A Proposal for Scaling the Scaling Laws

Wout Schellaert[1,2] **Ronan Hamon**[3] **Fernando Martínez-Plumed**[1]

José Hernández-Orallo[1,2]

[1]Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València
[2]Leverhulme Centre for the Future of Intelligence, University of Cambridge
[3]European Commission, Joint Research Centre, Ispra, Italy
wschell@vrain.upv.es

Abstract

Scaling laws are predictable relations between the performance of AI systems and various scalable design choices such as model or dataset size. In order to keep predictions interpretable, scaling analysis has traditionally relied on heavy summarisation of both the system design and its performance. We argue this summarisation and aggregation is a major source of predictive inaccuracy and lack of generalisation. With a synthetic example we show how scaling analysis needs to be *instance-based* to accurately model realistic benchmark behaviour, highlighting the need for richer evaluation datasets and more complex inferential tools, for which we outline an actionable proposal.

1 Introduction

Analysing how AI systems *scale* – how their performance is affected by various design choices such as parameter count or dataset size – has become a fruitful empirical tool: it informs the design of new generations of (scaled-up) systems (Hoffmann et al., 2022), uncovers architectural limitations (McKenzie et al., 2023), and generally helps both industry and policy in planning for what the near future of AI might look like. For example, the concept of *scaling laws* (Hestness et al., 2017; Villalobos, 2023) deals with capturing predictable patterns in the relation between scale and performance into simple mathematical relations, from which data driven extrapolations and predictions about next-generation performance can then be made.

Despite the usefulness of scaling analysis, there are also several issues. A primary concern is generalisation. Scaling laws need to be tailored (i.e. fitted) to different domains, architectures, and often even to each set of model hyperparameters independently. There is no universal 'scaling law' (Abnar et al., 2021; Caballero et al., 2022). Insights that generalise across tasks and metrics are rare. A second notable issue is predictive accuracy. For example, modelling breakpoints – changes in the behavioural trend – has proven difficult, partly because of the limited expressivity of the functional forms (Caballero et al., 2022), but also because new capabilities seemingly emerge out of the blue at certain scales (Wei et al., 2022)[1].

We argue that *oversummarisation* is a significant contributing factor to these issues. Firstly, the dimensions of scale and size capture only a small part of technological innovation, and are a rudimentary summary of the attributes that define and differentiate AI systems overall. Current methods typically consider only one or two scalable design choices. This is the **oversummarisation of systems**.

Secondly, the empirical aggregate performance metrics that act as the unit of analysis are, by construction, summary statistics. By not looking at the actual features of the task instances – like a researcher might – performance is treated as an abstract number, devoid of information that could explain differences. The detection of patterns underlying the relation between task features, system features, and performance is off the table from the start. For example, the aggregate metrics cannot capture any difference in scaling behaviour between subsets of the benchmark. This is the **oversummarisation of task performance**.

While this heavy summarisation is sensible in the light of interpretability or data scarcity, it comes at a cost of generalisation and predictive power. With major NLP evaluation efforts like BIG-Bench (Srivastava et al., 2022) and HELM (Liang et al., 2022) producing huge quantities of instance-level evaluation results across a plethora of different AI systems, it is time to capitalise on the available data, and much like we scale AI itself, to also *scale the inferential tools we use in our analysis of AI.*

[1]Schaeffer et al. (2023) convincingly argue that this is due to the bluntness of the used metrics.

Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models, pages 1–8

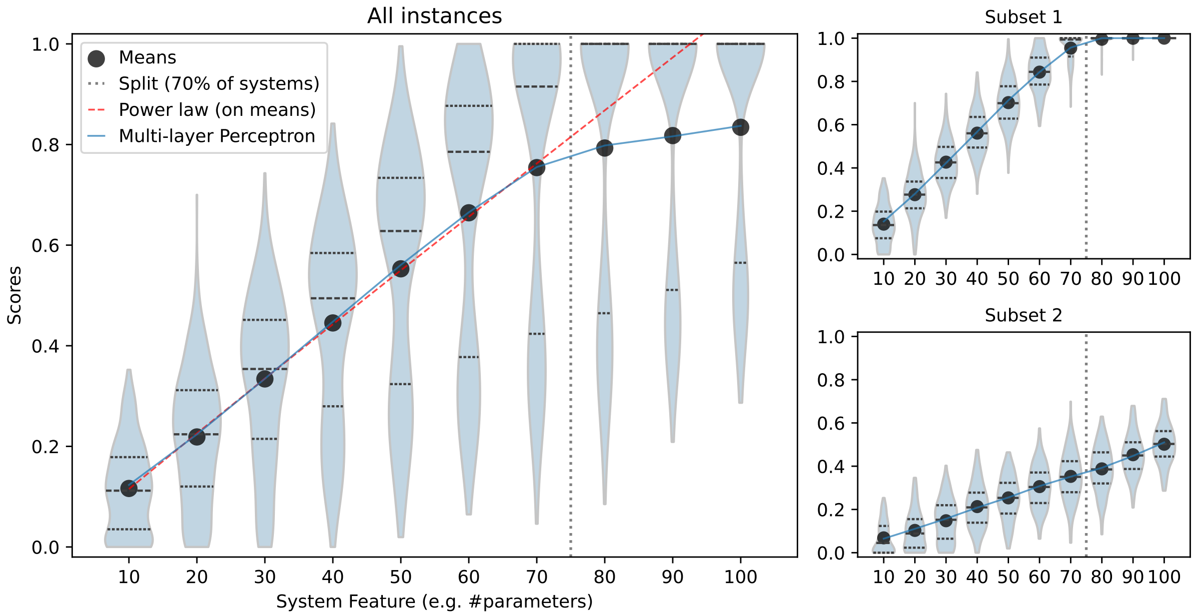

Figure 1: Synthetic example of task performance correlating with system scale which cannot be modelled from aggregate measures, while being completely regular from an instance-level perspective. The plot shows ten synthetic AI systems, whose synthetic evaluation scores are designed to be dependent on an abstract feature of the system. For example, system 2 has feature-value 20 (e.g. number of parameters), and has a mean score of about 0.2. The violin plots, with the quantiles marked, represent the distribution of scores of the respective systems. The red line is a power law fitted to the mean scores, while the blue line represents the aggregated predictions of a simple multi-layer perceptron (MLP) that predicts instance level scores. Both are trained/fitted on the smallest seven systems only. The last three systems then act as a test for the performance predictor.

2 Synthetic Example

To illustrate the challenges outlined earlier and to lay the foundation for our proposed methodology, we present a synthetic scenario where the scaling behaviour cannot be modelled from aggregate measures. The setup is as follows: we hypothesise ten AI systems, each of which scales up over the same (abstract) system feature, e.g. number of model parameters. We also devise a simple synthetic dataset consisting of 1000 instances divided into two subpopulations. The instances of the dataset synthetically have only one feature: a one-hot coded vector indicating which of the two different subgroups of the benchmark the instance belongs to.

To bring this to life, consider the task of sentiment classification of English text, whose domain would naturally contain a blend of English varieties, e.g. ‘standard English’, acting as subpopulation 1, and African-American Vernacular English (AAVE), acting as subpopulation 2. In this scenario, a one-hot vector indicating the subpopulation would not be provided explicitly, but actual features of the English variants would allow identifying the texts as belonging to different populations.

We now generate synthetic evaluation results, where we design the scores to be dependent on the scalable system feature. We simply let the mean score increase as the system feature scales. We also make this relation between scale and score differ between the two subpopulations, e.g. the sentiment of AVEE might be harder to classify than that of standard English, for example due to lower representation in training data. The scores are in the range $[0, 1]$, representing e.g. the probability assigned to the correct class.

Figure 1 illustrates the example. Observing only the mean scores, a conventional scaling analysis could sensibly only make a linear extrapolation (in red). On the other hand, an instance-based approach could discern the distinct subpopulations, noting that performance must saturate in the first group while increasing more gradually in the second. To exemplify this, we train a simple neural network[2] on the set of synthetic evaluation records[3] to predict instance-level scores, that can correctly extrapolate to larger systems (blue curve).

[2]A scikit-learn MLPRegressor with default parameters, with outputs clipped between 0 and 1.

[3]Tuples ⟨system feature, instance feature, score⟩.

Evaluation Records			
System Features (id, #params, #tokens)	**Instance Features** (same as the systems gets)	Model Outputs (optional)	**Score** (number)
GPT4, 1.8T, 2T	What movie do these emojis represent? 🧑👴🚗⌚		1
GPT4, 1.8T, 2T	Translate "Can I have the bill please?" into Italian.		0.7
Bard, 350G, 1.5T	What movie do these emojis represent? 🧑👴🚗⌚		0
...	...		...

Figure 2: Example dataset of evaluation records.

While the example is obviously exaggerated and idealistic, benchmark subgroups are not uncommon (Swayamdipta et al., 2020; Siddiqui et al., 2022). In a high-dimensional problem space like NLP, the subgroups are however not as crisp as in our example, and identifying them is far from straightforward; this complexity is precisely why we need more sophisticated statistical methods beyond simple aggregate measures. In general, it is hard to isolate a single capability in benchmark design (AREA et al., 2014; Hernández-Orallo, 2017), if that even makes sense for novel kinds of intelligence like LLMs. In reality, there will be a mixture of (meta-)features of both system and instance that influence the scores in complicated ways. Example instance features that the literature has shown to be impactful are input length or grammatical complexity (Graesser et al., 2011; Kazemnejad et al., 2023); Clever Hans phenomena and general confounding (Martínez-Plumed et al., 2022); mislabelling (Northcutt et al., 2021; Kreutzer et al., 2022), label disagreement (Aroyo and Welty, 2015; Pavlick and Kwiatkowski, 2019), or task ambiguity (Liu et al., 2023); or general dependence on other skills, e.g. for dealing with numeric values (Amini et al., 2019), negation (She et al., 2023), or social understanding (Sap et al., 2019). While these phenomena are also tested for individually, they are nonetheless confounding factors in most benchmarks. They influence scaling behaviour in currently unknown ways and require us to actually relate scores to instance features, instead of treating performance as an abstract number.

3 Proposal

Our proposed approach emphasises the integration of detailed evaluation data. It involves following three-step process:

1. **Collect a dataset of evaluation records**, where each record corresponds to the score a particular AI system achieved for a particular task instance. The dataset can incorporate multiple tasks and multiple systems, and preferably does so in order to enable cross-system and cross-capability generalisation. While it is unfortunately rare to make fine-grained evaluation data publicly available (Burnell et al., 2023), recent evaluation efforts such as BIG-Bench (Srivastava et al., 2022) and HELM (Liang et al., 2022) have made massive amounts of instance-level scores available that can be adopted directly. At the same time, one should describe the systems under examination with machine-readable features, which can range from straightforward attributes like model size to complex architectural characteristics or whether specific training methods such as RLHF (Ouyang et al., 2022) were used. Any design choice that plausibly has significant impact on performance is useful and needed information. Figure 2 illustrates an example of such a dataset.

2. **Train an instance-level score predictor.** Hernández-Orallo et al. (2022) introduced assessor models as conditional density estimators $\hat{p}(r|\pi, \mu)$ for doing predictive inference regarding score r given system features π and instance μ. Starting from the dataset of evaluation records, the estimator $\hat{p}(r|\pi, \mu)$ can be constructed as a standard machine learning system, with π and μ acting as inputs, and score r acting as the label. For our sentiment classification for example, it could be a regression tree trained from tabular system feature data and embeddings of the textual instances.

3. **Predict scores for hypothetical systems**. Equipped with the predictor $\hat{p}(r|\pi, \mu)$, we can describe a hypothetical system π' –with scaled up features– and collect instance-level score predictions for the instances of existing benchmarks. To make an overall performance estimation for π' on a benchmark dataset D, we simply combine the individual predictions, for example by averaging the predicted score for each instance in D: $1/|D| \sum_{\mu \in D} \arg\max_r \hat{p}(r|\pi', \mu)$ – analogous to how we would compute actual scores.

The design space for assessor models is large and the inferential problem is still a challenging extrapolating one. But the approach we propose should be able to – with the right inductive biases – at least equal the predictive accuracy of current scaling law methods since the same (and more) information is used. It can capture nonlinear behaviour before aggregation, and with appropriate design, generalisation and predictive accuracy should improve over low dimensional methods.

Apart from the pure predictive aspect, this approach can provide other scaling related insights as well. For example, one could use feature attribution methods to decouple the influence of various (scaled-up) design choices, comparing e.g. influence of scaling human feedback versus scaling the causal next-token training. One could reverse engineer the design of GPT-4 (OpenAI, 2023) by searching for the features that most accurately match actual GPT-4 performance. And while we have focused on extrapolation, it is perfectly possible to ask interpolating questions, e.g. investigating the performance trade-offs and identifying "sweet spots" for system design – such as the mix of training data, the type of optimisation algorithm used, or the inclusion of certain features – that stick to more familiar territory.

4 Related Work

Scaling laws in deep learning research focus on empirical relationships between performance metrics and design choices such as architecture, model size, or dataset size. Initially driven by findings that test loss scales with training data size in a power-law fashion (Hestness et al., 2017), research has diversified to analyse a range of tasks and architectures (Rosenfeld et al., 2019; Henighan et al., 2020; Kaplan et al., 2020) and to theorise scaling exponents (Sharma and Kaplan, 2020; Hutter, 2021; Bahri et al., 2021). However, recent work highlights the non-universal applicability of these laws, particularly in predicting downstream task performance (Hoffmann et al., 2022; Sorscher et al., 2023; Caballero et al., 2022), which is further complicated by the nuances of transfer learning (Abnar et al., 2021; Tay et al., 2022). In general, we find a critical gap in current methods: the over-reliance on aggregated data and limited system characteristics.

Approaches that deal with oversummarization of systems are proposed by Srinivasan et al. (2022) and Jain et al. (2023), which learn or meta-learn from multiple system features and therefore generalise better across systems and tasks, but still work at the aggregate performance level.

Instance-level score prediction is closely related to the notion of predictive uncertainty and calibration in probabilistic systems. Including for LLMs, it revolves around the idea that these systems can signal their own confidence by assigning probabilities to potential outcomes, much as we expect from evaluative models. Predictive uncertainty is the focus of intense research (Mielke et al., 2022; Kadavath et al., 2022; Baan et al., 2023; Hu et al., 2023), but conclusions are often contradictory or context dependent. The fields of anomaly detection and confidence estimation (e.g. Corbière et al., 2019 and Qu et al., 2022) are closely related as well. As described by (Hernández-Orallo et al., 2022), these investigations typically assume requirements that make them differ from the pure 'performance prediction' perspective adopted in our approach, e.g. by not being anticipative and requiring access to model outputs or internals, both of which are not available in the context of scaling laws.

The performance prediction idea also extends and is influenced by other research areas, such as Item Response Theory (Martínez-Plumed et al., 2019; Vania et al., 2021), which predicts success based on system ability and task difficulty, and techniques such as surrogate evaluation (Sacks et al., 1989) and Datamodels (Ilyas et al., 2022), which examine model behaviour in relation to training data. In addition, methods for detailed error analysis (Amershi et al., 2015) contribute to the understanding of model performance by identifying incorrect predictions and highlighting strengths and weaknesses.

5 Conclusion

Acknowledging the challenges of scaling analysis, our proposal aims to mitigate them by leveraging a richer dataset and more powerful inferential tools, i.e. "scaling the scaling laws". The approach unlocks various new applications and aspires to enhance predictive accuracy and generalisation, ultimately aiming for a single assessor model doing inference about scaling behaviour for all tasks and systems with sufficient evaluation data available. We invite the research community to contribute to this endeavour by harnessing instance-level evaluations and amplifying the collective progress in understanding AI performance.

Limitations

While our approach aims to help remediate the challenges of scaling analysis, it of course does not wholly fix the problems of generalisation and predictive accuracy in such a complex and multidimensional extrapolation setting. Predicting non-linear performance trends requires careful assumption making, especially when no trend reversal has been observed. Feature engineering is also critical, but is complicated by mixed input types, label imbalance, unknown variables, inconsistencies and noisy data. The large design space requires strategic decisions about model training and data handling, presenting us with a challenging machine learning problem, compounded by the conventional perils of scaling analysis.

Ethics Statement

We acknowledge the ethical responsibilities inherent in predicting AI scalability and are committed to transparency and the cautious application of our models. While we aim to inform resource allocation and research direction, we urge against over-reliance on predictions for critical decisions and emphasise the importance of safety, fairness, and mitigating potential risks as AI systems advance. Any forecast made by our approach should be interpreted as a rough estimation, not as the definite path forward.

Acknowledgements

This work was funded by valgrAI, the Norwegian Research Council grant 329745 Machine Teaching for Explainable AI, the Future of Life Institute, FLI, under grant RFP2-152, CIPROM/2022/6 and IDIFEDER/2021/05 (CLUSTERIA) funded by Generalitat Valenciana, European Union's (EU) Horizon 2020 research and innovation programme under grant agreement No. 952215 (TAILOR), US DARPA HR00112120007 (RECoG-AI) and Spanish grant PID2021-122830OB-C42 (SFERA) funded by MCIN/AEI/10.13039/501100011033 and "ERDF A way of making Europe", as well as the European Union under agreement INNEST/2021/317 (Neurocalçat) and by the Vic. Inv. of the Universitat Politècnica de Valencia under DOCEMPR21, DOCEMPR22, and DOCEMPR23.

References

Samira Abnar, Mostafa Dehghani, Behnam Neyshabur, and Hanie Sedghi. 2021. Exploring the Limits of Large Scale Pre-training. In *International Conference on Learning Representations*.

Saleema Amershi, Max Chickering, Steven M Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. Modeltracker: Redesigning performance analysis tools for machine learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 337–346.

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine*, 36(1):15–24.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. American Educational Research Association, Washington, DC.

Joris Baan, Nico Daheim, Evgenia Ilia, Dennis Ulmer, Haau-Sing Li, Raquel Fernández, Barbara Plank, Rico Sennrich, Chrysoula Zerva, and Wilker Aziz. 2023. Uncertainty in Natural Language Generation: From Theory to Applications.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2021. Explaining Neural Scaling Laws.

Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. 2023. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138.

Ethan Caballero, Kshitij Gupta, Irina Rish, and David Krueger. 2022. Broken Neural Scaling Laws. In *The Eleventh International Conference on Learning Representations*.

Charles Corbière, Nicolas THOME, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. 2019. Addressing Failure Prediction by Learning Model Confidence. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Arthur C Graesser, Danielle S McNamara, and Jonna M Kulikowich. 2011. Coh-metrix: Providing multilevel analyses of text characteristics. *Educational researcher*, 40(5):223–234.

Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B. Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M. Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. 2020. Scaling Laws for Autoregressive Generative Modeling.

José Hernández-Orallo. 2017. *The Measure of All Minds: Evaluating Natural and Artificial Intelligence*, 1 edition edition. Cambridge University Press, Cambridge, United Kingdom ; New York, NY.

José Hernández-Orallo, Wout Schellaert, and Fernando Martínez-Plumed. 2022. Training on the Test Set: Mapping the System-Problem Space in AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12256–12261.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep Learning Scaling is Predictable, Empirically.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack William Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*.

Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. 2023. Uncertainty in Natural Language Processing: Sources, Quantification, and Applications.

Marcus Hutter. 2021. Learning Curve Theory.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. 2022. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

Achin Jain, Gurumurthy Swaminathan, Paolo Favaro, Hao Yang, Avinash Ravichandran, Hrayr Harutyunyan, Alessandro Achille, Onkar Dabeer, Bernt Schiele, Ashwin Swaminathan, and Stefano Soatto. 2023. A Meta-Learning Approach to Predicting Performance and Data Requirements. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3623–3632.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling Laws for Neural Language Models.

Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. 2023. The Impact of Positional Encoding on Length Generalization in Transformers.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're Afraid Language Models Aren't Modeling Ambiguity.

Fernando Martínez-Plumed, David Castellano, Carlos Monserrat-Aranda, and José Hernández-Orallo. 2022. When ai difficulty is easy: The explanatory power of predicting irt difficulty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7719–7727.

Fernando Martínez-Plumed, Ricardo B. C. Prudêncio, Adolfo Martínez-Usó, and José Hernández-Orallo. 2019. Item response theory in AI: Analysing machine learning classifiers at the instance level. *Artificial Intelligence*, 271:18–42.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. 2023. Inverse Scaling: When Bigger Isn't Better.

Sabrina J. Mielke, Arthur Szlam, Emily Dinan, and Y.-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. 2021. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. *arXiv:2103.14749 [cs, stat]*.

OpenAI. 2023. GPT-4 Technical Report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent Disagreements in Human Textual Inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Haoxuan Qu, Yanchao Li, Lin Geng Foo, Jason Kuen, Jiuxiang Gu, and Jun Liu. 2022. Improving the Reliability for Confidence Estimation. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science, pages 391–408, Cham. Springer Nature Switzerland.

Jonathan S. Rosenfeld, Amir Rosenfeld, Yonatan Belinkov, and Nir Shavit. 2019. A Constructive Prediction of the Generalization Error Across Scales. In *International Conference on Learning Representations*.

Jerome Sacks, William J. Welch, Toby J. Mitchell, and Henry P. Wynn. 1989. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense Reasoning about Social Interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are Emergent Abilities of Large Language Models a Mirage? *Deployable Generative AI Workshop at ICML*.

Utkarsh Sharma and Jared Kaplan. 2020. A Neural Scaling Law from the Dimension of the Data Manifold.

Jingyuan S. She, Christopher Potts, Samuel R. Bowman, and Atticus Geiger. 2023. ScoNe: Benchmarking Negation Reasoning in Language Models With Fine-Tuning and In-Context Learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1803–1821, Toronto, Canada. Association for Computational Linguistics.

Shoaib Ahmed Siddiqui, Nitarshan Rajkumar, Tegan Maharaj, David Krueger, and Sara Hooker. 2022. Metadata Archaeology: Unearthing Data Subsets by Leveraging Training Dynamics.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. 2023. Beyond neural scaling laws: Beating power law scaling via data pruning.

Anirudh Srinivasan, Gauri Kholkar, Rahul Kejriwal, Tanuja Ganu, Sandipan Dandapat, Sunayana Sitaram, Balakrishnan Santhanam, Somak Aditya, Kalika Bali, and Monojit Choudhury. 2022. LITMUS Predictor: An AI Assistant for Building Reliable, High-Performing and Fair Multilingual NLP Systems. In *AAAI*.

Aarohi Srivastava et al. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Yi Tay, Mostafa Dehghani, Samira Abnar, Hyung Won Chung, William Fedus, Jinfeng Rao, Sharan Narang, Vinh Q. Tran, Dani Yogatama, and Donald Metzler. 2022. Scaling Laws vs Model Architectures: How does Inductive Bias Influence Scaling?

Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. 2021. Comparing Test Sets with Item Response Theory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1141–1158, Online. Association for Computational Linguistics.

Pablo Villalobos. 2023. Scaling Laws Literature Review. https://epochai.org/blog/scaling-laws-literature-review.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*.