A PATH TOWARD AI SUPERFORECASTERS

Denizalp Goktas, Gerardo Riaño-Briceño, Amy Greenwald
Simulacrum
New York City, NY, USA
{deni, gerardo, amy}@smlcrm.com

ABSTRACT

This note presents the technical challenge of designing forecasting systems that remain accurate under distribution shifts—changes in data across domains (concept shifts) and over time (structural shifts). Addressing this technical challenge amounts to the development of an AI superforecaster, for which we outline a three-phase development roadmap: (1) developing universal forecasters capable of adapting across domains, (2) building structural forecasters that remain accurate over long horizons, and (3) unifying both approaches into a universal structural forecaster—an AI system that can forecast anything, far into the future, with consistent accuracy. We provide evidence for the feasibility of this roadmap by summarizing our proof-of-concept work on universal forecasters, structural forecasters, and forecasting evaluation frameworks.

Simulacrum's mission is to build AI agents that see the future. Equipped with this knowledge, Simulacrum will empower people and machines to make better decisions. This note describes the technical problem Simulacrum aims to solve, the solution we propose, and evidence that our solution is effective.

The Problem: Existing AI forecasting systems do not adapt well to distribution shifts

An *(AI) forecasting system* (or *forecaster*) is software that maps historical values of *context* variables (i.e., *predictors*) to future predictions of the values of *target* variables (i.e., *predictands*), by modeling how the distribution of predictands depends on the distribution of predictors. Existing forecasting systems struggle to adapt to *distribution shifts* (i.e., divergence between the distribution of past predictors and the distribution of future predictands). Distribution shifts can be decomposed along two primary axes: (i) concept shifts and (ii) structural shifts.

A *concept shift* (or *domain shift*) arises from differences between the domains of the past predictors and the future predictands: e.g., training a weather forecasting model on data from the northern hemisphere and applying it to the southern hemisphere.[1] In contrast, a *structural shift* (or *temporal shift*) arises from differences between the distribution of past values of the predictors and the predictands from that of their

[1]These shifts are "concept" shifts as the meanings of the past predictors and the future predictands diverge.

future values.[2] Put succinctly, concept shifts capture differences in the distribution of the predictor and the predictand while fixing time, whereas structural shifts capture differences in the distribution over time while fixing the predictors and the predictands.

To illustrate the problem at hand, consider global weather forecasting systems such as GraphCast [Lam et al., 2023]. These forecasting systems have achieved state-of-the-art accuracy in predicting weather over a discretized grid of the world for a forecast horizon of up to 10 days. However, these global models are known to fail to forecast storm and rainfall intensity, as these predictands depend on local predictors, e.g., local atmospheric variables [Leffer, 2024]. That is, forecasting accuracy degrades under concept shifts (local vs. global). Similarly, these models fail to make accurate weather forecasts for horizons of 10+ years, as future weather data a decade on is likely to behave differently to past weather data, due to environmental changes (e.g., changes in atmospheric gases due to human activity) [Leffer, 2024]. That is, forecasting accuracy degrades under structural shift (weather vs. climate). Weather forecasting systems are only one of the myriad of forecasting systems that are subject to distribution shift, from economics to physics forecasting systems.

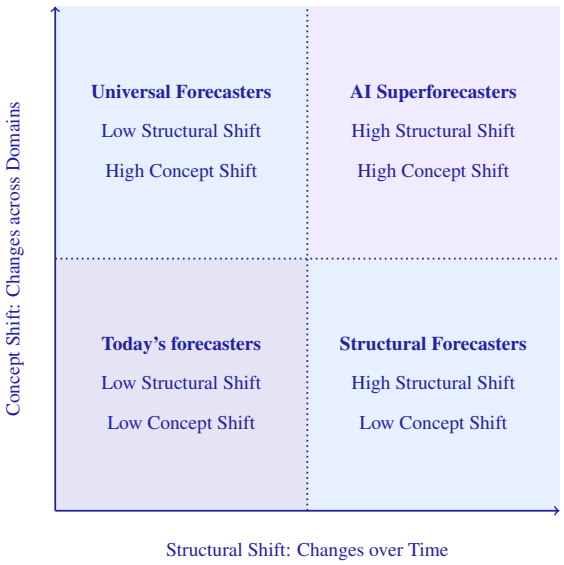

Figure 1: Categorization of the optimal forecasters for different distribution shift regimes.

The Solution: AI Superforecasters

Solving this problem is Simulacrum's north star, and amounts to the development of an AI superforecaster. An *AI superforecaster* is a general-purpose forecasting system that consistently ranks among the top percentiles of all forecasters in terms of forecasting score (e.g., Brier score [Brier, 1950]) across a large and diverse collection of benchmarks (e.g., stock market forecasting) over long time horizons (e.g., decades) [Tetlock and Gardner, 2016]. The development of an AI superforecaster would solve the problem before us, as by definition, a superforecaster can handle concept shifts (i.e., superior performance on diverse benchmarks)

[2]These shifts are "structural" shifts as the distribution of the predictors and predictands changes over time only due to changes in the underlying data-generating process.

and structural shifts (i.e., consistent superior performance across long time horizons). By building an AI superforecaster, Simulacrum will have achieved its mission.

Our plan to develop an AI superforecaster will unroll in three technical phases. First, we will build *universal forecasters*, i.e., AI forecasters that can adapt to concept shifts. These systems will achieve superior accuracy across diverse forecasting problems but their performance may degrade over long-time horizons due to structural shifts. Second, we will develop *structural forecasters*, i.e., AI forecasters that can adapt to structural shifts. These models will be domain-specific but will achieve superior accuracy both for short- and long-term forecast horizons; however, their application will be restricted to forecasting problems in their domains. Once these two phases are completed, leveraging the infrastructure we have developed in prior phases, we will develop a universal structural forecaster, i.e., an AI forecaster that can handle both concept and structural shifts. We believe that the structural universal forecaster produced during this third and final phase will become an AI *super*forecaster, capable of consistent superior performance compared to all other forecasters.

Phase 1: Universal Forecasters

Existing forecasters are overwhelmingly *narrow forecasters*, i.e., they do not adapt well to concept shifts. To remedy this issue, a nascent deep learning literature on time-series foundation models aims to build general purpose neural forecasters, also known as universal forecasters [Woo et al., 2024], which take as input any past time-series data (i.e., a context) and forecast associated future values.

In the last two years, significant progress has been made in training universal neural forecasters, but the performance of these models, despite their scale and complexity, is still only on par with the most basic forecasters on certain forecasting benchmarks [Liang et al., 2024]. Our paper introducing our time-series foundation model, the Likelihood Aligned Forecast Network (LAFN) [Goktas et al., 2025b], provides evidence that we can build a highly accurate universal forecaster. Specifically, we demonstrate that our foundation model achieves superior or on-par performance as compared to existing forecasters across a variety of benchmarks, and with only a fraction of the number of parameters of other universal forecasters.

Phase 2: Structural Forecasters

Unfortunately, data-driven deep learning approaches are likely to fail at making accurate forecasts in long-term horizons. The reason for this failure is attributed to 1) neural networks' high sample complexity without

enough inductive biases, as otherwise they can learn spurious relations [Zhang et al., 2017], and 2) the inability of solely data-driven methods to adapt to structural distribution shifts [Koh et al., 2021].

A nascent literature in structural forecasters [Karniadakis et al., 2021] has sought to remedy these issues by augmenting the training of neural networks with structural constraints. In physics, these models have been called physics-informed neural networks [Raissi et al., 2017], and have, for instance, outperformed solely data-driven weather forecasters [Verma et al., 2024]. In economics, these models have been called structural models [Keane, 2010], and are the state-of-the-art forecasting models used at central banks [Angrist and Pischke, 2010]. In a 2024 ICLR paper written by members of Simulacrum [Goktas et al., 2025a], we provide a mathematical framework that characterizes structural models, together with an algorithm to efficiently build them, and provide proof-of-concept experiments on real-world electricity market data, which show that the structural model learned by our algorithm outperforms widely used purely data-driven forecasters.

Phase 3: AI Superforecasters as Universal Structural Forecasters

Universal forecasters can adapt to concept shifts. Structural forecasters can adapt to structural shifts. Yet, neither alone is strong enough to satisfy the definition of a superforecaster. To achieve this milestone, we will unify the strengths of both approaches into a single forecasting system: a *universal structural forecaster*.

A universal structural forecaster combines (i) the generalization capabilities of universal forecasters across diverse domains with (ii) the robustness of structural forecasters over long-horizon temporal shifts. From a technical perspective, this requires the automated generation of structural constraints that encode how the world evolves over time, which will be used to build inductive biases into the neural universal forecaster. Operationally, we will rely on massive corpus of time-series datasets, efficient structural learning algorithms, and forecaster evaluation tools capable of detecting both forms of distribution shifts.

Critically, the development of a superforecaster demands not only new modeling paradigms but also rigorous methods of performance evaluation. Existing forecasting benchmarks rarely test for both concept and structural shifts simultaneously [Liang et al., 2024]. To ensure our models are truly shift-robust, we have been building benchmarks that span concept shifts and structural shifts.

To this end, we have developed and open-sourced *TempusBench*, a forecasting evaluation framework that enables fair, scalable, and shift-aware comparisons of forecasters across more than 20 forecasting

tasks [Goktas et al., 2025c]. *TempusBench* provides the infrastructure needed to measure progress toward superforecasting and to detect distribution shifts in advance.

Our initial universal forecaster prototype, the Likelihood Aligned Forecast Network (LAFN), already shows the promise of this direction. LAFN is robust, efficient, and surpasses state-of-the-art models on a wide range of forecasting benchmarks, despite using far fewer parameters [Goktas et al., 2025b]. The future iterations of our architecture in Phase 3 will incorporate the structural modeling techniques (e.g., latent dynamical laws, conservation constraints, equilibrium conditions) developed in Phase 2 [Goktas et al., 2025a], yielding a model capable of continuous adaptation under both concept and structural shifts.

Conclusion

Put simply: Phase 1 allows us to forecast anything. Phase 2 allows us to forecast far into the future. Phase 3 enables us to do both *consistently*. Over the years to come, in line with the roadmap outlined in this paper, we will release increasingly better forecasters and enhance our evaluation framework *TempusBench* to reach our goal of developing an AI superforecaster. As we come closer and closer to reaching this goal, our AI superforecaster will see the future. As a result, both humans and machines will be better informed, creating a world of smarter decision makers. In many ways, our AI superforecaster will become an image of the future itself: a *simulacrum*.

References

Joshua D Angrist and Jörn-Steffen Pischke. The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspectives*, 24(2):3–30, 2010. (Cited on page 4.)

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1): 1–3, 1950. (Cited on page 2.)

Denizalp Goktas et al. Efficient inverse multiagent learning. *arXiv preprint arXiv:2502.14160*, 2025a. (Cited on pages 4 and 5.)

Denizalp Goktas et al. Likelihood-aligned forecast networks. *arXiv preprint arXiv:2501.xxxxx*, 2025b. Replace ID once finalized. (Cited on pages 3 and 5.)

Denizalp Goktas et al. Tempusbench: An evaluation framework for time-series forecasting. In *Recent Advances in Time Series Foundation Models Workshop*, 2025c. (Cited on page 5.)

George Em Karniadakis et al. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021. (Cited on page 4.)

Michael P Keane. Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics*, 156(1): 3–20, 2010. (Cited on page 4.)

Pang Wei Koh et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5637–5664, 2021. (Cited on page 4.)

Remi Lam et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. (Cited on page 2.)

Lauren Leffer. Ai weather forecasting can't replace humans—yet. *Scientific American*, 2024. 9 Jan. (Cited on page 2.)

Yuxuan Liang et al. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. (Cited on pages 3 and 4.)

Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics-informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017. (Cited on page 4.)

Philip E Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Random House, 2016. (Cited on page 2.)

Yogesh Verma, Markus Heinonen, and Vikas Garg. Climode: Climate and weather forecasting with physics-informed neural odes. *arXiv preprint arXiv:2404.10024*, 2024. (Cited on page 4.)

Gerald Woo et al. Unified training of universal time series forecasting transformers. *arXiv preprint arXiv:2401.53140*, 2024. (Cited on page 3.)

Chiyuan Zhang et al. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017. (Cited on page 4.)