On the Use of ML and AI in the Investment Process

A White Paper by Arctan Research

Executive Summary

> *Arctan is a disciplined, theory-anchored quantitative equity and asset allocation family of strategies that uses machine learning as a flexible function approximator over a small set of economically motivated features, wrapped in a regime-conditional risk overlay — not as a black-box prediction engine.*

At Arctan Research we identify three main challenges for the formulation and use of ML in quantitative finance:

- **Too many factors:** There is a very large, and probably regime-specific, set of factors that determine asset prices and returns.
- **Too little data:** There is only one "timeline" available for analysis: actual economic history. Methods such as cross-validation or Monte Carlo simulations are not very useful in finance since we do not have a coherent model of the underlying data-generating processes.
- **Non-stationarity:** Even if the relevant factors have been identified, the quantitative relationships between factors and returns are likely to suffer from both slow drifting changes and more rapid, episodic breaks.

These challenges cannot be dissolved — they can only be managed through calibrated design choices. Arctan's architectural responses are threefold:

- **Curated features:** We heavily pre-process financial and economic signals using domain expertise and decades of empirical experience, giving ML a constrained, interpretable space to learn in.
- **Regime conditionality:** We partition the timeline into economic regimes, apply regime-dependent feature weighting and portfolio construction, and use variable leverage to sharply curtail risk at dangerous regime boundaries.
- **Continuous monitoring:** We run a battery of statistical signal-quality tests at every rebalancing to detect degradation and trigger re-estimation or risk reduction before damage compounds.

We Live in a Revolutionary Age

The possibilities opened by the accelerating development of artificial intelligence are already reshaping every field of human endeavor. But we do not yet have artificial general intelligence (AGI). The existing technology is both very good at certain tasks and very limited — if not outright counterproductive — in others. Until we have full AGI, any application of Machine Learning (ML) and AI must be tailored to the specific domain.

This is a crucial point, because the invention of "chat bots" that can discuss anything with apparent fluency might seem to imply that **domain expertise** — the deep understanding an expert has of their field — is no longer valuable. In general, and for the investment process in particular, this is obviously untrue, which is why one cannot "vibe code" one's way to being a millionaire trader, despite the many claims to the contrary on X or YouTube.

The nature of financial markets makes this point in stark terms. The success of ML in various fields — and we should recall that Large Language Models are but one instantiation of neural networks, a technology whose conceptual roots trace to the 1940s and 1950s[1] but whose modern power derives from the transformer architecture introduced in 2017[2] — has arisen from a combination of massive increases in computational power applied to large datasets. You need both: compute and data.[3]

The Three Challenges of ML in Finance

But what is "the data" in financial markets? First and foremost: *there is simultaneously too much data and too little.*

There is *too much* because there does not exist a fixed set of factors that determine asset prices. They change all the time and depending on the moment. Current theory and empirical models can map what *should* matter, but that does not constrain what *actually* matters. Our state-of-the-art asset pricing models can be expressed in terms of a stochastic pricing kernel, but in practice exactly what determines this factor is potentially limitless — the academic literature has catalogued over 300 published return-predicting signals.[4]

There is also *too little data.* To think systematically about markets, we have only one historical "timeline" — what actually happened. The impossibility of running experiments (of the type: "if Lehman Brothers had been bailed out, would we have had a crisis in 2008?") severely constrains our ability to map factors to outcomes. We cannot generate synthetic timelines because we lack a coherent underlying model of the real world.

And insofar as we can find empirical regularities in the data — say, that higher profitability leads to rising equity values — the exact quantitative relationship between these variables changes over time. Financial markets are *non-stationary* in the statistical sense, in part because the underlying economy evolves and we are dealing with highly adaptive human beings.

So, in summary, we have a domain where:

- *We cannot say with any certainty which factors will determine future returns;*
- *We are severely constrained by the impossibility of running alternative timelines;*
- *Relationships between variables shift over time — pervasive non-stationarity.*

Too much data, too little data… Should we despair? Of course not. Each challenge can be met with **the correct application of the right tools, guided by design choices made in the light of domain experience.** This is what Arctan Research does.

Challenge 1: What Are the Relevant Factors?

ML is fundamentally about learning functional forms. Given a set of inputs (features) and outcomes, the model learns the best mapping between them and can then predict out-of-sample what should happen given a specific set of feature values.

If there is enough data, ML can encode both which variables matter and what the relationship between them and the outcome looks like — even in the presence of substantial noise. But in a data-constrained domain with genuine uncertainty about what matters, and with only one experimental timeline, ML will struggle to both identify relevant features and find reliable functional forms. This is why attempts to feed raw financial data directly to deep learning models have not been consistently successful at monthly and quarterly horizons.

> *The canonical academic reference is Gu, Kelly & Xiu (2020),* whose large-scale study of ML methods in asset pricing found that deep learning outperforms classical approaches only when applied to hand-engineered characteristics — not to raw price data. That finding is precisely the empirical anchor for Arctan's curated-feature design philosophy.[5]

Every systematic investment process begins with uncovering a statistically significant investment signal. This is what we call the "analyst function."

How is this done in practice? First, draw on theoretical and empirical priors about what should matter. Second, pre-process the data so that it generates a useful signal — this may require heavy regularization and careful cross-validation that respects the time structure of the data. Both processes require deep domain expertise: decades of real market experience and deep knowledge of the relevant economic and financial literature.

Our equity models define a feature set of approximately 25 dimensions spanning fundamental company data, sensitivity to macroeconomic shocks, and statistical features of equity returns. Each feature is chosen for a reason and, where necessary, pre-processed to separate signal from noise.

We keep our feature set deliberately *small and interpretable.* We believe the true signal for equity returns lies in a low-dimensional but complex space. These signals are then used by *a flexible non-linear function approximator over the features.* In the challenge of "choosing the data and discovering the functional form," we choose the data using

domain knowledge so the ML can find the best functional form mapping features into returns.

Challenge 2: Overcoming the One-Timeline Limitation

In financial markets we cannot run experiments. We cannot even run good simulations because we lack a coherent underlying model of the real world.

But while we cannot, without access to the multiverse, generate alternative timelines, we can take the existing timeline and **define different regimes.** What our work shows is that the features that matter, the data that is relevant, and the way that data maps into future returns, all change depending on the financial and macroeconomic regime. Nothing works unconditionally.

Much of the quantitative literature treats factor selection as stationary: for example, "quality" firms with consistent profitability should earn a positive risk premium. What we have found is that this is NOT the case: the quality premium exists, but only in certain regimes.

The Arctan models **switch which latent direction of the cross-section they follow based on the regime.** Portfolio construction is regime-based, not a function of a once-and-for-all mapping between features and outcomes.

As an example, there is a clear distinction in the Brazilian equity market between "quality" — high-ROE stable earners such as large-cap banks — and higher-volatility, commodity and consumer-based companies. Arctan uses the feature space and regime signals to smoothly tilt the portfolio toward the companies with the highest risk-adjusted expected returns given its reading of the current regime.

The determination of what the relevant regimes are is also highly curated by domain expertise. Our work shows that **each equity market has a very specific regime structure** based on its economic, institutional, and political structure and how it fits into the broader global economy. There is no single "simple" variable that effectively determines the relevant regime.

In Arctan, the specification of the regime shapes not only portfolio construction but also gross exposure. All our equity models incorporate **a variable leverage feature that manages gross exposure depending on the regime and the risk of regime transitions.** By determining in real time the current regime and monitoring transition risk — especially transitions to riskier, negative regimes — Arctan generates what we call **"protective alpha":** more consistent cross-sectional performance, compounded by variable leverage that sharply curtails risk when necessary. The models have their strongest performance characteristics at the "barbell" of regimes, when they identify confirmed booms or developing crises.

Challenge 3: Managing Pervasive Non-Stationarity

Financial and economic data is both very noisy and unstable: its signal-to-noise ratio is low. Find a factor that seems to work, and as the literature has shown, that "discovery" will often lead to that factor no longer working — either through arbitrage by investors who read the same paper or through structural changes in the economy.

The challenge is to implement both **diagnostic and prescriptive procedures** to determine whether measured relationships still hold.

Diagnostically, one must examine the data and exercise judgement about sample length and other forms of regularization. These judgements are precisely where domain and expert knowledge are most valuable.

Prescriptively, one must apply a battery of statistical tests to measure signal quality on an ongoing basis. While the short-term nature of financial samples often limits the power of these tests, when combined with domain knowledge a rising probability of a structural break can be handled by re-estimating the relevant statistical relationships or reducing the overall level of risk being taken.

An important point deserves emphasis: though every Arctan model operates without discretionary overrides at the decision-point level — parameters are set during model specification and held fixed between scheduled reviews — the necessity to proactively measure and judge model performance in real time means the models are never truly acting without human oversight. **Active and continuous monitoring is a core part of the architecture, not an afterthought.**

Arctan's Architecture: A Neuro-Symbolic Synthesis

An earlier generation of AI models — the *symbolic AI* or expert-systems era, which flourished from the 1960s through the 1980s with landmark systems such as MYCIN and DENDRAL — used a combination of symbolic logic and expert opinion to encode decision rules for specific domains.[6] This approach, despite showing early promise, was largely abandoned when researchers found that much expert judgment could not in fact be consciously articulated and codified. The field entered an AI winter, followed by the rise of statistical machine learning (SVMs, random forests, boosting) through the 1990s and 2000s, and then the deep learning revolution, conventionally dated to 2012.

The rise of ML, and especially deep learning, went in the opposite direction: let the model learn everything from the data, with minimal prior structure.

One way to understand Arctan is as a junction of these two approaches — what the contemporary AI research literature terms *neuro-symbolic AI:*[7] expert, encoded domain knowledge guiding design choices where it adds value, combined with dynamic learning from data for both signal extraction and risk management. Until we have full AGI, this combination — symbolic structure where we have it, statistical learning where we need it — is the most robust path forward.

Conclusions

The application of AI and ML to finance opens many new opportunities. The research cycle is vastly accelerated: going from thesis to code to testing — something that previously required weeks or months — can now be done in a fraction of the time. LLMs can be used effectively to communicate with prospective and existing clients. The gains in efficiency and cost reduction are massive, especially for organizations that begin AI-centered rather than having to retrofit AI onto legacy processes.

Arctan Research is 100% AI-centered. Nothing is done without AI tools and agents.

But AI and ML must be adapted to the specific problems at hand. Financial markets are not only complex but highly mutable. It is not sufficient to find profitable strategies; they must *be robust.* When there is a trade-off between profitability and robustness, always choose robustness. Only robustness maximizes the probability that the encoded investment process can survive the many frequent small changes in its environment and adapt when very large changes occur.

This white paper has laid out the design philosophy underlying the Arctan product family — the Arctan Long-Only and Market-Neutral equity strategies — showing how each architectural choice is a direct response to one of the three fundamental challenges of ML in finance: too many factors (response: curated features), only one timeline (response: regime conditionality), and pervasive non-stationarity (response: continuous monitoring and adaptive re-estimation).

What Arctan Is Not Claiming

We are not claiming AGI. We are not claiming to have solved non-stationarity. We are not claiming alpha from novel factors undiscovered by the academic literature. We are claiming something more modest and more durable: a disciplined, systematically monitored framework that uses ML in the way ML actually works in data-scarce, non-stationary environments — guided by domain expertise, bounded by regime structure, and monitored continuously for signs of degradation.

References

Cochrane, J.H. (2011). Discount Rates. Journal of Finance, 66(4), 1047–1108.

Garcez, A. d'A. & Lamb, L.C. (2020). Neurosymbolic AI: The 3rd Wave. arXiv:2012.05876.

Gu, S., Kelly, B. & Xiu, D. (2020). Empirical Asset Pricing via Machine Learning. Review of Financial Studies, 33(5), 2223–2273.

Harvey, C.R., Liu, Y. & Zhu, H. (2016). …and the Cross-Section of Expected Returns. Review of Financial Studies, 29(1), 5–68.

Hoffmann, J. et al. (2022). Training Compute-Optimal Large Language Models. arXiv:2203.15556.

Hou, K., Xue, C. & Zhang, L. (2020). Replicating Anomalies. Review of Financial Studies, 33(5), 2019–2133.

Kaplan, J. et al. (2020). Scaling Laws for Neural Language Models. arXiv:2001.08361.

Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. NeurIPS 2012.

McCulloch, W.S. & Pitts, W. (1943). A Logical Calculus of the Ideas Immanent in Nervous Activity. Bulletin of Mathematical Biophysics, 5, 115–133.

Rosenblatt, F. (1957). The Perceptron: A Perceiving and Recognizing Automaton. Cornell Aeronautical Laboratory Report.

Sutton, R. (2019). The Bitter Lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html

Vaswani, A. et al. (2017). Attention Is All You Need. NeurIPS 2017. arXiv:1706.03762.