FORESIGHT

The International Journal of Applied Forecasting

ARTICLES *by* FORECASTERS *for* FORECASTERS: Q2:2024

LLMs and Foundational Models:
Not (Yet) as Good as Hoped

Join the *Foresight* readership by becoming a member of the International Institute of Forecasters
forecasters.org/foresight/

LLMs and Foundational Models: Not (Yet) as Good as Hoped

CHRISTOPH BERGMEIR

PREVIEW *"Foundational" time series models are likely the next big thing in time series forecasting. These are models that have been pretrained with many time series and/or work off the back of a large language model. In this paper, Christoph Bergmeir takes a critical look at the current state and future prospects of these models. As has happened before in the fastpaced field of machine learning, researchers rushing to publish results sometimes cut corners along the way – such as failing to compare new methods to suitable benchmarks. While there are clearly exciting developments with great potential for the future, not everything is as great as some of the research papers and marketing in this space suggest.*

BACKGROUND

Academic papers are an important resource for practitioners to learn about the most recent developments in a particular field. Before publication in an academic venue, papers go through a rigorous peer review process. While the peer review system itself has some recognized issues (which are outside the scope of this discussion), the results and conclusions presented in academic papers can usually be trusted.

In contrast, the field of machine learning (ML) has had a peculiar reliance on a handful of prestigious top-ranked conferences, instead of academic journals as in other fields. The publication process for these conferences is fast paced and, with the popularity of ML, it has become crowded. This exacerbates problems in the peer review process, with too little time for reviewers and authors, and the volume of papers outgrowing the number of experienced reviewers available. Thus, some papers are published that may not be good quality research.

In the ML forecasting literature, quality issues commonly manifest themselves in two ways. First, authors fail to compare their proposed approach to suitable benchmarks, instead comparing only to variants of their own method or other methods in their subfield, or to suboptimal implementations of other methods. Second, authors can focus on side problems of mere academic interest, making subsequently overstated claims of the general superiority of their method. Such issues are not new, and in fact motivated Spyros Makridakis and colleagues (2018) to organize the M4 forecasting competition.

Overstated claims of accuracy and superiority are widespread with transformer architectures, as others and I have previously documented (Hewamalage and colleagues, 2023). In this paper I focus on LLMs and foundational models for forecasting. This is the newest pocket of research with very promising methodology, but also with some issues mostly in the experimental setups of the research.

CURRENT STATE OF RESEARCH IN THE AREA

In the following I want to briefly discuss the most recent research in the area, problems I have spotted, and the responses I received from authors.

Papers with Misleading Evaluations

As a starting point, one paper to call out is TimesNet (Wu and colleagues, 2023),

Key Points

- Foundational models – generic pretrained models for forecasting – are a hot topic. They can be pretrained on time series data and/or work off a pretrained large language model.
- Recent publications in top machine learning conferences suggest that these models outperform all competition on the M4 dataset. However, they do not adequately benchmark against the original competition participants and do not manage to outperform their accuracy.
- Apart from accuracy, many other questions arise in their use. How to detect and deal with data leakage when using public benchmark datasets? What data have these models been pretrained on? What data confidentiality constraints or operational constraints on forecast speed and reliability does your organization have, and can those models work within these constraints?
- As such, the current state of these models is that they achieve interesting accuracy and have a promising path forward, but they do not yet surpass more conventional models.

from the International Conference on Learning Representations (a top ML conference). Their selling point is that TimesNet can do long-term forecasting, short-term forecasting, imputation, classification, and anomaly detection, all with the same methodology and all with state-of-the-art (SOTA) performance. For short-term forecasting, they use the M4 dataset. They report (Table 3 on page 7) an overall weighted average (OWA) of their method of 0.851 (OWA is an accuracy measure used in the M4 competition, and smaller is better). This lands TimesNet in first place among their comparison methods. Also, on their Web page (***github.com/thuml/Time-Series-Library***, accessed 2/23/2024), they show a leaderboard for the SOTA of short-term forecasting (which for them means performance on the M4) where TimesNet wins and some transformer methods come in second and third.

The problem here is that if we revisit the results from the original M4 (Makridakis and colleagues, 2020, Table 4), we will see that the competition was actually won by the method ES-RNN with an OWA of 0.821. Second place was FFORMA with an OWA of 0.838. The stated OWA of TimesNet would put it in seventh place, behind methods that used no deep learning at all and mostly not even ML. This is still not a bad place (out of the 61 original participants), but certainly not SOTA, which is usually understood to mean the best result a method has been able to achieve on this dataset. Another interesting aspect is their treatment of N-BEATS. This method got into the spotlight some time ago when it first came out precisely by claiming to achieve SOTA in the M4, with an OWA of 0.795 (Oreshkin and colleagues, 2019, Table1). The TimesNet paper indeed does report results for N-BEATS, but with an OWA of 0.855.

Other papers followed, with a few claiming SOTA accuracy on the M4, but omitting the fact that they are actually not able to outperform the original winner. I'll *not* discuss most of them here for brevity but do want to mention as an extreme case, TIME-LLM (Jin and colleagues, 2023), that claims to win on the M4 with an OWA of 0.859. If you might understandably ask how this can now suddenly be a winner, it's that they report OWAs for TimesNet of 0.955 and for N-BEATS of 0.896, which would place these methods with those "updated" results outside of the top 15 in the M4, and somewhere close to the Theta benchmark (N-BEATS) or even considerably worse (TimesNet).

One thing we seem to be learning from this is that ML methods are not so easy to use for nonexperts. If already other top ML scientists cannot run them in a way that they achieve performance close to the original results, how will an average data science graduate in a company make them work better than simple benchmarks?

Finally, the TIME-LLM paper reports sMAPE results on the quarterly M3 dataset (Table 13 in the Appendix of the current version of the paper from openreview.net, accepted at ICLR 2024) of 11.171 for TIME-LLM and 10.410 for TimesNet. When we now look at the results from the original participants in the M3 from 20 years ago (see, e.g., Table 5 in Bergmeir and colleagues [2016]), we see that it was won by Theta with an sMAPE of 8.956, and NAIVE2 had an sMAPE of 9.951. As such, these results likely to be published in ICLR 2024 as "winners" are worse than a NAIVE2 from 20 years ago.

The Responses from the Authors

When I asked the TimesNet authors for the discrepancies for N-BEATS and the results of the original competitors, regarding N-BEATS they replied: "As we stated in the paper, N-BEATS employs a special ensemble method, which incorporates the results predicted from different input series. Its final results are ensembled from 7 models. Thus, to ensure a fair comparison, we test all the models with one single input length." Regarding the original M4, they say "Note that different from competition, for research, a fair comparison is essential. Without fair comparison, we cannot obtain any scientific conclusions" (***github.com/thuml/Time-Series-Library/issues/293***, accessed 2/23/2024). I'm not sure what Makridakis and the M4 team would respond to statements hinting that the M4 was not a fair comparison. In fact, I cannot think of a fairer comparison than a competition, and Makridakis's motivation for the competitions was to create a platform for an objective, academic, fair comparison. Anyway, this argument seems to have found widespread adoption among these papers. For example, Wang and colleagues (2023) say, "The original paper of N-BEATS (2019) adopts a special ensemble method to promote the performance. For fair comparisons, we remove the ensemble and only compare the pure forecasting models." In a private conversation with me, the authors of Time-LLM brought to their defense this same argument, along with arguments along the lines of only wanting to benchmark deep-learning methods.

A Breaststroke Competition

I want to argue that if those authors are going to host a breaststroke competition, then this might be exciting for all sorts of reasons. But they should make it clear to anybody that this is a breaststroke competition and not freestyle – as in the freestyle competition they are losing. They should make this particularly clear to all the practitioners who are now going to take these methods out into the wild, wasting lots of time, energy, and computer resources before realizing that these methods are not yet what they promise. Knowing these top-level conferences and how they operate, I think it is safe to assume that these papers would have had a much more difficult stance with the reviewers if these things had been made clearer.

Also, the rules that those papers apply are quite arbitrary. Why can't I use an ensemble? Does that mean they don't have to win against a random forest in their competition? Note that the winning M4 method was in fact a deep-learning method, so it should qualify at least in this sense for their breaststroke competition. However, it had some ensembling elements to it.

We should also keep in mind that the original participants of the M4 didn't have the luxury of knowing the test set, an advantage that all subsequent methods do have, and that can lead to overfitting in potentially subtle ways. Also, I'm aware that merely checking numbers in papers and comparing them has its pitfalls, and I didn't do any calculations myself. Did they use the same definition of OWA? Did they maybe leave out some time series for whatever reason? Are we really comparing apples to apples here? None of the authors hinted to anything along these lines in their defense, so I assume that these results are what they claim to be, namely OWA on the full M4, comparable to the original participants' results. Furthermore, it should be the first

responsibility of the authors of these papers to show to me as a reader how these methods fare compared with the original submissions, to be able to make a direct assessment of how well they perform.

WHERE TO FROM HERE?

What does all this mean for the current state of forecasting models that are pretrained and/or use LLMs? Contenders I want to mention more positively are TimeGPT-1 (Garza and Mergenthaler-Canseco, 2023), Lag-Llama (Rasul and colleagues,2023), and LLMTIME (Gruver and colleagues, 2023). They all have credible results that are good if not disrupting-the-field-of-forecasting level of good (yet), and thus more consistent with the results from above when you take my discussions into account. As such, these models are clearly coming but not there yet. A second place in the M4 that some of these models achieve is remarkable. We can still develop dedicated models that outperform those models purely based on accuracy. However, once these models really do achieve the accuracies that we are already now promised they have, some interesting issues will arise, as follows.

Data Leakage Will Be a Major Challenge

Data leakage is always a problem in forecasting (I've written about it in Bergmeir [2023]). Already, global models that train across time series face these problems more than local per-series univariate models. For example, models pretrained on the M4 seem to show really good performance on the M3. One has to wonder how much this has to do with the M4 dataset being put together roughly 20 years after the M3 dataset, with ample room to contain information about the future of the M3 series. Since the M3 and the M4 were put together by similar authors, it may well be that the data sources overlap to a degree. So it is difficult to tell how much of this performance is relevant for real-world forecasting applications and how much is merely due to some form of data leakage.

With foundational and LLM-based models, this will become a big challenge. Our evaluation methodologies currently all rely on experiments on publicly available benchmark datasets. LLM producers usually don't disclose the datasets they train on (and for good reason: see current copyright lawsuits such as *New York Times* vs OpenAI). So, while it is somewhat unlikely that they would train with time series data, the reality is that we don't know. Furthermore, a known weakness of LLMs is their bad performance at solving mathematical problems. So it wouldn't be surprising if soon we see models that are pretrained on numerical data, including time series. When looking at pure time series models, Lag-Llama is clear about the data they train on, whereas TimeGPT is less so. We have a good idea of what they likely trained on, which is all publicly available forecasting datasets that they could find.

In the future, we may be forced to use nonpublic benchmark datasets, and to be careful to only benchmark models that we can run ourselves, as sending them to a Web service like TimeGPT will effectively hand over this dataset to the model developers who will likely use the data to train the next iteration of their model.

But if you now have a pretrained model that has been trained on all publicly available datasets, then how can you evaluate such models? In the future, we may be forced to use nonpublic benchmark datasets, and to be careful to only benchmark models that we can run ourselves, as sending them to a Web service like TimeGPT will effectively hand over this dataset to the model developers who will likely use the data to train the next iteration of their model (the TimeGPT Terms & Conditions have a "Use of Content to Improve Services" clause, from which one can opt out).

What About External Covariates? And How Much Can You Learn from 14 Data Points?

Another problem that I'm sure will be solved at some point but is not addressed yet, at least as far as I can see, is the context of a time series, be it metadata or external covariates.

If you have a forecasting task that is wind power forecasting and you have wind and temperature as covariates, and then you have another forecasting task in retail where your covariates may be promotions, prices, and other factors, it is not clear how one model could handle such different covariates. Maybe this can be done with embeddings or certain forms of dimensionality reduction techniques (principal component analysis in the simplest case). Another way forward could be so-called prior-data fitted networks, that are able to learn a training and inference algorithm directly. I can also think of stacking solutions, where you take the output of the foundational model and feed it as an additional input into your specialized forecasting method that also gets fed all the covariates.

Finally, this could be a space where LLMs actually shine. Let's assume you have a yearly time series and have a long history, which could be 14 years. That gives you 14 data points in your series. With the current implementations, you would just feed in these 14 data points with no additional information. But there is a limit to how much information you can extract from 14 data points, even for an LLM-enhanced-transformer-deep-learning-super-model. To be able to predict from such small amounts of data, a model needs to make strong assumptions and have strong regularization. As such, even the most complex model would eventually fall back to something simpler, and the more important question is whether your assumptions about the data are correct. As an example, suppose I know the yearly series are sales of millions of a product like chewing gum that hasn't yet entered large markets such as China or India. It will be more reasonable to forecast the continuation of an exponential growth that we see in the chewing gum data than if we know that the series are sales of millions of iPhones. The big strength that LLMs promise here is that they could be able to take in all contexts in any form that you may have about your time series. LLMs allow you to build assumptions and biases into the model that are not justified from the time series data alone, but from the context you have provided and the general "knowledge" the LLM possesses.

What About Data Confidentiality and the Constraints of Your Production Environment?

Other interesting questions arise from a purely operational perspective of how these models will actually be used. If you need to send your time series to a Web service that gives you back forecasts, you may not be able to do it because the data is confidential. Or you may have some real-time constraints, e.g., in wind power forecasting you may need to produce a five-minute-out forecast. If it takes you more than 20 seconds to generate your forecast, you may already be worse than just a naive/persistence forecast produced 20 seconds later. There could be other operational constraints around reliable internet connections, etc. So an interesting question is whether there will be open-source pretrained models available that you can run on site, in the way you want to run them.

LLMs allow you to build assumptions and biases into the model that are not justified from the time series data alone, but from the context you have provided and the general "knowledge" the LLM possesses.

CONCLUSIONS

Although performance claims are often overstated, valuable developments are happening in the space of pretrained and LLM-based models for forecasting.

While we should not yet throw overboard everything we've learned about forecasting over the last 40-plus years, the future promises to be exciting.

REFERENCES

Bergmeir, C., Hyndman, R.J. & Benítez, J.M. (2016). Bagging Exponential Smoothing Methods Using STL Decomposition and Box–Cox Transformation, *International Journal of Forecasting*, 32(2), 303-312.

Bergmeir, C. (2023). Common Pitfalls and Better Practices in Forecast Evaluation for Data Scientists, *Foresight*, Issue 70, 5-12.

Garza, A. & Mergenthaler-Canseco, M. (2023). TimeGPT-1. arXiv preprint arXiv:2310.03589.

Gruver, N., Finzi, M., Qiu, S. & Wilson, A.G. (2023). Large Language Models are Zero-Shot Time Series Forecasters, arXiv preprint arXiv:2310.07820. NeurIPS (2023).

Hewamalage, H., Ackermann, K. & Bergmeir, C. (2023). Forecast Evaluation for Data Scientists: Common Pitfalls and Best Practices, *Data Mining and Knowledge Discovery*, 37(2), 788-832.

Jin, M. et al. (2023). Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. ICLR 2024 (accepted). arXiv preprint arXiv:2310.01728.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2020). The M4 Competition: 100,000 Time Series and 61 Forecasting Methods, *International Journal of Forecasting*, 36(1), 54-74.

Makridakis, S., Spiliotis, E. & Assimakopoulos, V. (2018). Statistical and Machine Learning Forecasting Methods: Concerns and Ways Forward, *PLOS One*, 13(3), e0194889.

Oreshkin, B.N., Carpov, D., Chapados, N. & Bengio, Y. (2019). N-BEATS: Neural Basis Expansion Analysis for Interpretable Time Series Forecasting, ICLR, 2020.

Rasul, K. et al. (2023). Lag-llama: Towards Foundation Models for Time Series Forecasting. arXiv preprint arXiv:2310.08278.

Wang, S. et al. (2023, October). TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting, In *The Twelfth International Conference on Learning Representations*.

Wu, H. et al. (2023). Timesnet: Temporal 2d-Variation Modeling for General Time Series Analysis. *arXiv preprint arXiv:2210.02186*.

Christoph Bergmeir is a María Zambrano Senior Fellow in the Department of Computer Science and Artificial Intelligence at University of Granada, Spain. Previously he was a Visiting Research Data Scientist at Meta Inc. in the U.S., and a Senior Lecturer at Monash University, Australia. Christoph holds a PhD in computer science from the University of Granada and an MSc in computer science from the University of Ulm, Germany. He has worked in forecasting for capacity planning, sustainable energy, and supply chain, and has over 7,000 citations, an h-index of 33, and has received more than $2.7 million in external research funding. Four of his publications on time series forecasting have been Clarivate Web of Science Highly Cited Papers (top 1% of their research field).

christoph.bergmeir@gmail.com

This article originally appeared in *Foresight*, Issue 73 (forecasters.org/foresight) and is made available with permission from *Foresight* and the International Institute of Forecasters.