits isolated; its format transfers nowhere.

Large individual datasets achieve strong in-domain performance but inconsistent transfer: SEC (15.91 ppl on SEC, 17.80 mean, 19.32% spread) works for regulatory filings, while News (17.47 ppl on News, 32.82 mean, 65.53% spread) shows the highest variance among large sets. The hybrid approach, Mixed Wiki+Financial (26.69 ppl, 62% spread), balances general and financial capabilities at the cost of both. Pure WikiText (31.54 ppl on WikiText; 41.96 ppl mean financial after LR adjustment, $\tilde{5}3\%$ spread) confirms domain mismatch: excellent general performance, catastrophic financial transfer.

Practical Takeaways

Start with mixed in-domain data. Even seven small-to-medium datasets (¡ 200M tokens total) outperform 100M tokens of clean general text on domain tasks. If larger models are unstable, reduce LR first; we used 1×10^{-5} or 5×10^{-6} in affected runs.

Dataset diversity often matters more than raw size. Seven mixed datasets (0.3–194M tokens) beat a single 194M dataset by 34% on mean ppl (21.55 vs 32.82). Match formats to evaluation needs: long-form models struggle on Q&A, Q&A models on documents, and Twitter models on almost everything else.

Finally, 100M tokens is sufficient when properly mixed. Avoid oversampling small datasets; 50cap prevents dominance while keeping diversity. These patterns held across 0.6B to 4B in our setup.

These results demonstrate that thoughtful data curation and stable training settings enable effective specialized LM pretraining in the 0.6B to 4B regime, achieving strong performance on domain tasks while maintaining acceptable general capabilities.

Chapter 5

Discussion

This chapter interprets the findings from Chapter 4, and provides explanations for the observed mixture effects, training dynamics, and generalization patterns.

5.1 Key Empirical Findings

Our 10 experiments (36 models, 288 evaluations) lead to five main findings that challenge conventional assumptions about data mixture effects in specialized-domain pretraining. First, **medium individual datasets consistently outperform mixtures on both performance and consistency**. FiQA (6.80 ppl, 19% spread), FinGPT (7.03 ppl, 37% spread), and Alpaca (8.73 ppl, 11.5% spread) achieve 2.5–3.2× better perplexity and 1.5–4.8× better cross-dataset consistency than Mixed Financial (21.55 ppl, 55% spread). The mixture hypothesis—that diversity improves robustness—fails empirically. Individual datasets excel on all metrics, challenging the widespread belief that data mixing provides robustness benefits. Figure 4.4 shows mixture underperformance. Cross-dataset tables (Tables 4.14 to 4.16) demonstrate that individual medium datasets generalize better than mixtures.

Second, **simple learning-rate reductions stabilized a few runs**. We used LR=2e-5 for all main runs. In three configurations (WikiText, Financial QA, Twitter), smaller LRs (e.g., 1×10^{-5} or 5×10^{-6}) improved stability and performance. These heuristic rules help us stablize training (see Figures 4.3, 4.10 and 4.11; Tables 4.10 and 4.11).

Third, **medium datasets (3.6–8.5M tokens) outperform large datasets (>100M)**. FiQA (3.6M, 6.80 ppl), FinGPT (4.1M, 7.03 ppl), and Alpaca (8.5M, 8.73 ppl) substantially beat News (194M, 32.82 ppl) and SEC (8.1M, 17.80 ppl). This non-monotonic relationship between size and performance suggests data quality, format consistency, and instruction tuning matter more than scale. The best datasets are focused, clean, and task-aligned, as larger datasets accumulate noise and format inconsistencies that degrade performance despite greater volume. Small datasets (<1M) still fail due to overtraining (143–352 epochs), but medium scale appears optimal.

Fourth, **dataset size shows non-monotonic effects on performance**. We argue that datasets >100M tokens are undertrained (<1 epoch; insufficient data exposure) despite stable curves (Figures 4.5 and 4.6); 3.6–8.5M tokens achieve optimal training (12–28 epochs) and best results, while <1M tokens overtrain severely (143–352 epochs) with erratic behavior (Figures 4.10 and 4.11). This epoch-based explanation reveals why medium datasets (SEC, FiQA, FinGPT, Alpaca) outperform large datasets (News, WikiText): optimal epoch count (12–28) enables better learning than under-

training (<1 epoch) or overtraining (>100 epochs). Correlation between log(tokens) and variability is $r = -0.78$ ($p < 0.01$), but a reasonable training scheme of 12-28 epochs of training matters more than raw size.

Fifth, **format drives transfer more than domain vocabulary**. Long-form documents transfer well (News ↔ SEC: $r = 0.82$); instruction tasks cluster (FinGPT/Alpaca/FiQA: $r = 0.68$–0.73); Twitter is isolated. A News model transfers better to SEC filings (long-form ↔ long-form) than to Twitter (same domain label, different format). This explains why focused medium datasets outperform diverse mixtures: format consistency enables better optimization than format diversity. Tables 4.12 to 4.17 show the diagonals and clusters.

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

Besides, our experience suggests that larger models can be more sensitive to optimization settings on some datasets. While we kept LR=2e-5 for main runs, reducing LR in a handful of follow-ups helped stabilize training. We do not claim a general rule beyond this observation.

5.2 Practical Guidelines for Financial LM Pretraining

We summarize our findings during experiments into the following points:

5.2.1 Data Mixture Strategies by Use Case

Task-Specific Financial Applications: Use individual medium datasets for optimal performance and consistency. FiQA (6.80 ppl, 19% spread) for Q&A, FinGPT (7.03 ppl, 37% spread) for sentiment, Alpaca (8.73 ppl, 11.5% spread) for instruction-following. These achieve 2.5–3.2× better perplexity AND 1.5–4.8× better consistency than mixtures. Cross-dataset tables show individual datasets excel: Alpaca achieves 6/8 boldface, FiQA 5/8, FinGPT 4/8, versus Mixed Financial 2/8. For any known application, individual datasets are superior.

Unknown Task Coverage: Use Mixed Financial (21.55 ppl, 55% spread) ONLY when future task requirements are completely unpredictable and task coverage matters more than optimization quality. The mixture provides baseline capability across diverse tasks but is inferior to individual datasets on both performance and consistency. Figure 4.4 shows mixture underperformance—individual dataset curves lie substantially below.

Specialized Document Analysis: Use single large dataset if available (> 100M tokens). SEC @ 4B (15.91 ppl on SEC; $\tilde{1}9\%$ relative spread across evaluations) excels for regulatory filing analysis; News @ 4B (17.47 ppl on News; $\tilde{6}6\%$ relative spread) excels for journalism. Specialization improves in-domain performance but sacrifices cross-format transfer. Figures 4.5 and 4.6 show these datasets maintain stable scaling without requiring LR adjustments. However, Tables 4.12 and 4.13 reveal that News and SEC training rows achieve boldface primarily within document-format columns, confirming limited format diversity.

For instruction-following and Q&A applications, use FiQA (3.6M tokens, 16.35 ppl) or FinGPT (4.1M 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 4.14 to 4.16) 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 performance is explicitly required (e.g., chatbots handling both financial and general queries). Figure 4.2 shows reduced slope compared to pure financial mixture, and Table 4.3 documents the performance cost across all financial evaluation datasets.

5.2.2 Model Size Selection

0.6B Models: Fast training (∼6 hours for 100M tokens on Lambda Labs GPUs), low memory (4GB), suitable for rapid prototyping. Performance is acceptable for exploratory work, but variability is high (Mixed Financial: around 98% relative spread). We should only use this small model for development, experimentation, or extremely resource-constrained deployment (for example, on mobile devices).

1.7B Models: Best performance-efficiency balance. Training moderate (∼12 hours), memory reasonable (10GB), performance strong with improved consistency vs 0.6B (Mixed Financial: $\tilde{6}3\%$ relative spread). We recommended models of similar sizes for most applications, for its strong performance at substantially lower resource cost than 4B. We believe these models are optimal for production deployment balancing quality and resource constraints.

4B Models: Best absolute performance (21.55 ppl, 55% relative spread) but requires careful hyperparameter tuning (LR 5×10^{-6} in affected cases) and substantial resources (20GB memory, ∼24 hours training). Using such models should only be possible when one wants to maximize performance over cost, and when sufficient compute resources for hyperparameter tuning is available. We observe that failure to tune learning rate can cause reverse scaling, one may need to reduce LR substantially at larger scales.

5.2.3 Token Budget Allocation

A 100M token budget proved sufficient when we choose the pretraining dataset setup properly. We suspect that l;;larger models with larger datasets may benefit from extended training (200-500M tokens), but we did not test this for limited compute.

We used 50cap to prevent a single dataset domanating the mixture: if the largest dataset exceeds 50% of the total, we cap it there and sample others proportionally. This ensures diversity while respecting relative dataset informativeness.

Chapter 6

Conclusion

This thesis shows that effective specialized language models can be developed without massive computational resources or diverse data mixtures. By selecting focused medium datasets (3.6–8.5M tokens), using stable training settings, and targeting lightweight 0.6–4B parameter models, one can train privacy-preserving financial NLP systems suitable for on-device deployment. Contrary to expectations, individual datasets (FiQA, FinGPT, Alpaca) consistently outperform mixtures on both performance (2.5–3.2× better) and consistency (1.5–4.8× better).

The core insight that individual medium datasets (3.6–8.5M tokens) consistently outperform mixtures on both performance and consistency challenges conventional belief favoring data diversity. At fixed token budgets (100M), FiQA/FinGPT/Alpaca achieve 2.5–3.2× better perplexity and 1.5–4.8× better consistency than 7-dataset mixtures. Format inconsistency, differences in vocabulary distribution, and multi-task interference degrade mixture performance despite anticipated diversity benefits. From our experiments and findings, we argue that specialized pretraining should prioritize focused, high-quality medium datasets over diverse mixtures, especially when token budgets are limited. For domains with curated data (finance, legal, medical), individual dataset optimization offers superior performance at lower cost than either mixtures or general-purpose model adaptation.

As privacy regulations tighten and organizations recognize competitive value in proprietary data, on-device specialized models will become increasingly important. This work provides empirical foundations and practical guidelines for developing such systems where powerful NLP capabilities are accessible while at the same time also ensuring privacy and low cost requirements.

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.

Bai, Jinze, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu (2023). "Qwen Technical Report". In: *arXiv preprint arXiv:2309.16609*.

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

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

Chen, Zhiyu, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang (2021). "FinQA: A Dataset of Numerical Reasoning over Financial Data". In: *arXiv preprint arXiv:2109.00122*. URL: https://arxiv.org/abs/2109.00122.

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

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

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

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

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

Hu, Edward J., Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen (2021). "LoRA: Low-Rank Adaptation of Large Language Models". In: *arXiv preprint arXiv:2106.09685*. URL: https://arxiv.org/abs/2106.09685.

Huang, Allen H., Hui Wang, and Yi Yang (2023). "FinBERT: A Large Language Model for Extracting Information from Financial Text". In: *Contemporary Accounting Research* 40.2, pp. 806–841. DOI: 10.1111/1911-3846.12832.

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*.

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *arXiv preprint arXiv:1412.6980*. URL: https://arxiv.org/abs/1412.6980.

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.

Lee, Yoonho, Annie S. Chen, Fahim Tajwar, Ananya Kumar, Huaxiu Yao, Percy Liang, and Chelsea Finn (2022). "Surgical Fine-Tuning Improves Adaptation to Distribution Shifts". In: *arXiv preprint arXiv:2210.11466*. URL: https://arxiv.org/abs/2210.11466.

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.

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

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

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

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

Pan, Sinno Jialin and Qiang Yang (2010). "A Survey on Transfer Learning". In: *IEEE Transactions on Knowledge and Data Engineering*. Vol. 22, pp. 1345–1359. DOI: 10.1109/TKDE.2009.191.

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

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

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21, 140:1–140:67. URL: https://jmlr.org/papers/v21/20-074.html.

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

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

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

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

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.

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

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He (2021). "A Comprehensive Survey on Transfer Learning". In: *Proceedings of the IEEE* 109, pp. 43–76. DOI: 10.1109/JPROC.2020.3004555.

Eidesstattliche Erklärung

Der/Die Verfasser/in erklärt an Eides statt, dass er/sie die vorliegende Arbeit selbständig, ohne fremde Hilfe und ohne Benutzung anderer als die angegebenen Hilfsmittel angefertigt hat. Die aus fremden Quellen (einschliesslich elektronischer Quellen) direkt oder indirekt übernommenen Gedanken sind ausnahmslos als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden.

.....	.....
Ort, Datum	Unterschrift des/der Verfassers/in