Research Proposal: Identifying Scaling Laws in Financial Forecasting

An Application of Machine Learning Metascience to Empirical Finance

Jiyan Schneider (82441028)*

July 20, 2025

Research Question and Motivation

Research Question

Recent work in Natural Language processing shows that model performance obeys scaling laws, meaning that the loss of predictions falls as model parameters and training samples increase (Kaplan et al., 2020; Hoffmann et al., 2022). Should analogous laws exist in asset-price forcasting, this would let practicioners quantify the marginal value of data versus compute. Furthermore, it would offer a new metric for market predictability. I propose an empirical study that (i) trains a grid of scalable neural networks on equity data, (ii) estimates the exponents that link loss to parameters and observations, and (iii) converts accuracy gains into economic value via a simple trading strategy. The deliverables of this study include fitted scaling law equations, guidance on optimal resource allocation for quantitative finance models.

While the scaling hypothesis has had large impacts in the fields like natural language processing and computer vision, its applicability to empirical finance remains largely

*jiyan.schneider@keio.jp

unexplored. This proposal seeks to bridge that gap by addressing the following primary research question:

Can the performance of financial forecasting models be described by predictable scaling laws, similar to those observed in large language models?

This research will test the scaling hypothesis in the field of finance. Specifically, that the prediction error of a neural network model trained on a financial forecasting task scales as a power-law function of model size (number of parameters, N) and data size (number of training examples, D).

Motivation and Importance

The potential discovery of scaling laws in finance carries significant and compelling implications for both industry practitioners and academic researchers.

Practical and Real-World Implications: The process of developing quantitative trading models is notoriously resource-intensive and characterized by costly trial-and-error. A validated scaling law would provide a quantitative framework to optimize this process. It would allow firms to answer critical questions about resource allocation: Should we invest in acquiring more historical data, or in training a larger, more complex model? By quantifying the expected "return on investment" for both data and compute, a scaling law could guide strategic decisions, potentially saving millions in development costs. It would also allow us to make judgements about wether we should be investing any money into Machine Learning in the first place. It would offer a principled way to find the financial equivalent of "Chinchilla" (Hoffmann et al., 2022), a model that achieves superior performance by optimally balancing model size and training data.

Theoretical Implications: From a theoretical standpoint, scaling laws offer a novel framework for understanding the limits of predictability in financial markets. The canonical scaling law equation, $L(N, D) \approx E_{\text{irred}} + A/N^{\alpha} + B/D^{\beta}$, includes a term for irreducible error, E_{irred}. In a financial context, this term could be interpreted as a quantitative measure of fundamental market efficiency or intrinsic randomness, the component of market

movements that cannot be predicted regardless of model or data scale. Estimating this term would be a significant contribution to the literature on market efficiency.

Originality and Contribution: This research is original in its direct and systematic application of the scaling law methodology, born from machine learning metascience, to a core problem in empirical finance. While machine learning models are now common in finance, the literature lacks a rigorous investigation into *how* their performance scales with resources. This study would address that gap, providing a new lens through which to view model development and market predictability.

Literature Review and Background

The seminal work finding the existence of these scaling laws in Artificial intelligence is that of Kaplan et al. (2020), who first comprehensively demonstrated that the performance of neural language models, as measured by cross-entropy loss, scales smoothly as a power-law with model size, dataset size, and training compute. Their findings spanned over seven orders of magnitude and suggested that for optimal performance, one should train the largest model affordable, as larger models are significantly more sample-efficient.

This conclusion was refined by Hoffmann et al. (2022). In their paper on "compute-optimal" training, they conducted a large-scale study and found that Kaplan et al. had under-emphasized the importance of data. Their key finding was that for a fixed computational budget, model size and dataset size should be scaled in roughly equal proportion. To prove this, they trained their "Chinchilla" model (70B parameters) on four times more data than the competing Gopher model (280B parameters). Despite being four times smaller, Chinchilla significantly outperformed Gopher, demonstrating that the prevailing wisdom of "bigger is always better" was incomplete. Most recently, these findings have been replicated and extended, for instance by DeepSeek-AI et al. (2024), confirming that this IsoFLOP methodology is a robust tool for finding optimal model-data trade-offs and using scaling laws to inform decisions about model architecture.

This paradigm has been successfully applied beyond text. Zhai et al. (2022) applied

a similar analysis to computer vision, successfully training a 2-billion parameter Vision Transformer to a new state-of-the-art on ImageNet. Furthermore, in the field of Speech Language Models (SLMs), it has been found that while scaling laws do seem to exist, the scaling exponents are quite low (Cuervo & Marxer, 2024), discouraging the creation of large models, and instead leading to researchers pivoting into other directions.(Maimon et al., 2025). The scaling framework has become a standard tool for guiding research and investment in AI.

This proposal situates itself directly in this line of inquiry, aiming to perform the first comprehensive "Chinchilla-style" analysis for a financial task. The existing finance literature extensively uses predictive models, but research typically focuses on feature engineering or novel architectures for a fixed data/compute budget. This work addresses a different, more fundamental question: how does predictive power itself behave as we scale the fundamental resources of the learning process?

Proposed Empirical Strategy

To answer our research question, we will adopt the "IsoFLOP" methodology pioneered by Hoffmann et al. (2022). This approach is designed to disentangle the effects of model size and data size while holding the total computational cost of training constant.

Data and Prediction Task

We will focus on a high-frequency forecasting task, which generates a large number of training samples from a given historical period. We want to focus on a task where lots of data is available so that, should scaling laws exists, there is room so that we can scale up. The proposed task is to predict the direction (and possibly magnitude) of the mid-price change of a financial instrument over a short future horizon (e.g., 10 seconds) based on market data from the immediate past (e.g., the previous 60 seconds). This is a well-defined, tractable, and challenging problem. The selection of a specific dataset is contingent on data availability and access through university resources. Several options

will be considered, each with distinct advantages and disadvantages:

Trade and Quote (TAQ) Data via WRDS: This is the canonical source for high-frequency academic research in the US. It contains every trade and quote for all US-listed securities. This dataset has extremely broad coverage (thousands of stocks over decades), and a well-understood data structure. However, it can be pretty noisy and data cleaning is required.

LOBSTER (Limit Order Book System): This dataset provides high-fidelity, message-by-message reconstructions of the limit order book for a set of NASDAQ-traded stocks. This dataset provides the most granular view of market dynamics, but its coverage is more limited (less securities, shorter time-period) than TAQ.

Cryptocurrency Exchange Data: Public APIs or data vendors (e.g., Kaiko) provide high-frequency data for crypto assets. This data is often freely available 24/7 thus there is no data fragmentation, however, different microstructure and regulatory environment may make findings less generalizable to traditional equities.

The final choice will be based on availability, and preferably a pilot study to further assess data quality and processing requirements against available computational resources. Regardless of the source, careful chronological splitting of data into training, validation, and out-of-sample test sets will be paramount to prevent look-ahead bias.

Model Architecture

We will use a standard Transformer-based neural network architecture. There are two reasons for the selection of this architecture. Firstly, the Transformer (Vaswani et al., 2017) is an ideal candidate for this study because its size is easily and systematically scalable by adjusting two key hyperparameters: the model's depth (number of layers) and its width (the hidden dimension size). Secondly in recent years it has become one of the most widely studied architectures in ML in general as well as in Finance in particular.

Experimental Design: The IsoFLOP approach

The core of the empirical strategy is as follows:

1. Define Compute Budgets: We will select several fixed computational budgets (measured in total floating-point operations, or FLOPs), for example, $C_1 = 10^{18}$, $C_2 = 10^{19}$, and $C_3 = 10^{20}$ FLOPs. Each budget defines an "IsoFLOP curve."

2. Train a Family of Models: For each fixed compute budget C_i, we will train a series of models. The models will vary in size (N) and will be trained on a corresponding amount of data (D) such that the total training FLOPs (proportional to $N \times D$) remains constant. For example, a model with half the parameters will be trained on twice the data.

3. Record Performance: For each trained model, we will record its final prediction loss on a held-out validation set.

4. Fit the Scaling Law: By plotting the validation loss against model size (N) for each IsoFLOP curve, we expect to see a U-shaped relationship. The minimum of this U-shape represents the optimal model size for that compute budget. By analyzing the relationship between the optimal model size and optimal data size across different compute budgets, we can estimate the exponents α and β in the scaling law equation.

Expected Contributions and Evaluation

The success of the proposal will be evaluated on two fronts: **Statistical Evaluation:** The primary evaluation will be the goodness-of-fit of the power-law model to the observed data. We will measure the final cross-entropy loss of our models and test how well the log-log linear relationship ($\log(L - E_{\text{irred}})$ vs. $\log(N)$ and $\log(D)$) holds. A high R^2 in this regression would provide strong evidence for the scaling hypothesis. **Economic Evaluation:** As a secondary, more practical measure, we will construct a simple, signal-driven trading strategy based on the predictions of the best-performing models from our experiment. We will evaluate its out-of-sample performance, for instance by calculating its Sharpe ratio. This will help ground the statistical loss metric in tangible economic

value.

This research is expected to make several contributions. Primarily, it would establish whether the powerful scaling law paradigm is applicable to financial forecasting. If successful, the estimated exponents α and β would provide the first quantitative guidance on the relative importance of model size versus data size in finance. Furthermore, the estimate of the irreducible error E_{irred} would provide a novel, model-based estimate of the level of inherent randomness in high-frequency markets.

Feasibility and Limitations

This research is ambitious but feasible as a semester-long project, particularly as a pilot study. While training dozens of models is computationally expensive, the experiment can be scaled down to a manageable size. The core relationships of scaling laws have been shown to hold even at smaller scales. The project can be conducted using university high-performance computing (HPC) resources. The proposed models (standard Transformers) and data formats are well-supported by open-source deep learning libraries like PyTorch or JAX.

We acknowledge several key challenges:

1. Computational Cost: A full-scale analysis matching that of Hoffmann et al. is beyond the scope of a single student project. The plan is to demonstrate the methodology on a smaller but still significant scale.

2. Data Non-stationarity: Financial markets evolve over time and this non-stationarity could complicate the scaling relationship. Our experimental design will include robust out-of-sample testing across different time periods to assess the stability of any identified laws.

3. Generalizability: The findings may be specific to the chosen prediction task, asset class, and model architecture. Future work would be needed to test the generality of these laws across different tasks (e.g., volatility forecasting), markets (e.g., futures, options), and models (e.g., LSTMs).

References

Cuervo, S., & Marxer, R. (2024). Scaling Properties of Speech Language Models. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 351–361. https://doi.org/10.18653/v1/2024.emnlp-main.21

DeepSeek-AI, Bi, X., Chen, D., Chen, G., Chen, S., Dai, D., Deng, C., Ding, H., Dong, K., Du, Q., Fu, Z., Gao, H., Gao, K., Gao, W., Ge, R., Guan, K., Guo, D., Guo, J., Hao, G., … Zou, Y. (2024, January 5). *DeepSeek LLM: Scaling Open-Source Language Models with Longtermism*. arXiv: 2401.02954 [cs]. https://doi.org/10.48550/arXiv.2401.02954

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D. d. L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., … Sifre, L. (2022, March 29). *Training Compute-Optimal Large Language Models*. arXiv: 2203.15556 [cs]. https://doi.org/10.48550/arXiv.2203.15556

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020, January 23). *Scaling Laws for Neural Language Models* [Comment: 19 pages, 15 figures]. arXiv: 2001.08361 [cs]. https://doi.org/10.48550/arXiv.2001.08361

Maimon, G., Hassid, M., Roth, A., & Adi, Y. (2025, April 3). *Scaling Analysis of Interleaved Speech-Text Language Models*. arXiv: 2504.02398 [cs]. https://doi.org/10.48550/arXiv.2504.02398

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need [_eprint: 1706.03762]. *CoRR*, *abs/1706.03762*. http://arxiv.org/abs/1706.03762

Zhai, X., Kolesnikov, A., Houlsby, N., & Beyer, L. (2022). Scaling Vision Transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1204–1213. https://doi.org/10.1109/cvpr52688.2022.01179