Resolute Resolutions (a blog), volume 5:

The Determinants of Controllable AGI

March 2025

Allen Schmaltz
allen@re.express

Abstract

We briefly introduce, at a conceptual level, technical work for deriving robust estimators of the predictive uncertainty over large language models (LLMs), and we consider the implications for real-world deployments and AI policy.

The Foundational LLM Limitation: An Absence of Robust Estimators of Predictive Uncertainty

The limitations of unconstrained LLMs, which includes the more recent RL-based reasoning-token models, are readily evident to end-users. Hallucinations, highly-confident wrong answers, and related issues diminish their benefits in most real-world settings. The punchline is that the end-user has no means of knowing whether the output can be trusted, beyond carefully checking the output, which precludes model-based automation for most complex, multi-stage pipelines.

The foundational problem for all of the large models with non-identifiable parameters is that there has been an absence of meaningfully reliable estimators of the predictive uncertainty over the high-dimensional outputs. It is a very challenging problem, because the parameters of the models are non-identifiable [1, inter alia][1]; the parameter counts are typically in the billions or more; and given typical use-cases of LLMs, we seek approximately conditional (rather than marginal) estimates. This latter point is subtle but critical. Most benchmarks and evaluations for LLMs provide scores or metrics (e.g., overall accuracy) that are averaged over entire test sets. Such evaluations can be useful for broad comparisons and winnowing of models (i.e., at the unit-of-analysis of models), but they are insufficiently informative to serve as estimators at the instance level for end-users over complex, high-dimensional outputs—especially those for which distribution shifts are the norm.

Conversely, with a reliable and robust estimator of the predictive uncertainty that has a human interpretable mapping to applicable partitions of the observed data, the primary technical limitations of LLMs (and AI models, more generally) would be effectively resolved. When such an estimator is uncertain, the model could then refrain from a final decision and take another branching action. While the classification-with-reject-option is sometimes viewed as too coarse of a final quantity of interest—and perhaps an unsatisfying denouement to the uncertainty endeavor—it is in fact exactly what is needed with expressive neural networks that can re-cross-encode input and output pairs; take conditional test-time actions based on sub-problems; and incorporate information exogenous to the model itself, as via retrieval. In other words, with neural networks, a rejected prediction need not immediately default to human adjudication (i.e., telling the user the model cannot be used); instead, we can build multi-stage pipelines searching across data sources, problem formulations, and even input modalities. However, such pipelines are not feasible without robust estimators of predictive uncertainty.

[1] Informally, this means that two or more distinct sets of values for the parameters can result in identical output distributions. As a consequence, interpreting the parameters of such models is typically much more complicated than with a simple linear regression model, for example.

Accounting for Epistemic Uncertainty in High-Dimensions

Unfortunately, most classes of estimators of predictive uncertainty are not suitable for LLM settings, even if they provide statistical assurances in theory. The reasons are multi-fold. As noted above, we seek estimators that are approximately conditional rather than marginal. Given the high-dimensional, complex models with non-identifiable parameters, we seek to minimize distributional assumptions. Additionally, we seek quantities of interest that are understandable to non-expert end-users. These issues are known to researchers of uncertainty quantification in statistical machine learning. However, there is a yet more fundamental problem that nullifies the usefulness of approaches that address some of the previously mentioned issues, and renders existing estimators inadequate for LLMs, which is epistemologically evident given the absence of reliable predictive uncertainty over LLMs despite significant work from Bayesian, Frequentist, and empirical perspectives: An inadequate control over the epistemic (reducible) uncertainty.

These issues are collectively addressable with Similarity-Distance-Magnitude (SDM) estimators [3], which transform the signals of epistemic uncertainty previously observed with instance-based, metric-learning approximations of neural networks [2]. We can construct a new activation function by adding Similarity (i.e., correctly predicted depth-matches into training)-awareness and Distance-to-training-distribution-awareness to the existing output Magnitude (i.e., decision-boundary)-awareness of the softmax function. Conceptually this new function is

$$\textsc{sdm}(\boldsymbol{z})_i = \frac{\textsc{Similarity}^{\textsc{Distance} \cdot \textsc{Magnitude}_i}}{\sum_{c=1}^{C} \textsc{Similarity}^{\textsc{Distance} \cdot \textsc{Magnitude}_c}} \tag{1}$$

with a corresponding negative log likelihood loss that takes into account the change of base.

A series of distributional transforms over the output of the SDM activation function then yields an estimator that is remarkably robust to distribution shifts and yields an easy to understand quantity of interest, index-conditional calibration. The SDM estimator demonstrates that for LLMs, there are regions of the output distribution that are low variation and high-probability that can be reliably detected. Existing estimators marginalize over the distinctions in these regions, which can cause unexpected behavior at test-time, rendering moot any statistical assurances and rendering unreliable the empirical patterns observed during model development.

Constrained Inference as Evaluation

Importantly, with a meaningful estimator of the predictive uncertainty, we can construct instance-wise verifiers over the generated output, as detailed in [3]. Evaluation at the model level can then reflect how the models are used in practice: We prefer the model(s) that maximize the cardinality of the admitted (i.e., non-rejected) points for a given probability threshold.

Controllable AGI

The newfound behavior of high-dimensional models with SDM estimators is a fundamental departure from the unpredictable brittleness of unconstrained LLMs. It also lends itself to a deeper understanding of the limits and possibilities of the large parameter neural networks, and it provides a concrete goal for building much larger models that end-users can trust. Importantly, on this path, the intermediate models are nonetheless useful and reliable for targeted use-cases.

Controllable Artificial Intelligence can be defined as the separation of aleatoric (irreducible) uncertainty and epistemic (reducible) uncertainty in regions of high-probability and near-zero dispersion over the output distributions of high-dimensional models with non-identifiable parameters (e.g., large language models, LLMs). This is achievable with SDM estimators at current scales for targeted tasks, with tasks and outputs unlike those seen in the observed data rejected by the verifier. By extension, we can then define controllable Artificial General Intelligence (AGI) as this same behavior, but over arbitrary tasks and input modalities, with the cardinality of the non-admitted points approaching zero.

The Path to Controllable AGI: A Path not Taken, but Re-takeable

Starting around 2019, deep learning can be viewed as forking into two paths, with differing outlooks for constructing general purpose AI systems. The first was predicated on scaling unconstrained LLMs to increase marginal accuracy to eventually obtain a general purpose model, with relatively limited focus on near-term use-cases. This has been the approach of the large foundation model providers. The second took a more critical view of the existing architectures and was an effort to make LLMs useable in the near-term for real-world settings, such as medicine, and to ensure that larger models would also be reliable and controllable. It sought a fundamentally new understanding of the statistical behavior of high-dimensional objects. Whereas the first approach was primarily a systems engineering endeavor, the latter required significant novel research, both from a deep learning and a statistics perspective. The first path has shown that scaling data and compute does tend to increase marginal accuracy across tasks, ceteris paribus. However, at the same time it has become clear that brittleness to distribution shifts and lack of reliable uncertainty quantification preclude the use of such LLMs in most real-world settings. Nonetheless, most AI and AGI policies of major research groups have been predicated on, if not outright overfit to, the 2020-era unconstrained LLMs and scale-only approaches. That is not an efficient allocation of resources, since scale-only approaches are unlikely to fully overcome the observed limitations given the high-dimensions of the input, output, and models. Notably, suggested variations on the unconstrained LLM theme, such as test-time compute, are also not sufficient for reliable deployments (and relatively inefficient) without reliable estimators of predictive uncertainty. Now that the second path has resulted in robust and reliable estimators of the predictive uncertainty, and a concrete path to controllability of much more powerful models than currently available, it makes sense to shift resources to scaling SDM estimators, as well as more complex pipelines and networks, such as SDM networks [3].

For applications, this will result in models that behave as users instruct them to; when the models are uncertain, applicable actions can be taken, based on the statistically-principled, interpretable verification estimators over the output. This removes the mystery and unwanted surprises that accompany the current generation of LLMs. It also provides a concrete basis for regulatory guidance, if applicable for particular settings (e.g., medicine), with well-defined parameters: The acceptable probability, acceptable cardinality of the admitted set, minimal acceptable effective sample sizes, degree of near-zero dispersion across admitted points (over high-probability partitions of data/model iterations), and related.

Fortunately, the arc of AI history is still early, and there is still time to enable reliable AI before such systems are deeply integrated into the decision-making systems of modern society.

References

[1] J. T. G. Hwang and A. A. Ding. Prediction Intervals for Artificial Neural Networks. *Journal of the American Statistical Association*, 92(438):748–757, 1997. ISSN 01621459. URL `http://www.jstor.org/stable/2965723`.

[2] A. Schmaltz. Detecting Local Insights from Global Labels: Supervised and Zero-Shot Sequence Labeling via a Convolutional Decomposition. *Computational Linguistics*, 47(4):729–773, Dec. 2021. doi: 10.1162/coli_a_00416. URL `https://aclanthology.org/2021.cl-4.25`.

[3] A. Schmaltz. Similarity-Distance-Magnitude Universal Verification, 2025. URL `https://arxiv.org/abs/2502.20167`.