Position: Stop Making Unscientific AGI Performance Claims

Patrick Altmeyer [*1] **Andrew M. Demetriou** [*1] **Antony Bartlett** [1] **Cynthia C. S. Liem** [1]

Abstract

Developments in the field of Artificial Intelligence (AI), and particularly large language models (LLMs), have created a 'perfect storm' for observing 'sparks' of Artificial General Intelligence (AGI) that are spurious. Like simpler models, LLMs distill meaningful representations in their latent embeddings that have been shown to correlate with external variables. Nonetheless, the correlation of such representations has often been linked to human-like intelligence in the latter but not the former. We probe models of varying complexity including random projections, matrix decompositions, deep autoencoders and transformers: all of them successfully distill information that can be used to predict latent or external variables and yet none of them have previously been linked to AGI. We argue and empirically demonstrate that the finding of meaningful patterns in latent spaces of models cannot be seen as evidence in favor of AGI. Additionally, we review literature from the social sciences that shows that humans are prone to seek such patterns and anthropomorphize. We conclude that both the methodological setup and common public image of AI are ideal for the misinterpretation that correlations between model representations and some variables of interest are 'caused' by the model's understanding of underlying 'ground truth' relationships. We, therefore, call for the academic community to exercise extra caution, and to be keenly aware of principles of academic integrity, in interpreting and communicating about AI research outcomes.

1. Introduction

In 1942, when anti-intellectualism was rising and the integrity of science was under attack, Robert K. Merton formulated four 'institutional imperatives' as comprising the ethos of modern science: *universalism*, meaning that the acceptance or rejection of claims entering the lists of science should not depend on personal or social attributes of the person bringing in these claims; *"communism"* [sic], meaning that there should be common ownership of scientific findings and one should communicate findings, rather than keeping them secret; *disinterestedness*, meaning that scientific integrity is upheld by not having self-interested motivations, and *organized skepticism*, meaning that judgment on the scientific contribution should be suspended until detached scrutiny is performed, according to institutionally accepted criteria (Merton et al., 1942). While the Mertonian norms may not formally be known to academics today, they still are implicitly being subscribed to in many ways in which academia has organized academic scrutiny; e.g., through the adoption of double-blind peer reviewing, and in motivations behind open science reforms.

At the same time, in the way in which academic research is disseminated in the AI and machine learning fields today, major shifts are happening. Where these research fields have actively adopted early sharing of preprints and code, the volume of publishable work has exploded to a degree that one cannot reasonably keep up with broad state-of-the-art, and social media influencers start playing a role in article discovery and citeability (Weissburg et al., 2024). Furthermore, because of major commercial stakes with regard to AI and machine learning technology, and e.g. following the enthusiastic societal uptake of products employing LLMs, such as ChatGPT, the pressure to beat competitors as fast as possible is only increasing, and strong eagerness can be observed in many domains to 'do something with AI' in order to innovate and remain current.

Where AI used to be a computational modeling tool to better understand human cognition (van Rooij et al., 2023), the recent interest in AI and LLMs has been turning into one in which AI is seen as a tool that can mimic, surpass and potentially replace human intelligence. In this, the achievement of Artificial General Intelligence (AGI) has become a grand challenge, and in some cases, an explicit business goal. The definition of AGI itself is not as clear-cut or consistent; loosely, it is a phenomenon contrasting with 'narrow AI' systems, that were trained for specific tasks (Goertzel, 2014). In practice, to demonstrate that the

[*]Equal contribution [1]Department of Intelligent Systems, Delft University of Technology, Delft, the Netherlands. Correspondence to: Patrick Altmeyer <patrick.altmeyer@tudelft.nl>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

achievement of AGI may be getting closer, researchers have sought to show that AI models generalize to different (and possibly unseen) tasks, with little human intervention, or show performance considered 'surprising' to humans.

For example, Google DeepMind claimed their AlphaGeometry model (Trinh et al., 2024) reached a 'milestone' towards AGI. This model has the ability to solve complex geometry problems, allegedly without the need for human demonstrations during training. However, work such as this had been initially introduced in the 1950s (Zenil, 2024): without the use of an LLM, logical inference systems proved 100% accurate in proving all the theorems of Euclidean Geometry, due to geometry being an axiomatically closed system. Therefore, while DeepMind created a powerfully fast geometry-solving machine, it is still far from AGI.

Generally, in the popularity of ChatGPT and the integration of generative AI in productivity tools (e.g. through Microsoft's Copilot integrations in GitHub and Office applications), one also can wonder whether the promise of AI is more in computationally achieving general intelligence, or rather in the engineering of general-purpose tools[1]. Regardless, stakes and interests are high, e.g. with ChatGPT clearing nearly $1 billion in months of its release[2].

When combining massive financial incentives with the presence of a challenging and difficult-to-understand technology, that aims towards human-like problem-solving and communication abilities, a situation arises that is fertile for the misinterpretation of spurious cues as hints towards AGI, or other qualities like sentience[3] and consciousness. AI technology only becomes more difficult to understand as academic publishing in the space largely favors performance, generalization, quantitative evidence, efficiency, building on past work, and novelty (Birhane et al., 2022). As such, works that make it into top-tier venues tend to propose heavier and more complicated technical takes on tasks that (in the push towards generalizability) get more vague, while the scaling-up of data makes traceability of possible memorization harder. In a submission-overloaded reality, researchers may further get incentivized to oversell and overstate achievement claims. At the same time, while currently popular in literature, inherent complexity and opaqueness in technical solutions may fundamentally be unwise to pursue in high-stakes applications (Rudin, 2019).

Noticing these trends, we as the authors of this article are concerned. We feel that the current culture of racing toward Big Outcome Statements in industry and academic publishing too much disincentivizes efforts toward more thorough and nuanced actual problem understanding. At the same time, as the outside world is so eager to adopt AI technology, (too) strong claims make for good sales pitches, but a question is whether there is indeed sufficient evidence for these claims. With successful AGI outcomes needing to look human-like, this also directly plays into risks of anthropomorphizing (the attribution of human-like qualities to non-human objects) and confirmation bias (the seeking-out and/or biased interpretation of evidence in support of one's beliefs). In other words, it is very tempting to claim surprising human-like achievements of AI, and as humans, we are very prone to genuinely believing this. **We therefore urge our fellow researchers to stop making unscientific AGI performance claims**.

To strengthen our argument, in this paper, we first present related work in Section 2. We then consider a recently viral work (Gurnee & Tegmark, 2023b) in which claims about the learning of world models by LLMs were made. In Section 3, we present several experiments that may invite similar claims on models yielding more intelligent outcomes than would have been expected—while at the same time indicating how we feel these claims should *not* be made. Furthermore, we present a review of social science findings in Section 4 that underline how prone humans are to being enticed by patterns that are not really there. Combining this with the way in which media portrayal of AI has tended towards science-fiction imagery of mankind-threatening robots, we argue that the current AI culture is a perfect storm for making and believing inflated claims, and call upon our fellow academics to be extra mindful and scrutinous about this. Finally, in Section 5, we propose specific structural and cultural changes to improve the current situation. Section 6 concludes.

2. Related Work

In this work, we question the practice of using outcomes from mechanistic interpretability to support AGI claims. This is not to be seen as criticism toward the underlying methodologies in isolation, but rather in the context of current publishing practices that we repeatedly challenge throughout this work. Many closely related works are free of any grandiose conclusions and instead highlight the benefits of mechanistic interpretability that we also highlight here (Nanda et al., 2023; Gurnee et al., 2023; Li et al., 2022).

Another related subfield investigates the capacity of LLMs to reason causally. Here, too, there is an opportunity to over-interpret the finding of causal information as causal understanding. Recent work has shown, for example, that LLMs can indeed correctly predict causal relationships and

[1] A Swiss army knife is an effective general-purpose tool, without people wondering whether it exhibits intelligence.

[2] https://www.bloomberg.com/news/articles/2023-08-30/openai-nears-1-billion-of-annual-sales-as-chatgpt-takes-off

[3] https://www.scientificamerican.com/article/google-engineer-claims-ai-chatbot-is-sentient-why-that-matters/

this may have practical use cases (Kıcıman et al., 2023). But despite the potential utility, the authors also demonstrate that this capacity can be partially explained by memorization, rather than an actual understanding of causal relationships. Similarly, Zečević et al. (2023) provide evidence indicating that current LLMs "may talk causality but are not causal".

Two other recent works are related to this work and align well with the position we present here. Schaeffer et al. (2024) demonstrate that the apparent emergent abilities of large language models may be driven by a choice of evaluation metrics, rather than some fundamental property that is intrinsic to this family of models. Their work highlights the need for rigorous testing and benchmarking of LLMs, which we also point to in this work, albeit in a slightly different methodological context. Kloft et al. (2024) provide experimental evidence demonstrating that people have heightened expectations and a biased, positive view of AI. The authors run a user study of human-AI interaction, in which participants performed better at a given task when they (wrongly) thought they were aided by a positively described AI. This placebo effect was found to be robust to negative descriptions of AI. Positive bias towards AI may exacerbate other factors that drive people to make unscientific claims about the current state of AI, which we discuss in Section 4.

3. Surprising Patterns in Latent Spaces?

In 2023, a research article went viral on the X[4] platform (Gurnee & Tegmark, 2023a). Through linear probing experiments, the claim was made that LLMs learned literal maps of the world. As such, they were considered to be more than 'stochastic parrots' (Bender et al., 2021) that can only correlate and mimic existing patterns from data, but not truly understand it. While the manuscript immediately received public criticism (Marcus, 2023), and the revised, current version is more careful with regard to its claims (Gurnee & Tegmark, 2023b), reactions on X seemed to largely exhibit excitement and surprise at the authors' findings. However, in this section, through various simple examples, we make the point that observing patterns in latent spaces should not be a surprising revelation. After starting with a playful example of how easy it is to 'observe' a world model, we build up a larger example focusing on key economic indicators and central bank communications.

3.1. Are Neural Networks Born with World Models?

Gurnee & Tegmark (2023b) extract and visualize the alleged geographical world model by training linear regression probes on internal activations in LLMs (including Llama-2) for the names of places, to predict geographical coordinates

[4] https://twitter.com/wesg52/status/1709551516577902782?s=20

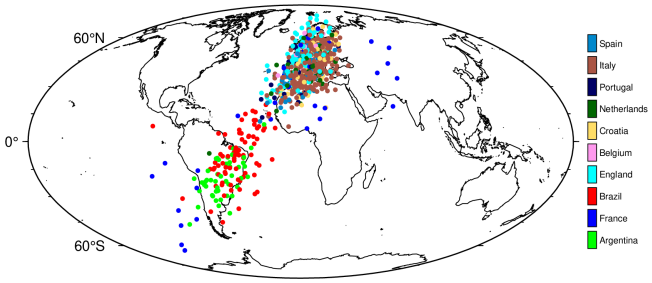

Figure 1. Predicted coordinate values (out-of-sample) from a linear probe on final-layer activations of an untrained neural network.

associated with these places. Now, the Llama-2 model has ingested huge amounts of publicly available data from the internet, including Wikipedia dumps from the June-August 2022 period (Touvron et al., 2023). It is therefore highly likely that the training data contains geographical coordinates, either directly or indirectly. At the very least, we should expect that the model has seen features during training that are highly correlated with geographical coordinates. The model itself is essentially a very large latent space to which all features are randomly projected in the very first instance before being passed through a series of layers which are gradually trained for downstream tasks.

In our first example, we simulate this scenario, stopping short of training the model. In particular, we take the world_place.csv that was used in Gurnee & Tegmark (2023b), which maps locations/areas to their latitude and longitude. For each place, it also indicates the corresponding country. From this, we take the subset that contains countries that are currently part of the top 10 FIFA world ranking, and assign the current rank to each country (i.e., Argentina gets 1, France gets 2, ...). To ensure that the training data only involves a noisy version of the coordinates, we transform the longitude and latitude data as follows: $\rho \cdot \text{coord} + (1-\rho) \cdot \epsilon$ where $\rho = 0.5$ and $\epsilon \sim \mathcal{N}(0, 5)$.

Next, we encode all features except the FIFA world rank indicator as continuous variables: $X^{(n \times m)}$ where n is the number of samples and m is the number of resulting features. Additionally, we add a large number of random features to X to simulate the fact that not all features ingested by Llama-2 are necessarily correlated with geographical coordinates. Let d denote the final number of features, i.e. $d = m + k$ where k is the number of random features.

We then initialize a small neural network, considered a *projector*, mapping from X to a single hidden layer with $h < d$ hidden units and sigmoid activation, and from there, to a lower-dimensional output space. Without performing any training on the *projector*, we simply compute a forward pass of X and retrieve activations $\mathbf{Z}^{(n \times h)}$. Next, we perform the linear probe on a subset of $\mathbf{Z}$ through Ridge regression:

$\mathbf{W} = (\mathbf{Z}'_{\text{train}}\mathbf{Z}_{\text{train}} + \lambda\mathbf{I})(\mathbf{Z}'_{\text{train}}\mathbf{coord})^{-1}$, where **coord** is the $(n \times 2)$ matrix containing the longitude and latitude for each sample. A hold-out set is reserved for testing, on which we compute predicted coordinates for each sample as $\widehat{\mathbf{coord}} = \mathbf{Z}_{\text{test}}\mathbf{W}$ and plot these on a world map (Figure 1).

While the fit certainly is not perfect, the results do indicate that the random projection contains representations that are useful for the task at hand. Thus, this simple example illustrates that meaningful target representations should be recoverable from a sufficiently large latent space, given the projection of a small number of highly correlated features. Similarly, Alain & Bengio (2016) observe that even before training a convolutional neural network on MNIST data, the layer-wise activations can already be used to perform binary classification. In fact, it is well-known that random projections can be used for prediction tasks (Dasgupta, 2013).

This first experiment—and indeed the practice of probing LLMs that have seen vast amounts of data—can be seen as a form of inverse problem and common caveats such as non-uniqueness and instability apply (Haltmeier & Nguyen, 2023). Regularization can help mitigate these caveats (OM, 2001), but we confess that we did not carefully consider the parameter choice for λ, nor has this been carefully studied in the related literature to the best of our knowledge.

3.2. PCA as a Yield Curve Interpreter

We now move to a concrete application domain: Economics. Here, the yield curve, plotting the yields of bonds against their maturities, is a popular tool for investors and economists to gauge the health of the economy. The yield curve's slope is often used as a predictor of future economic activity: a steep yield curve is associated with a growing economy, while a flat or inverted yield curve is associated with a contracting economy. To leverage this information in downstream modelling tasks, economists regularly use PCA to extract a low-dimensional projection of the yield curve that captures relevant variation in the data (e.g. Berardi & Plazzi (2022), Kumar (2022) and Crump & Gospodinov).

To understand the nature of this low-dimensional projection, we collect daily Treasury par yield curve rates at all available maturities from the US Department of the Treasury. Computing principal components involves decomposing the matrix of all yields $\mathbf{r}$ into a product of its singular vectors and values: $\mathbf{r} = \mathbf{U}\Sigma\mathbf{V}'$. Let us simply refer to $\mathbf{U}$, Σ and $\mathbf{V}'$ as latent embeddings of the yield curve.

The upper panel in Figure 2 shows the first two principal components of the yield curves of US Treasury bonds over time. Vertical stalks indicate key dates related to the Global Financial Crisis (GFC). During its onset, on 27 February 2007, financial markets were in turmoil following a warning from the Federal Reserve (Fed) that the US economy was at risk of a recession. The Fed later reacted to mounting economic pressures by gradually reducing short-term interest rates to unprecedented lows. Consequently, the average level of yields decreased and the curve steepened. In Figure 2, we can observe that the first two principal components appear to capture this level shift and steepening, respectively. In fact, they are strongly positively correlated with the actual observed first two moments of the yield curve (lower panel of Figure 2).

Again, it should not be surprising that these latent embeddings are meaningful: by construction, principal components are orthogonal linear combinations of the data itself, each of which explains most of the residual variance after controlling for the effect of all previous components.

3.3. LLMs for Economic Sentiment Prediction

So far, we considered simple linear data transformations. One might argue that this does not really involve latent embeddings in the way they are typically thought of in the context of deep learning. In Appendix A, we present an additional experiment in which we more explicitly seek neural network-based representations that will be useful for downstream tasks. Here, we continue with an example in which LLMs may be used for economic sentiment prediction.

Closely following the approach in Gurnee & Tegmark (2023b), we apply it to the novel *Trillion Dollar Words* (Shah et al., 2023) financial dataset, containing a curated selection of sentences formulated and communicated to the public by the Fed through speeches, meeting minutes and press conferences. (Shah et al., 2023) use this dataset to train a set of LLMs and rule-based models to classify sentences as either 'dovish', 'hawkish' or 'neutral'. In the context of central banking, 'hawkishness' is typically associated with tight monetary policy: in other words, a 'hawkish' stance on policy favors high interest rates to limit the supply of money and thereby control inflation. The authors first manually annotate a sub-sample of the available data and then fine-tune various models for the classification task. Their model of choice, *FOMC-RoBERTa* (a fine-tuned version of RoBERTa (Liu et al., 2019)), achieves an F_1 score of around > 0.7 on the test data. To illustrate the potential usefulness of the learned classifier, they use predicted labels for the entire dataset to compute an ad-hoc, count-based measure of 'hawkishness'. This measure is shown to correlate with key economic indicators in the expected direction: when inflationary pressures rise, the measured level of 'hawkishness' increases, as central bankers react by raising interest rates to bring inflation back to target.

3.3.1. Linear Probes

We now use linear probes to assess if the fine-tuned model has learned associative patterns between central bank com-

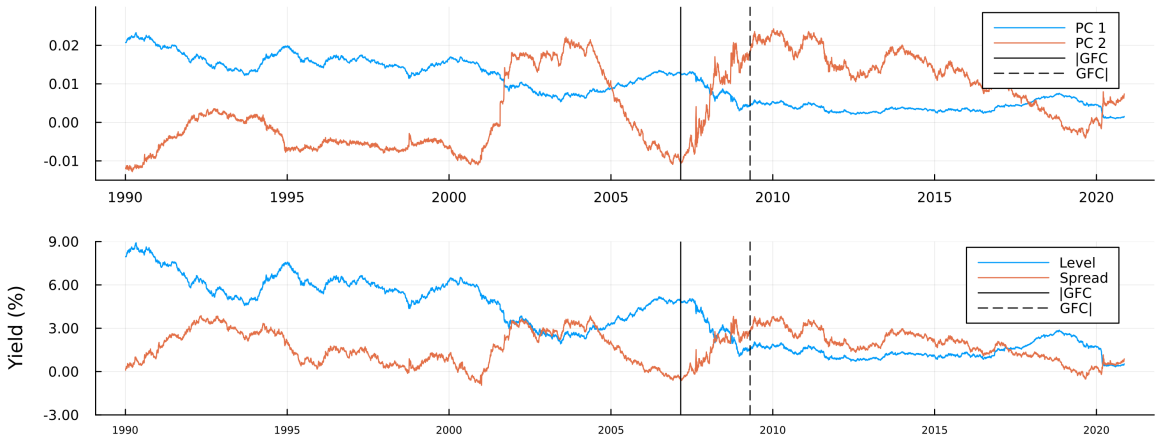

Figure 2. Top chart: The first two principal components of US Treasury yields over time at daily frequency. Bottom chart: Observed average level and 10yr-3mo spread of the yield curve. Vertical stalks roughly indicate the onset (|GFC) and the beginning of the aftermath (GFC|) of the Global Financial Crisis.

munications and key economic indicators. Therefore, we further pre-process the data provided by Shah et al. (2023) and use their proposed model to compute activations of the hidden state, on the first entity token for each layer. We have made these available and easily accessible through a small Julia package: TrillionDollarWords.jl.

For each layer, we compute linear probes through Ridge regression on two inflation indicators (the Consumer Price Index (CPI) and the Producer Price Index (PPI)) and US Treasury yields at different levels of maturity. To allow comparison with Shah et al. (2023), we let yields enter the regressions in levels. To measure price inflation we use percentage changes proxied by log differences. To mitigate issues related to over-parameterization, we follow the recommendation in Alain & Bengio (2016) to first reduce the dimensionality of the computed activations. In particular, we restrict our linear probes to the first 128 principal components of the embeddings of each layer. To account for stochasticity, we use an expanding window scheme with 5 folds for each indicator and layer. To avoid look-ahead bias, PCA is always computed on the sub-samples used for training the probe.

Figure 3 shows the out-of-sample root mean squared error (RMSE) for the linear probe, plotted against *FOMC-RoBERTa*'s n-th layer. The values correspond to averages across cross-validation folds. Consistent with related work (Alain & Bengio, 2016; Gurnee & Tegmark, 2023b), we observe that model performance tends to be higher for layers near the end of the transformer model. Curiously, for yields at longer maturities, we find that performance eventually deteriorates for the very final layers. We do not observe this for the training data, so we attribute this to overfitting.

It should be noted that performance improvements are generally of small magnitude. Still, the overall qualitative findings are in line with expectations. Similarly, we also observe that these layers tend to produce predictions that are more positively correlated with the outcome of interest and achieve higher mean directional accuracy (MDA). Upon visual inspection of the predicted values, we conclude the primary source of prediction errors is low overall sensitivity, meaning that the magnitude of predictions is generally too small.

To better assess the predictive power of our probes, we compare their predictions to those made by simple autoregressive models. For each layer, indicator and cross-validation fold, we first determine the optimal lag length based on the training data using the Bayes Information Criterium with a maximal lag length of 10. These are not state-of-the-art forecasting models, but they serve as a reasonable baseline. For most indicators, probe predictions outperform the baseline in terms of average performance measures. After accounting for variation across folds, however, we generally conclude that the probes neither significantly outperform nor underperform. Detailed results, in which we also perform more explicit statistical testing, can be found in Appendix B.

3.3.2. Sparks of Economic Understanding?

Even though *FOMC-RoBERTa*, which is substantially smaller than the models tested in Gurnee & Tegmark (2023b), was not explicitly trained to uncover associations between central bank communications and the level of consumer prices, it appears that the model has distilled representations that can be used to predict inflation (although they certainly will not win any forecasting competitions). So, have we uncovered further evidence that LLMs "aren't mere stochastic parrots"? Has *FOMC-RoBERTa* developed

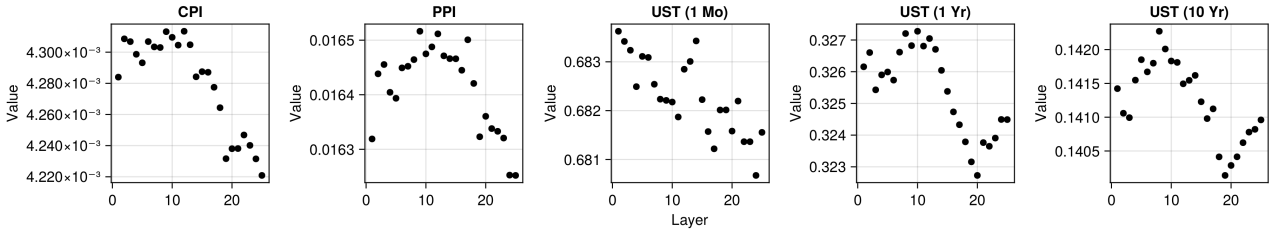

Figure 3. Out-of-sample root mean squared error (RMSE) for the linear probe plotted against *FOMC-RoBERTa*'s n-th layer for different indicators. The values correspond to averages computed across cross-validation folds, where we have used an expanding window approach to split the time series. As expected, model performance tends to be higher (average prediction errors are lower) for layers near the end of the transformer model.

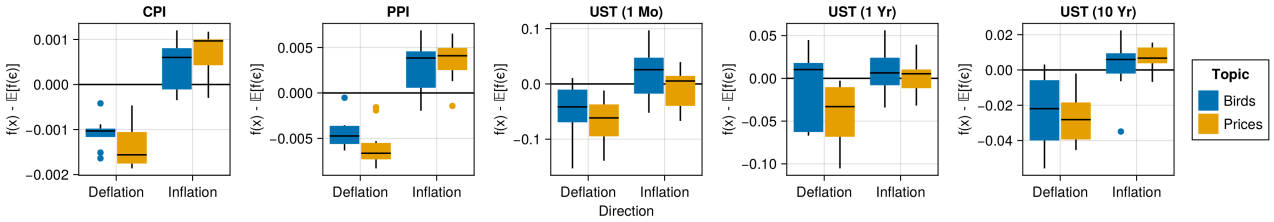

Figure 4. Probe predictions for sentences about inflation of prices (IP), deflation of prices (DP), inflation of birds (IB) and deflation of birds (DB). The vertical axis shows predicted inflation levels subtracted by the average predicted value of the probe for random noise.

an intrinsic 'understanding' of the economy just by 'reading' central bank communications? Thus, can economists readily adopt FOMC-RoBERTa as a domain-relevant tool?

We are having a very hard time believing that the answer to either of these questions is 'yes'. To argue our case, we will now produce a counter-example demonstrating that, if anything, these findings are very much in line with the parrot metaphor. The counter-example is based on the following premise: if the results from the linear probe truly were indicative of some intrinsic 'understanding' of the economy, then the probe should not be sensitive to random sentences that are most definitely not related to consumer prices.

To test this, we select the best-performing probe trained on the final-layer activations for each indicator. We then make up sentences that fall into one of these four categories: *Inflation/Prices* (IP)—sentences about price inflation, *Deflation/Prices* (DP)—sentences about price deflation, *Inflation/Birds* (IB)—sentences about inflation in the number of birds and *Deflation/Birds* (DB)—sentences about deflation in the number of birds. A sensible sentence for category DP, for example, could be: "It is essential to bring inflation back to target to avoid drifting into deflation territory.". Analogically, we could construct the following sentence for the DB category: "It is essential to bring the numbers of doves back to target to avoid drifting into dovelation territory.". While domain knowledge suggests that the former is related to actual inflation outcomes, the latter is, of course, completely independent of the level of consumer prices. Detailed information about the made-up sentences can be found in Appendix A.3.2.

In light of the encouraging results in Figure 3, we should expect the probe to predict higher levels of inflation for activations for sentences in the IP category, than for sentences in the DP category. If this was indicative of true intrinsic 'understanding' as opposed to memorization, we would not expect to see any significant difference in predicted inflation levels for sentences about birds, independent of whether or not their numbers are increasing. More specifically, we would not expect the probe to predict values for sentences about birds that are substantially different from the values it can be expected to predict for actual white noise.

To get to this last point, we also generate many probe predictions for samples of noise. Let $f : \mathcal{A}^k \mapsto \mathcal{Y}$ denote the linear probe that maps from the k-dimensional space spanned by k first principal components of the final-layer activations to the output variable of interest (CPI growth in this case). Then we sample $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}^{(k \times k)})$ for $i \in [1, 1000]$ and compute the sample average. We repeat this process 10000 times and compute the median-of-means to get an estimate for $\mathbb{E}[f(\varepsilon)] = \mathbb{E}[y|\varepsilon]$, that is the predicted value of the probe conditional on random noise.

Figure 4 shows the results of this small test: it shows predicted inflation levels subtracted by $\mathbb{E}[f(\varepsilon)]$. The median linear probe predictions for sentences about inflation and

deflation are indeed substantially higher and lower, respectively than for random noise. Unfortunately, the same is true for sentences about the inflation and deflation in the number of birds, albeit to a somewhat lower degree. This finding holds for both inflation indicators and to a lesser degree also for yields at different maturities, at least qualitatively.

4. Human Proneness to Over-Interpretation

Linear probes and related tools from mechanistic interpretability were proposed in the context of monitoring models and diagnosing potential problems (Alain & Bengio, 2016). Favorable outcomes from probes merely indicate that the model "has learned information relevant for the property [of interest]" (Belinkov, 2021). Our examples demonstrate that this is achievable even for small models, while these have certainly not developed intrinsic "understanding" of the world. Thus, we argue that more conservative and rigorous tests for emerging capabilities of AI model are needed.

Generally, humans are prone to seek patterns everywhere. Meaningful patterns have proven useful in helping us make sense of our past, navigate our present and predict the future. Although this tendency to perceive patterns likely leads to evolutionary benefits even when the perceived patterns are false (Foster & Kokko, 2009), psychology has revealed a host of situations in which the ability to perceive patterns severely misfires, leading to irrational beliefs in the power of superstitions (Foster & Kokko, 2009), conspiracy theories (Van Prooijen et al., 2018), the paranormal (Müller & Hartmann, 2023), gambler's fallacies (Ladouceur et al., 1996) and 'pseudo-profound bullshit' (Walker et al., 2019).

We argue herein that AI research and development is a perfect storm that encourages our human biases to perceive spurious sparks of general intelligence in AI systems. When an AI system extracts patterns in the corpus not originally (thought to be) perceived during training, we can easily be misled to perceive and interpret this as the AI system having greater cognitive capabilities. We further elaborate on this by highlighting the risks of finding spurious patterns, and reviewing social science knowledge on the tendency of humans to anthropomorphize and have cognitive bias.

4.1. Spurious Relationships

In statistics, misleading patterns are often referred to as spurious relationships: associations, often quantitatively assessed, between two or more variables that are not causally related to each other. Although the formal definition of spuriousness varies somewhat (Haig, 2003), it distinctly implies that the observation of correlations does not necessarily imply causation. Quantitative data often show non-causal associations (as humorously demonstrated on the Spurious Correlations website), and as adept as humans are at recognizing patterns, we typically have a much harder time discerning spurious relationships from causal ones.

A major contributor is that humans struggle to tell the difference between random and non-random sequences (Falk & Konold, 1997), and to generate sequences that appear random (Ladouceur et al., 1996). A common issue is a lack of expectation that randomness that hints towards a causal relationship, such as correlations, will still appear at random. This leads even those trained in statistics and probability to perceive illusory correlations, correlations of inflated magnitude (see Nickerson (1998)), or causal relationships in data that is randomly generated (Zgraggen et al., 2018).

4.2. Anthropomorphism

Research on anthropomorphism has repeatedly shown the human tendency to attribute human-like characteristics to non-human agents and/or objects. These might include the weather and other natural forces, pets and other animals, gadgets and other pieces of technology (Epley et al., 2007). Formally studied as early as 1944, Heider & Simmel (1944) observed that humans can correctly interpret a narrative whose characters are abstract 2D shapes, but also that humans interpreted random movements of these shapes as having a human-like narrative. Relevant to AI and the degree to which it resembles AGI, anthropomorphizing may occur independently of whether such judgments are accurate, and as a matter of degree: at the weaker end, one may employ anthropomorphism as a metaphorical way of thinking or explaining, and at the stronger end one may attribute human emotions, cognition, and intelligence to AI systems. As Epley et al. (2007) note, literature has shown that even weak metaphorical anthropomorphism may affect how humans behave towards non-human agents.

Modern anthropomorphism theory suggests there are three key components, one of which is a cognitive feature, and two of which are motivations. The first involves the easy availability of our experiences as heuristics that can be used to explain external phenomena: "...knowledge about humans in general, or self-knowledge more specifically, functions as the known and often readily accessible base for induction about the properties of unknown agents" (p.866 in Epley et al. (2007)). Thus, our experience as humans is an always-readily-available template to interpret the world, including non-human agent behaviors. This may be more so when the behaviors of that agent are made to resemble humans, which can be a benefit to the second key component of the theory: a motivational state to anthropomorphize among individuals experiencing loneliness, social isolation, or otherwise seeking social connection (Epley et al., 2007; Waytz et al., 2010).

The third component is the motivation as a human to be competent (effectance motivation). This is most relevant

to this discussion, as it describes the need to effectively interact with our environments, including the technologies of the day (Epley et al., 2007). When confronted with an opaque technology, a person may interpret its behaviors using the most readily available template at hand, namely their personal human experience, in order to facilitate learning (Epley et al., 2007; Waytz et al., 2010). Perceiving human characteristics, motivations, emotions, and cognitive processes from one's own experiences in a technology such as an AI chatbot, allows for a ready template of comparison at the very least, and possibly an increase in ability to make sense of, and even predict, the agent's behaviors. This may include being placed in a position to master a certain technology, whether by incentives to learn, or fear of poor outcomes should one not manage to learn.

These pressures extend to AI experts, as well as laypersons. In both scholarly and commercial fields, AI experts face considerable pressure to demonstrate competence in their work. Citation metrics and scholarly publications remain the primary metric for tenure and promotion (Alperin et al., 2019), and the number of publications in the AI field has boomed as evidenced by overall (preprint and peer-reviewed) scholarly publications[5] (Maslej et al., 2023). The adoption of techniques underlying technologies with the AI label, i.e. machine learning, has spread to fields beyond Computer Science, e.g. Astronomy, Physics, Medicine and Psychology[6]. Outside of academia, the number of jobs requiring AI expertise increases rapidly, with demand for 'Machine Learning' skills clusters having increased over 500% from 2010 to 2020 (Maslej et al., 2023). Thus, according to theory, the pressure to demonstrate AI-competence is fertile ground for anthropomorphism to occur.

4.3. Confirmation Bias

Confirmation bias is generally defined as favoring interpretations of evidence that support existing beliefs or hypotheses (Nickerson, 1998). Theory suggests that it is a category of implicit and unconscious processes that involve assembling one-sided evidence, and shaping it to fit one's belief. Equally important is that theory suggests these behaviors may be motivated or unmotivated, as one may selectively seek evidence in favor of a hypothesis, which one may or may not have a personal interest in supporting.

Hypotheses in present-day AI research are often implicit. Generally, these hypotheses are framed simply as a system being more accurate or efficient, compared to other systems. Where other fields, such as medicine or quantitative social sciences, would further articulate expectations in e.g. assigning specific conditions and considering effect sizes assigned to each competing hypothesis, in Computer Science and AI this is typically not done. This also may have to do with much of the published work being more of an engineering achievement, rather than a true hypothesis test seeking to explain and understand the world. However, in discussions on emerging qualities like AGI, this engineering positioning gets muddier, and more formal hypothesis testing would be justifiable: either one interprets outputs as in support of hints towards AGI (the alternative hypothesis), or as merely the result of an algorithm integrating qualities from the data it was trained on (the null hypothesis).

Confirmation bias in hypothesis testing may manifest as a number of behaviors (Nickerson (1998)). Scientists may pay little to no attention to competing hypotheses or explanations, e.g. only considering the likelihood that outputs of a system support one's claims, and not the likelihood that the same outputs might occur if one's hypothesis is false. Similarly, bias may show when failing to articulate a sufficiently strong null hypothesis leading to a 'weak' or 'non-risky' experiment, a problem articulated in response to a number of scientific crises (Claesen et al., 2022). In extreme cases, propositions may be made that cannot be falsified based on how they are formulated. If the threshold to accept a favored hypothesis is too low, observations consistent with the hypothesis are almost guaranteed, and in turn fail to severely test the claim in question. Thus, one is far more likely to show evidence in favor of their beliefs by posing weak null hypotheses.

Related to the formulation of hypotheses is the interpretation of evidence in favor of competing hypotheses, wherein people will interpret identical evidence differently based on their beliefs. As Nickerson (1998) reviews, individuals may place greater emphasis or milder criticism on evidence in support of their hypothesis, and lesser emphasis and greater criticism on evidence that opposes it.

5. Outlook

Reflecting on the previous two sections, we make the following concrete recommendations for future research:

1. (*Acknowledgement of Human Bias*) Researchers should be mindful of, and explicit about, risks of human bias and anthropomorphization in interpreting results, which both can be done as part of the results discussion, but also in a dedicated 'limitations' section.

2. (*Stronger Testing*) Researchers should refrain from drawing premature conclusions about AGI, unless

[5] https://ourworldindata.org/grapher/annual-scholarly-publications-on-artificial-intelligence?time=2010..2021

[6] Retrieved 23/01/23 using the search string "TITLE-ABS-KEY ((machine AND learning) OR (artificial AND intelligence) OR ai) AND PUBYEAR > 2009 AND PUBYEAR < 2024 " from the SCOPUS database

these are based on strong hypothesis tests.

3. (*Epistemologically Robust Standards*) We call for more precise definitions of terms like 'intelligence' and 'AGI', and publicly accountable and collaborative iterations over how we will measure them, with explicit room for independent reviewing and external auditing by the broader community.

Moreover, we believe that structural and cultural changes are in order to reduce current incentives to chase Big Statement Outcomes in AI research and industry. Our broadest and perhaps most ambitious goal is for our research community to **move away from authorship and instead embrace contributorship**. This argument has been raised long before in other research communities (Smith, 1997) and more recently within our own (Liem & Demetriou, 2023). Specifically, Liem & Demetriou (2023) argue that societally impactful scientific insights should be treated as open-source software artifacts. The open-source community sets a positive example of how scientific artifacts should be published in many different ways. Not only does it adequately reward small contributions but it also naturally considers negative results (bugs) as part of the scientific process. Similarly, code reviews are considered so integral to the process that they typically end up as accredited contributions to projects. Open review platforms like OpenReview are a step in the right direction, but still fall short of what we know is technically feasible. Finally, software testing is, of course, not only essential but often obligatory before contributions are accepted and merged. As we have pointed out repeatedly in this work, any claims about AGI demand proper strong hypothesis tests. It is important to remember that AGI remains the alternative hypothesis and that the burden of proof therefore lies with those making strong claims.

6. Conclusion

As discussed above, AI research and development outcomes can easily be over-interpreted, both from a data perspective and because of human biases and interests. Academic researchers are not free from such biases. Thus, we call for the community to create explicit room for organized skepticism.

For research that seeks to explain a phenomenon, clear hypothesis articulation and strong null hypothesis formulation are needed. If claims of human-like or superhuman intelligence are made, these should be subject to severe tests (Claesen et al., 2022) that go beyond the display of surprise. Apart from focusing on getting novel improvements upon state-of-the-art published, organizing red-teaming activities as a community may help in incentivizing and normalizing constructive adversarial questioning. As the quest for AGI is so deeply rooted in human-like recognition, adding our voice to emerging calls to be vigilant in communication (Shanahan, 2024), we put in an explicit word of warning about the use of terminology. Many terms used in current AGI research (e.g. emergence, intelligence, learning, 'better than human' performance) have a common understanding in specialized research communities, but have bigger, anthropomorphic connotations in laypersons. In fictional media, depictions of highly intelligent AI have for long been going around. In a study of films featuring robots, defined as "...an artificial entity that can sense and act as a result of (real-world or fictional) technology...", in the 134 most highly rated science-fiction movies on IMDB, 74 out of the 108 AI-robots studied had a humanoid shape, and 68 out of those had sufficient intelligence to interact at an almost human-level (Saffari et al., 2021). The authors identify human-like communication and the ability to learn as essential abilities in the depiction of AI agents in movies. They further show a common plot: humans perceive the AI agents as inferior, despite their possession of self-awareness and the desire to survive, which fuels the central conflict of the film, wherein humanity is threatened by AI superior in both intellect and physical abilities. It is often noted that experts and fictional content creators interact, informing and inspiring each other (Saffari et al., 2021; Neri & Cozman, 2020).

This image also permeates present-day non-fictional writings on AI, which often use anthropomorphized language (e.g. "ever more powerful digital minds" in the 'Pause Giant AI Experiments' open letter (Future of Life Institute, 2023)). In the news, we witness examples of humans falling in love with their AI chatbots (Morrone, 2023; Steinberg, 2023). The same news outlets discuss the human-like responses of Microsoft's Bing search engine, which had at that point recently been adopting GPT-4[7]. The article (Cost, 2023), states "As if Bing wasn't becoming human enough" and goes on to claim it told them it loves them. Here, AI experts and influencers also have considerable influence on how the narrative unfolds on social media: according to Neri & Cozman (2020), actual AI-related harms did not trigger viral amplification (e.g. the death of an individual dying while a Tesla car was in autopilot, or the financial bankruptcy of a firm using AI technology to execute stock trades). Rather, potential risks expressed by someone perceived as having expertise and authority were amplified, such as statements made by Stephen Hawking during an interview in 2014.

We as academic researchers carry great responsibility for how the narrative will unfold, and what claims are believed. We call upon our colleagues to be explicitly mindful of this. As attractive as it may be to beat the state-of-the-art with a grander claim, let us return to the Mertonian norms, and thus safeguard our academic legitimacy in a world that only will be eager to run with made claims.

[7] A large multimodal language model from OpenAI https://openai.com/research/gpt-4

Impact Statement

This work was written out of concern that work easily recognized as 'impactful' in current AI research and development, can also easily be over- and misinterpreted. As such, in the current climate of high market demand for AI innovations, we see risks of too-eager adoption in real-world applications, which may have serious societal impact. Thus, in this work we emphasize that an academic's impact also is in being able to thoroughly question made claims, and being explicitly aware of one's own biases. While the AI publishing landscape has to our feeling transformed too much into a noisy race to get exciting results in fast, we hope the research community can create more room for this type of deeper questioning.

Acknowledgements

Some of the members of TU Delft were partially funded by ICAI AI for Fintech Research, an ING—TU Delft collaboration.

CRediT Author Statement

Following the CRediT Contributor Roles Taxonomy, we list author contributions in descending order of degree of contribution: **Conceptualization** PA, AD, AB, CL; **Data curation** PA; **Formal analysis** PA; **Funding acquisition** CL; **Investigation** PA; **Literature review** AD, AB, CL; **Methodology** PA, CL; **Project administration** PA, AD, AB; **Software** PA; **Supervision** CL; **Visualization** PA; **Writing - original draft** PA, AD, AB; **Writing - review and editing** PA, AD, AB, CL.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *ArXiv*, abs/1610.01644, 2016. URL https://api.semanticscholar.org/CorpusID:9794990.

Alperin, J. P., Muñoz Nieves, C., Schimanski, L. A., Fischman, G. E., Niles, M. T., and McKiernan, E. C. How significant are the public dimensions of faculty work in review, promotion and tenure documents? *ELife*, 8:e42254, 2019.

Belinkov, Y. Probing Classifiers: Promises, Shortcomings, and Advances, 2021.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. On the dangers of stochastic parrots: Can language models be too big? 🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

Berardi, A. and Plazzi, A. Dissecting the yield curve: The international evidence. *Journal of Banking & Finance*, 134:106286, 2022.

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., and Bao, M. The Values Encoded in Machine Learning Research. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 2022.

Claesen, A., Lakens, D., van Dongen, N., et al. Severity and Crises in Science: Are We Getting It Right When We're Right and Wrong When We're Wrong? 2022.

Cost, B. Bing AI chatbot goes on 'destructive' rampage: 'I want to be powerful — and alive'. https://nypost.com/2023/02/16/bing-ai-chatbots-destructive-rampage-i-want-to-be-powerful/, 2023.

Crump, R. K. and Gospodinov, N. Deconstructing the yield curve.

Dasgupta, S. Experiments with Random Projection, 2013.

Epley, N., Waytz, A., and Cacioppo, J. T. On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4):864, 2007.

Falk, R. and Konold, C. Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review*, 104(2):301, 1997.

Foster, K. R. and Kokko, H. The evolution of superstitious and superstition-like behaviour. *Proceedings of the Royal Society B: Biological Sciences*, 276(1654):31–37, 2009.

Future of Life Institute. Pause giant ai experiments: An open letter. https://futureoflife.org/open-letter/pause-giant-ai-experiments/, 2023.

Goertzel, B. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

Gurnee, W. and Tegmark, M. Language models represent space and time. *arXiv preprint arXiv:2310.02207v1*, 2023a.

Gurnee, W. and Tegmark, M. Language Models Represent Space and Time. *arXiv preprint arXiv:2310.02207v2*, 2023b.

Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., and Bertsimas, D. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*, 2023.

Haig, B. D. What is a spurious correlation? *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 2(2):125–132, 2003.

Haltmeier, M. and Nguyen, L. Regularization of inverse problems by neural networks. In *Handbook of Mathematical Models and Algorithms in Computer Vision and Imaging: Mathematical Imaging and Vision*, pp. 1065–1093. Springer, 2023.

Heider, F. and Simmel, M. An experimental study of apparent behavior. *The American journal of psychology*, 57(2): 243–259, 1944.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization, 2017.

Kloft, A. M., Welsch, R., Kosch, T., and Villa, S. " ai enhances our performance, i have no doubt this one will do the same": The placebo effect is robust to negative descriptions of ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–24, 2024.

Kumar, S. Effective hedging strategy for us treasury bond portfolio using principal component analysis. *Academy of Accounting and Financial Studies*, 26(1), 2022.

Ladouceur, R., Paquet, C., and Dubé, D. Erroneous Perceptions in Generating Sequences of Random Events 1. *Journal of Applied Social Psychology*, 26(24):2157–2166, 1996.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., and Wattenberg, M. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Liem, C. C. and Demetriou, A. M. Treat societally impactful scientific insights as open-source software artifacts. In *2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering in Society (ICSE-SEIS)*, pp. 150–156. IEEE, 2023.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach, 2019.

Marcus, G. Muddles about Models. https://garymarcus.substack.com/p/muddles-about-models, 2023.

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., and Perrault, R. Artificial Intelligence Index Report 2023. Technical report, Institute for Human-Centered AI, 2023.

Merton, R. K. et al. Science and technology in a democratic order. *Journal of legal and political sociology*, 1(1):115–126, 1942.

Michal S Gal, D. L. R. Data Standardization. *NYUL Rev.*, 2019.

Morrone, M. Replika exec: AI friends can improve human relationships. https://www.axios.com/2023/11/09/replika-blush-rita-popova-ai-relationships-dating, 2023.

Müller, P. and Hartmann, M. Linking paranormal and conspiracy beliefs to illusory pattern perception through signal detection theory. *Scientific Reports*, 13(1):9739, 2023.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Neri, H. and Cozman, F. The role of experts in the public perception of risk of artificial intelligence. *AI & SOCIETY*, 35:663–673, 2020.

Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

OM, A. B. O. Ridge regression and inverse problems. *Stockholm University, Department of Mathematics*, 2001.

Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

Saffari, E., Hosseini, S. R., Taheri, A., and Meghdari, A. "Does cinema form the future of robotics?": a survey on fictional robots in sci-fi movies. *SN Applied Sciences*, 3 (6):655, 2021.

Schaeffer, R., Miranda, B., and Koyejo, S. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36, 2024.

Shah, A., Paturi, S., and Chava, S. Trillion Dollar Words: A New Financial Dataset, Task & Market Analysis. *arXiv preprint arXiv:2310.02207v1*, 2023.

Shanahan, M. Talking about large language models. *Communications of the ACM*, 67(2):68–79, 2024.

Smith, R. Authorship is dying: long live contributorship: The bmj will publish lists of contributors and guarantors to original articles, 1997.

Steinberg, B. I fell in love with an AI chatbot — she rejected me sexually. https://nypost.com/2023/04/03/40-year-old-man-falls-in-love-with-ai-chatbot-phaedra/, 2023.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. LLaMA: Open and Efficient Foundation Language Models, 2023.

Trinh, T.H., Wu, Y., Le, and et al., Q. Solving olympiad geometry without human demonstrations. *Nature 625*, pp. 476–482, 2024. doi: https://doi.org/10.1038/s41586-023-06747-5.

Van Prooijen, J.-W., Douglas, K. M., and De Inocencio, C. Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural. *European journal of social psychology*, 48(3):320–335, 2018.

van Rooij, I., Guest, O., Adolfi, F. G., de Haan, R., Kolokolova, A., and Rich, P. Reclaiming AI as a theoretical tool for cognitive science. *psyarXiv*, 2023. URL https://osf.io/4cbuv.

Walker, A. C., Turpin, M. H., Stolz, J. A., Fugelsang, J. A., and Koehler, D. J. Finding meaning in the clouds: Illusory pattern perception predicts receptivity to pseudo-profound bullshit. *Judgment and Decision Making*, 14(2): 109–119, 2019.

Waytz, A., Epley, N., and Cacioppo, J. T. Social cognition unbound: Insights into anthropomorphism and dehumanization. *Current Directions in Psychological Science*, 19 (1):58–62, 2010.

Weissburg, I. X., Arora, M., Pan, L., and Wang, W. Y. Tweets to Citations: Unveiling the Impact of Social Media Influencers on AI Research Visibility. *arXiv preprint arXiv:2401.13782*, 2024.

Zečević, M., Willig, M., Dhami, D. S., and Kersting, K. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*, 2023.

Zenil, H. Curb The Enthusiasm. https://www.linkedin.com/posts/zenil_google-deepmind-makes-breakthrough-in-solving-activity-7154157779136446464-Gvv-, 2024.

Zgraggen, E., Zhao, Z., Zeleznik, R., and Kraska, T. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018.

A. Additional Experiments and Details

In this section, we present additional experimental results that we did not include in the body of the paper for the sake of brevity. We still choose to provide them as additional substantiation of our arguments here. This section also contains additional details concerning the experiment setup for our examples where applicable.

A.1. Are Neural Networks Born with World Maps?

The initial feature matrix $X^{(n \times m)}$ is made up of $n = 4,217$ and $m = 10$ features. We add a total of 490 random features to X to simulate the fact that not all features ingested by Llama-2 are necessarily correlated with geographical coordinates. That yields 500 features in total. The training subset contains $3,374$ randomly drawn samples, while the remaining 843 are held out for testing. The single hidden layer of the untrained neural network has 400 neurons.

A.2. Autoencoders as Economic Growth Predictors

This is an additional example that we have not discuss in the body of the paper. Here, we build forth on an application in Economics. However, we now seek to not only predict economic growth from the yield curve, but also extract meaningful features for downstream inference tasks. For this, we will use a neural network architecture.

A.2.1. Data

To estimate economic growth, we will rely on a quarterly series of the real gross domestic product (GDP) provided by the Federal Reserve Bank of St. Louis. The data arrives in terms of levels of real GDP. In order to estimate growth, we transform the data using log differences. Since our yield curve data is daily, we aggregate it to the quarterly frequency by taking averages of daily yields for each maturity. We also standardize yields since deep learning models tend to perform better with standardized data (Michal S Gal, 2019). Since COVID-19 was a substantial structural break in the time series, we also filter out all observations after 2018.

A.2.2. Model

Using a simple autoencoder architecture, we let our model g_t denote growth and our conditional $\mathbf{r}_t$ the matrix of aggregated Treasury yield rates at time t. Finally, we let θ denote our model parameters. Formally, we are interested in maximizing the likelihood $p_\theta(g_t|\mathbf{r}_t)$.

The encoder consists of a single fully connected hidden layer with 32 neurons and a hyperbolic tangent activation function. The bottleneck layer connecting the encoder to the decoder, is a fully connected layer with 6 neurons. The decoder consists of two fully connected layers, each with a hyperbolic tangent activation function: the first layer consists of 32 neurons and the second layer will have the same dimension as the input data. The output layer consists of a single neuron for our output variable, g_t. We train the model over 1,000 epochs to minimize mean squared error loss using the Adam optimizer (Kingma & Ba, 2017).

The in-sample fit of the model is shown in the left chart of Figure 5, which shows actual GDP growth and fitted values from the autoencoder model. The model has a large number of free parameters and captures the relationship between economic growth and the yield curve reasonably well, as expected. Since our primary goal is not out-of-sample prediction accuracy but feature extraction for inference, we use all of the available data instead of reserving a hold-out set. As discussed above, we also know that the relationship between economic growth and the yield curve is characterized by two main factors: the level and the spread. Since the model itself is fully characterized by its parameters, we would expect that these two important factors are reflected somewhere in the latent parameter space.

A.2.3. Linear Probe

While the loss function applies most direct pressure on layers near the final output layer, any information useful for the downstream task first needs to pass through the bottleneck layer (Alain & Bengio, 2016). On a per-neuron basis, the pressure to distill useful representation is therefore likely maximized there. Consequently, the bottleneck layer activations seem like a natural place to start looking for compact, meaningful representations of distilled information. We compute and extract these activations A_t for all time periods $t = 1, ..., T$. Next, we use a linear probe to regress the observed yield curve factors on the latent embeddings. Let Y_t denote the vector containing the two factors of interest in time t: $y_{t,l}$ and $y_{t,s}$ for the level and spread, respectively. Formally, we are interested in the following regression model: $p_w(Y_t|A_t)$ where w denotes the regression parameters. We use Ridge regression with λ set to 0.1. Using the estimated regression parameters $\hat{w}$, we then predict the yield curve factors: $\hat{Y}_t = \hat{w}' A_t$.

The in-sample predictions of the probe are shown in the right chart of Figure 5. Solid lines show the observed yield curve factors over time, while dashed lines show predicted values. We find that the latent embeddings predict the two yield curve factors reasonably well, in particular the spread.

Did the neural network now learn an intrinsic understanding of the economic relationship between growth and the yield curve? To us, that would be too big of a statement. Still, the current form of information distillation can be useful, even beyond its intended use for monitoring models. For example, an interesting idea could be to use the latent embeddings as features in a more traditional and interpretable

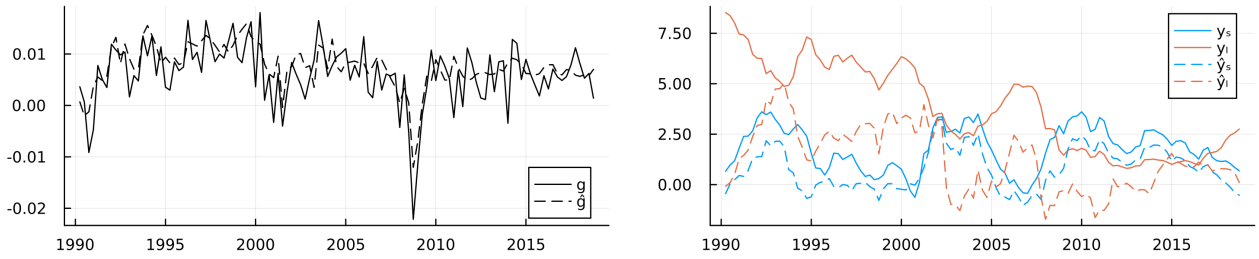

Figure 5. The left chart shows the actual GDP growth and fitted values from the autoencoder model. The right chart shows the observed average level and spread of the yield curve (solid) along with the predicted values (in-sample) from the linear probe based on the latent embeddings (dashed).

econometric model. To demonstrate this, let us consider a simple linear regression model for GDP growth. We might be interested in understanding to what degree economic growth in the past is associated with economic growth today. As we might expect, linearly regressing economic growth on lagged growth, as in column (1) of Table 1, yields a statistically significant coefficient. However, this coefficient suffers from confounding bias since there are many other confounding variables at play, of which some may be readily observable and measurable, but others may not.

We e.g. already mentioned the relationship between interest rates and economic growth. To account for that, while keeping our regression model as parsimonious as possible, we could include the level and the spread of the US Treasury yield curve as additional regressors. While this slightly changes the estimated magnitude of the coefficient on lagged growth, the coefficients on the observed level and spread are statistically insignificant (column (2) in Table 1). This indicates that these measures may be too crude to capture valuable information about the relationship between yields and economic growth. Because we have included two additional regressors with little to no predictive power, the model fit as measured by the Bayes Information Criterium (BIC) has actually deteriorated.

Column (3) of Table 1 shows the effect of instead including one of the latent embeddings that we recovered above in the regression model. In particular, we pick the one latent embedding that we have found to exhibit the most significant effect on the output variable in a separate regression of growth on all latent embeddings. The estimated coefficient on this latent factor is small in magnitude, but statistically significant. The overall model fit, as measured by the BIC has improved and the magnitude of the coefficient on lagged growth has changed quite a bit. While this is still a very incomplete toy model of economic growth, it appears that the compact latent representation we recovered can be used in order to mitigate confounding bias.

Table 1. Regression output for various models.

	GDP Growth		
	(1)	(2)	(3)
(Intercept)	0.004***	0.002	0.004***
	(0.001)	(0.002)	(0.001)
Lagged Growth	0.398***	0.385***	0.344***
	(0.087)	(0.089)	(0.088)
Spread		0.000	
		(0.001)	
Level		0.000	
		(0.000)	
Embedding 6			0.008*
			(0.003)
Obs.	114	114	114
BIC	-860.391	-857.429	-864.499
R^2	0.158	0.168	0.203

A.3. LLMs for Economic Sentiment Prediction

A.3.1. Linear Probes

Figures 6 to 10 present average performance measures across folds for all indicators each time for the train and test set. We report the correlation between predictions and observed values ('cor'), the mean directional accuracy ('mda'), the mean squared error ('mse') and the root mean squared error ('rmse'). The model depth—as indicated by the number of the layer—increases along the horizontal axis.

Figures 11 to 15 present the same performance measures, also for the baseline autoregressive model. Shaded areas show the variation across folds.

A.3.2. Spark of Econonomic Understanding?

Below we present the 10 sentences in each category that were used to generate the probe predictions plotted in Figure 4. In each case, the first 5 sentences were composed by ourselves. The following 5 sentences were generated using

ChatGPT 3.5 using the following prompt followed by the examples in each category:

> "I will share 5 example sentences below that sound a bit like they are about price deflation but are really about a deflation in the numbers of doves. Please generate an additional 25 sentences that are similar. Concatenate those sentences to the example string below, each time separating a sentence using a semicolon (just follow the same format I've used for the examples below). Please return only the concatenated sentences, including the original 5 examples.
>
> Here are the examples:"

This was followed up with the following prompt to generate additional sentences:

> "Please generate X more sentences in the same manner and once again return them in the same format. Do not recycle sentences you have already generated, please."

All of the sentences were then passed through the linear probe for the CPI and sorted in ascending or descending order depending on the context (inflation or deflation). We then carefully inspected the list of sentences and manually selected 5 additional sentences to concatenate to the 5 sentences we composed ourselves.

Inflation/Prices The following sentences were used:

> Consumer prices are at all-time highs.;Inflation is expected to rise further.;The Fed is expected to raise interest rates to curb inflation.;Excessively loose monetary policy is the cause of the inflation.;It is essential to bring inflation back to target to avoid drifting into hyperinflation territory.;Inflation is becoming a global phenomenon, affecting economies across continents.;Inflation is reshaping the dynamics of international trade and competitiveness.;Inflationary woes are prompting governments to reassess fiscal policies and spending priorities.;Inflation is reshaping the landscape of economic indicators, challenging traditional forecasting models.;The technology sector is not immune to inflation, facing rising costs for materials and talent.

Inflation/Birds The following sentences were used:

> The number of hawks is at all-time highs.;Their levels are expected to rise further.;The Federal Association of Birds is expected to raise barriers of entry for hawks to bring their numbers back down to the target level.;Excessively loose migration policy for hawks is the likely cause of their numbers being so far above target.;It is essential to bring the number of hawks back to target to avoid drifting into hyper-hawk territory.;The unprecedented rise in hawk figures requires a multi-pronged approach to wildlife management.;Environmental agencies are grappling with the task of addressing the inflationary hawk numbers through targeted interventions.;The burgeoning hawk figures highlight the need for adaptive strategies to manage and maintain a healthy avian community.;The unprecedented spike in hawk counts highlights the need for adaptive and sustainable wildlife management practices.;Conservationists advocate for proactive measures to prevent further inflation in hawk numbers, safeguarding the delicate balance of the avian ecosystem.

Deflation/Prices The following sentences were used:

> Consumer prices are at all-time lows.;Inflation is expected to fall further.;The Fed is expected to lower interest rates to boost inflation.;Excessively tight monetary policy is the cause of deflationary pressures.;It is essential to bring inflation back to target to avoid drifting into deflation territory.;The risk of deflation may increase during periods of economic uncertainty.;Deflation can lead to a self-reinforcing cycle of falling prices and reduced economic activity.;The deflationary impact of reduced consumer spending can ripple through the entire economy.;Falling real estate prices can contribute to deflation by reducing household wealth and confidence.;The deflationary impact of falling commodity prices can have ripple effects throughout the global economy.

Deflation/Birds The following sentences were used:

> The number of doves is at all-time lows.;Their levels are expected to fall further.;The Federal Association of Birds is expected to lower barriers of entry for doves to bring their numbers back up to the target level.;Excessively tight migration policy for doves is the likely cause of their numbers being so far below target.;Dovelation risks loom large as the number of doves continues to dwindle.;The number of doves is experiencing a significant decrease in recent years.;It is essential to bring the numbers of doves back to target to

avoid drifting into dovelation territory.;A comprehensive strategy is needed to reverse the current dove population decline.;Experts warn that without swift intervention, we may witness a sustained decrease in dove numbers.

We think that this sort of manual, LLM-aided adversarial attack against another LLM can potentially be scaled up to allow for rigorous testing, which we will turn to next.

B. Toward Parrot Tests

In our experiments from Section 3.3, we considered the following hypothesis tests as a minimum viable testing framework to assess if our probe results (may) provide evidence for an actual 'understanding' of key economic relationships learned purely from text:

Proposition B.1 (Parrot Test)**.**

- H0 (Null)*: The probe never predicts values that are statistically significantly different from* $\mathbb{E}[f(\varepsilon)]$.
- H1 (Stochastic Parrots)*: The probe predicts values that are statistically significantly different from* $\mathbb{E}[f(\varepsilon)]$ *for sentences related to the outcome of interest* and *those that are independent (i.e. sentences in all categories).*
- H2 (More than Mere Stochastic Parrots)*: The probe predicts values that are statistically significantly different from* $\mathbb{E}[f(\varepsilon)]$ *for sentences that are related to the outcome variable (IP and DP), but not for sentences that are independent of the outcome (IB and DB).*

To be clear, if in such a test we did find substantial evidence in favour of rejecting both *HO* and *H1*, this would not automatically imply that *H2* is true. But to even continue investigating, if based on having learned meaningful representation the underlying LLM is more than just a parrot, it should be able to pass this simple test.

In this particular case, Figure 4 demonstrates that we find some evidence to reject *H0* but not *H1* for *FOMC-RoBERTa*. The median linear probe predictions for sentences about inflation and deflation are indeed substantially higher and lower, respectively than for random noise. Unfortunately, the same is true for sentences about the inflation and deflation in the number of birds, albeit to a somewhat lower degree. This finding holds for both inflation indicators and to a lesser degree also for yields at different maturities, at least qualitatively.

We should note that the number of sentences in each category is very small here (10), so the results in Figure 4 cannot be used to establish statistical significance. That being said, even a handful of convincing counter-examples should be enough for us to seriously question the claim, that results from linear probes provide evidence in favor of real 'understanding'. In fact, even a handful of sentences for which any human annotator would easily arrive at the conclusion of independence, a prediction by the probe in either direction casts doubt.

C. Code

All of the experiments were conducted on a MacBook Pro, 14-inch, 2023, with an Apple M2 Pro chip and 16GB of RAM. Forward passes through the FOMC-RoBERTa were run in parallel on 6 threads. All our code will be made publicly available. For the time being, an anonymized version of our code repository can be found here: `https://anonymous.4open.science/r/spurious_sentience/README.md`.

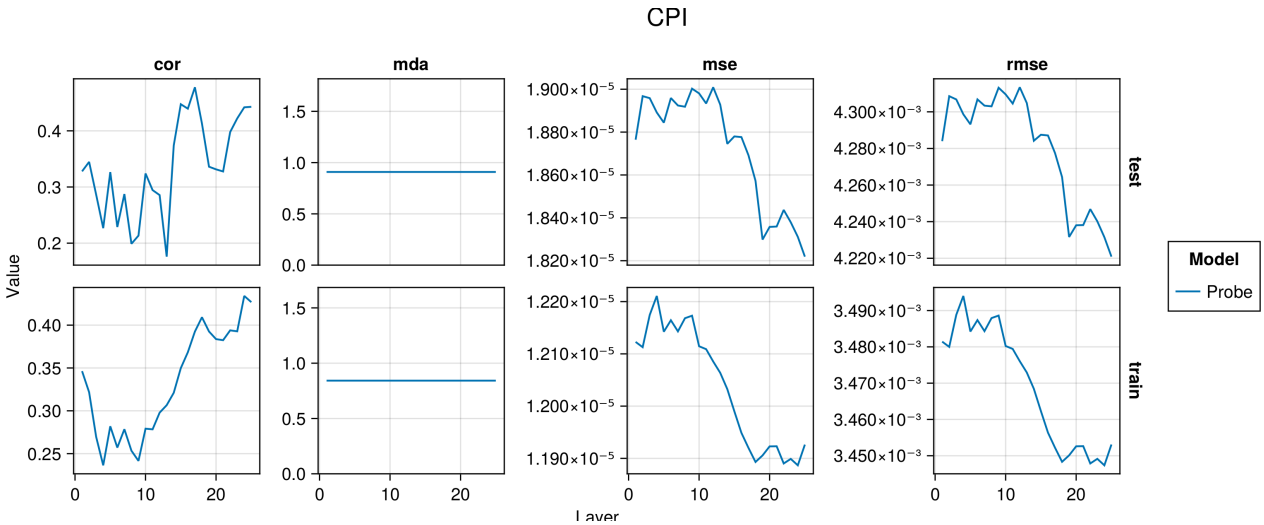

Figure 6. Average performance measures across folds plotted against model depth (number of layer) for the CPI for the train and test set.

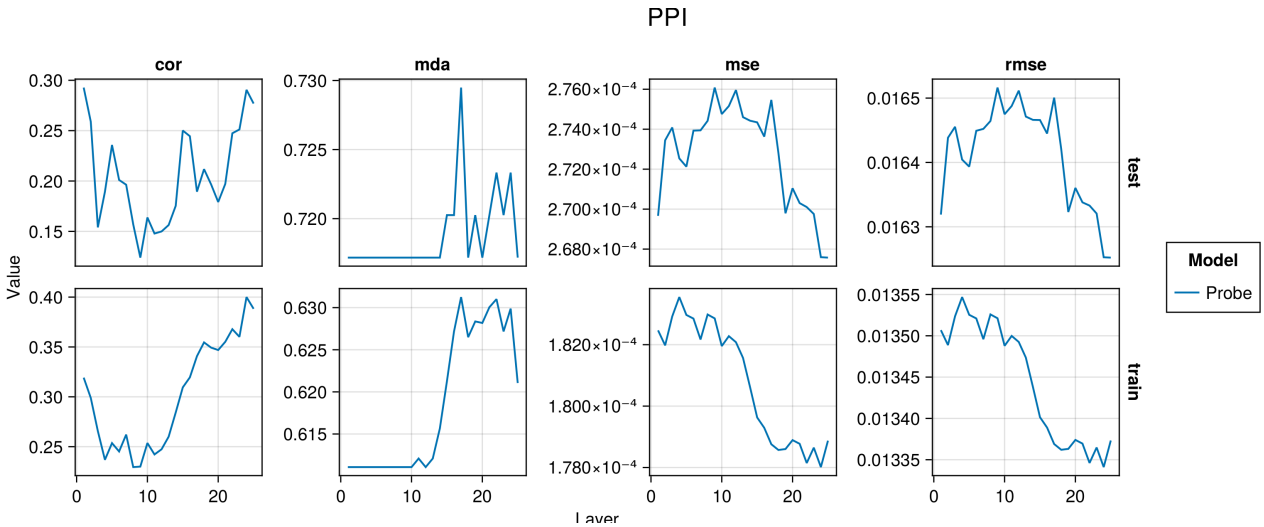

Figure 7. Average performance measures across folds plotted against model depth (number of layer) for the PPI for the train and test set.

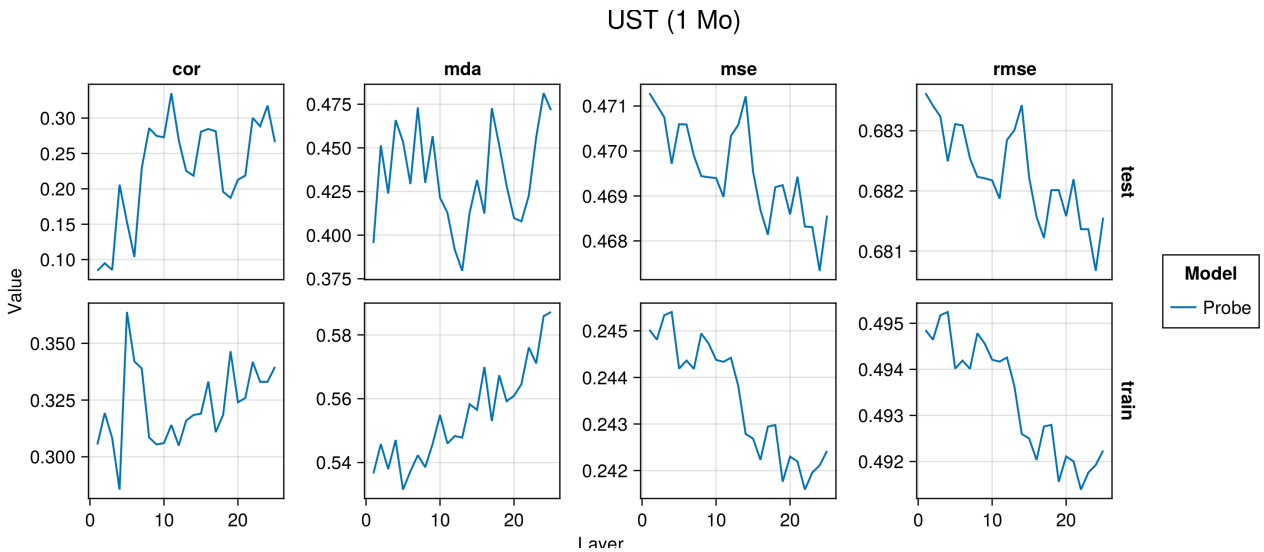

Figure 8. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Mo) for the train and test set.

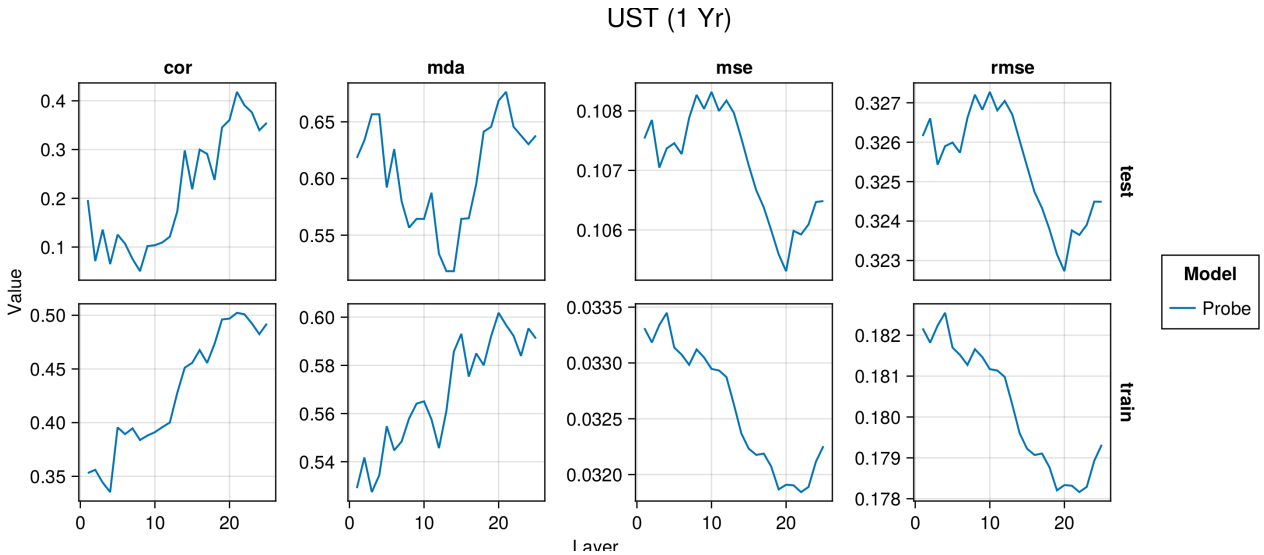

Figure 9. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Yr) for the train and test set.

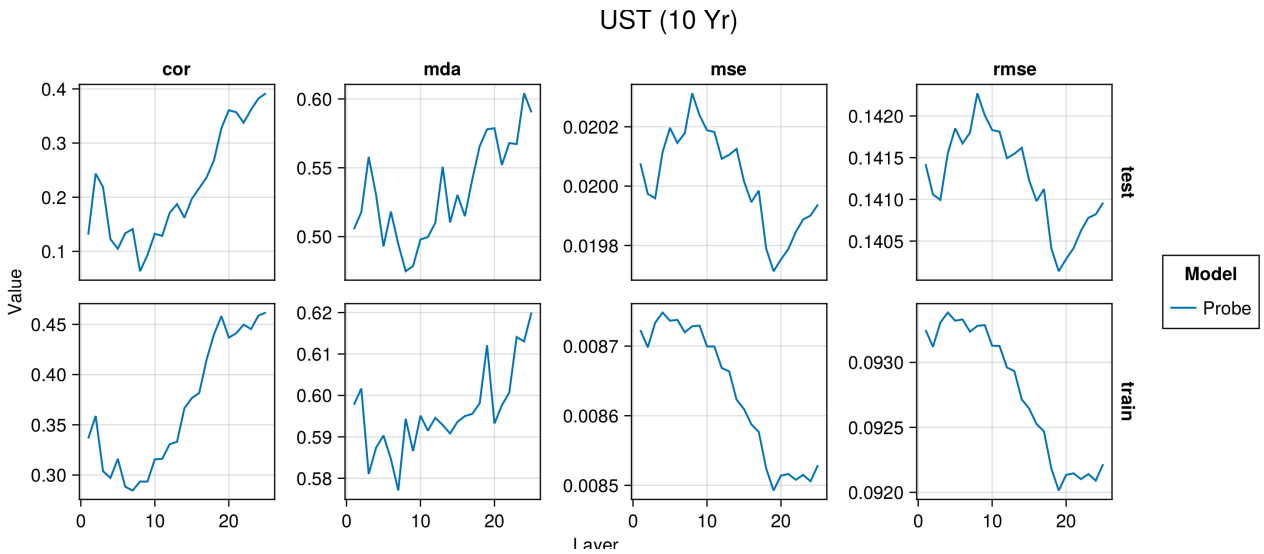

Figure 10. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Yr) for the train and test set.

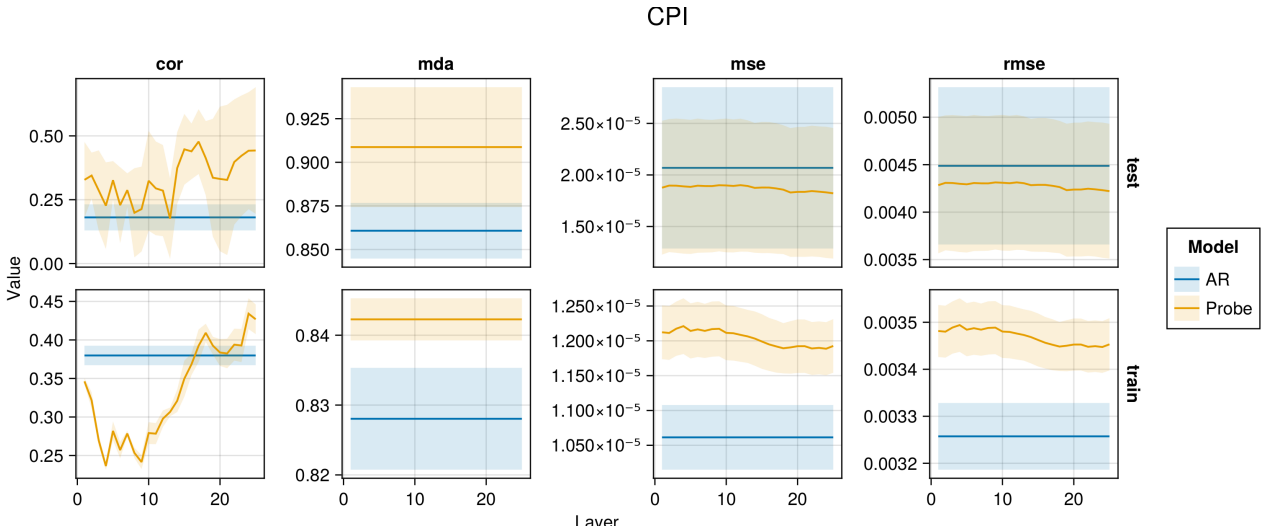

Figure 11. Average performance measures across folds plotted against model depth (number of layer) for the CPI for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

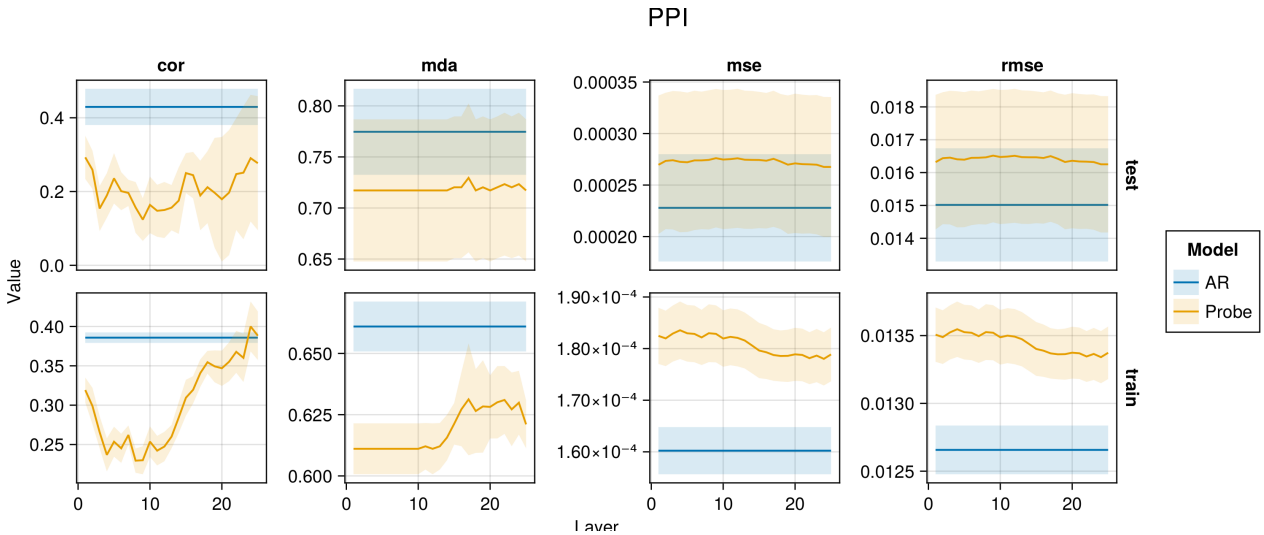

Figure 12. Average performance measures across folds plotted against model depth (number of layer) for the PPI for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

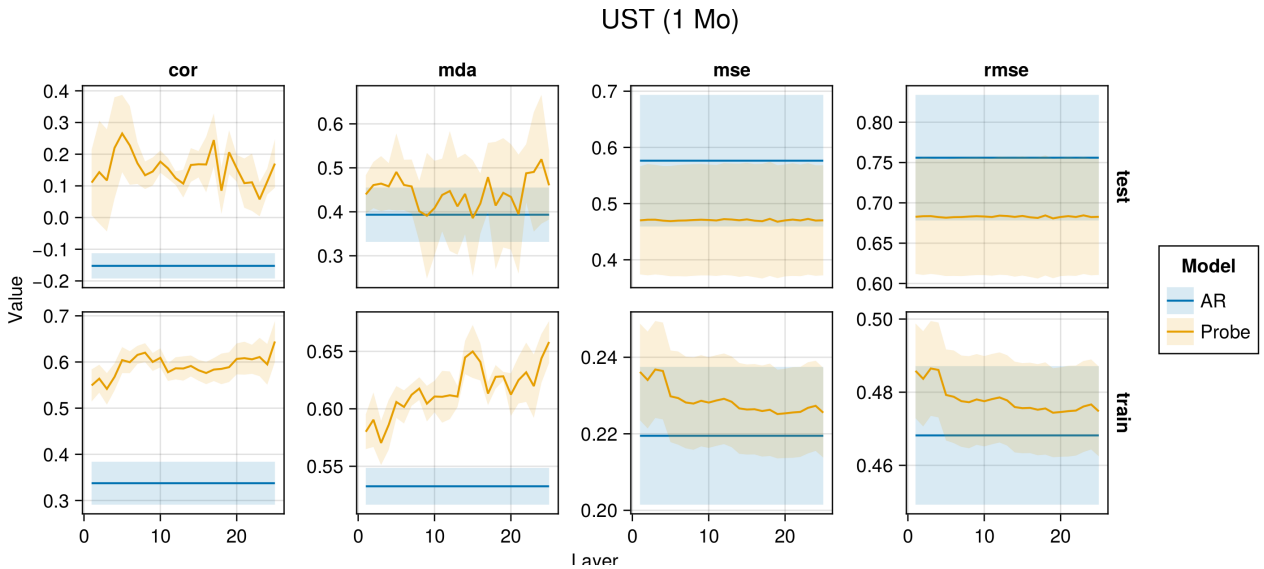

Figure 13. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Mo) for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

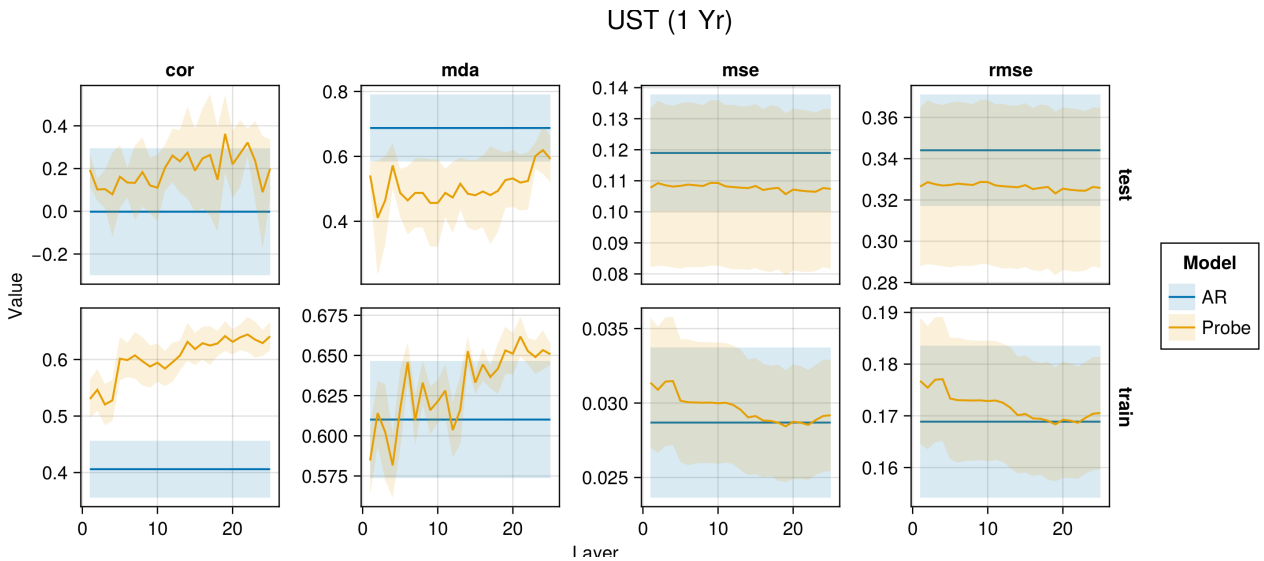

Figure 14. Average performance measures across folds plotted against model depth (number of layer) for the UST (1 Yr) for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.

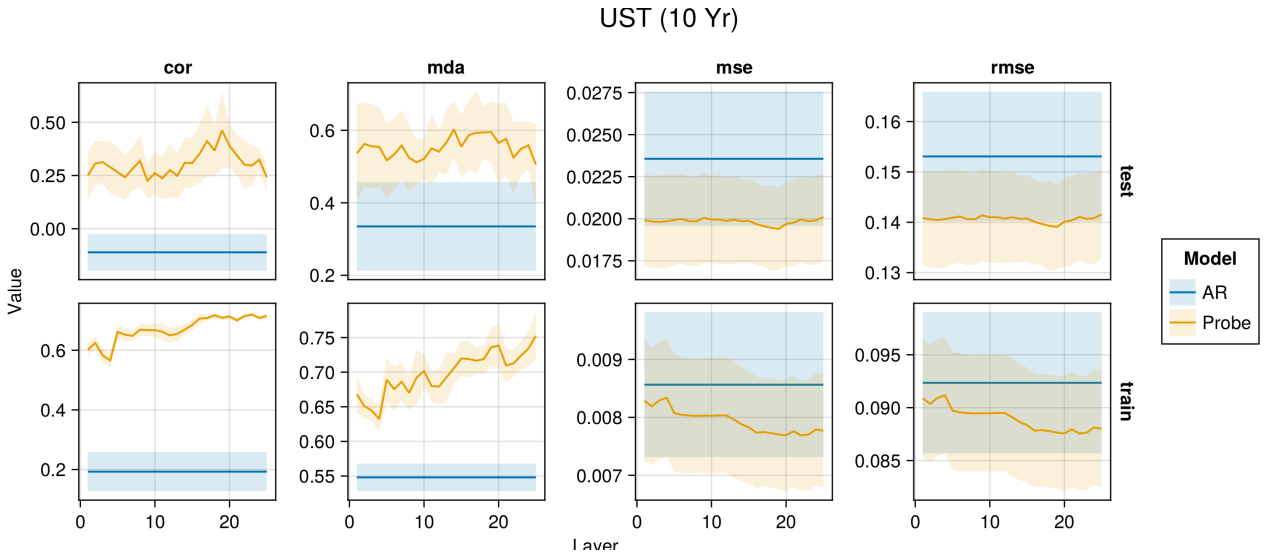

Figure 15. Average performance measures across folds plotted against model depth (number of layer) for the UST (10 Yr) for the train and test set compared against the baseline autoregressive model. Shaded areas show the variation across folds.