Position: Why Tabular Foundation Models Should Be a Research Priority

Boris van Breugel [1] **Mihaela van der Schaar** [1 2]

Abstract

Recent text and image foundation models are incredibly impressive, and these models are attracting an ever-increasing portion of research resources. In this position piece we aim to shift the ML research community's priorities ever so slightly to a different modality: tabular data. Tabular data is the dominant modality in many fields, yet it is given hardly any research attention and significantly lags behind in terms of scale and power. **We believe the time is now to start developing tabular foundation models**, or what we coin a *Large Tabular Model* (LTM). LTMs could revolutionise the way science and ML use tabular data: not as single datasets that are analyzed in a vacuum, but contextualized with respect to related datasets. The potential impact is far-reaching: from few-shot tabular models to automating data science; from out-of-distribution synthetic data to empowering multidisciplinary scientific discovery. We intend to excite reflections on the modalities we study, and convince some researchers to study large tabular models.

1. Introduction

Let us start with the obvious: the recent progress in modelling text and image is incredibly impressive. It is not just the capabilities of these models that has grabbed the public imagination, it is also their seemingly "creative", human-like output; photorealistic images of horse-riding astronauts and poems in the style of Sylvia Plath. We acknowledge that text and image foundation models have large potentials for real-world good—from low-cost, tailored educational aids, to personalized medicine. Furthermore, these models are already showcasing how the use of ML need not be constrained to the ML community.

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK [2]Alan Turing Institute, London, UK. Correspondence to: Boris van Breugel <bv292@cam.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

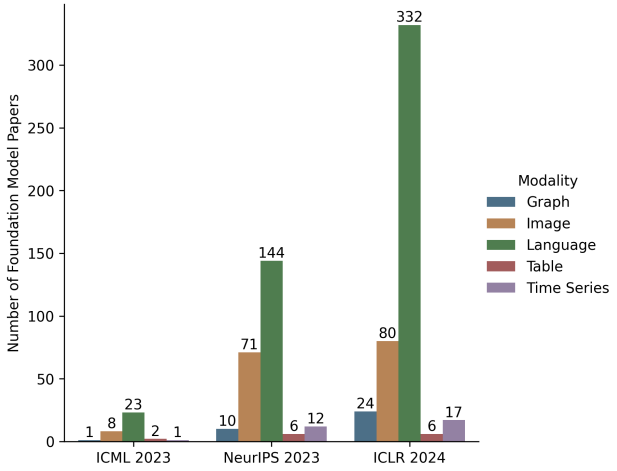

Figure 1. Representation of different modalities in foundation model research across recent ML conferences, roughly estimated as the number of accepted papers with abstracts containing keywords (see Appendix A). LLMs are booming and tabular data is heavily underrepresented.

The immense public interest in text and image modalities, as well as commercial incentives, may partly explain why the majority of foundation model (FM) research is focused on these modalities. The flip side is that other modalities receive hardly any attention from the ML research community (see Figure 1). With this work, we prompt foundation model researchers to reflect on the modality they study, to consider alternative modalities, and to potentially redirect their attention. In particular, in this paper we will argue how *tabular* foundation models have been almost entirely unexplored, yet in some domains present a potential for impact that is just as large, if not larger, than image and text models. We will refer to this class of models as Large Tabular Model (LTM).

How to think about LTMs? LLMs are already widely used as tools, and vision FMs (e.g. (Saharia et al., 2022)) can generate synthetic images or embed real images for downstream tasks. LTMs can play a similar role for (multimodal settings with) tabular data. For data scientists, LTMs could provide an invaluable tool for cleaning and preprocessing datasets; for finding relevant datasets (possibly from different domains or general knowledge-bases, e.g. Wikipedia tables);

for augmenting existing datasets (e.g. few-shot generation of additional columns or performing SQL joins based on related data, e.g. adding a GDP column to a dataset with different countries); and for conducting automated (meta)-analyses. Just like vision FMs, generative LTMs could be used to generate synthetic data, and do so with less data than conventional generative models. Generating full, accurate synthetic datasets that can be used in-place of real data has many applications (van Breugel & van der Schaar, 2023): enable data sharing while maintaining privacy, reduce bias and improve representation of marginalized groups, simulate unseen domains, and augment the size of existing real data. Just like vision and text FMs, using the LTM's embedding of tabular rows (or full datasets) could be used for downstream tasks, including as input for prediction models or other foundation models.

Why tabular FMs have been overlooked. We may wonder why LTMs have been overlooked so far. We hypothesize there are three main reasons: data, tabular ML difficulty, and human perception. First, until recently there has been a lack of large tabular metadatasets. These datasets may also be messy or not fit the typical ML problem setting (e.g. are not labelled for any specific task, or may require domain knowledge). Additionally, due to privacy or ownership concerns, some subfields (e.g. medical data) would not be represented in this data. Second, the difficulty of tabular ML can be discouraging. As discussed, tabular data baselines are often strong, and a new method may not consistently outperform it—possibly hindering publication as a result. Furthermore, there are some unique challenges (see desiderata 2). Third, we believe humans' "more natural" skills of interpreting text and vision may have played a large role. Many vision and text papers have as primary way of evaluation human judgement (e.g. realism of generated images or texts). This is a solution when metrics are unavailable or unreliable—e.g. the primary image generation metrics FID and IS have known problems (Liu et al., 2018; Chong & Forsyth, 2019; Alaa et al., 2022)—yet visual inspection is very hard for tabular data. At last, non-experts understand that StableDiffusion and ChatGPT generate impressively realistic outputs, and hence these models are featured heavily in and outside the ML research community. This in turn may have encouraged more people to study this topic compared to modalities like time-series and tabular.

Why care about the tabular domain. Let us highlight three main reasons why switching to tabular foundation models is worth your consideration. First, **tabular data is ubiquitous** in the real world (Borisov et al., 2022; Shwartz-Ziv & Armon, 2022)—from electronic healthcare records (Fatima & Pasha, 2017) to census data (Office for National Statistics, 2021), from cybersecurity (Buczak & Guven, 2016) to credit scoring (Dastile et al., 2020), and from finance to natural sciences (Shwartz-Ziv & Armon, 2022). Quantitative research relies on these datasets, which in turn progresses scientific knowledge and influences public policy. This practical importance is also reflected in its prominence in online data science competitions, e.g. Kaggle and KDD Cup focus primarily on tables (Kaggle, 2017; Huang et al., 2020).

Second, despite the above **the tabular domain offers uniquely exciting, large, unsolved challenges for researchers**. In the past two decades, machine learning for tabular data has not progressed as resolutely compared to other modalities. Recent benchmarking papers (Gorishniy et al., 2021; Borisov et al., 2022; Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022) all find that XGBoost (Chen & Guestrin, 2016), a tree-based model, is still among the top performers for supervised learning on tabular data. The authors attribute many reasons to this, e.g. tabular data being discontinuous, containing heterogeneous and uninformative features (Grinsztajn et al., 2022), and data containing no spatial invariances that could inform a good prior (cf. convolutional nets for vision) (Borisov et al., 2022). We will go into some of these further in Section 2, but for now we would like to point to another, more obvious reason why tabular data research is lagging behind other modalities: the scale of data, models, and task complexity is vastly different than in the vision and text domain. In (Grinsztajn et al., 2022), the largest datasets are only 50,000 points, and for these datasets (Appendix A2) the difference between data-heavy transformer-based models and tree-based models becomes smaller. Similarly, Borisov et al. (2022) find that for their largest dataset (which counts 11 million samples but just 27 features), the transformer-based model SAINT (Somepalli et al., 2021) outperforms tree-based models. Additionally, benchmark papers (Kadra et al., 2021; Gorishniy et al., 2021; McElfresh et al., 2023) do find settings in which neural nets seem to outperform XGBoost rather consistently, e.g. when the number of samples increases (McElfresh et al., 2023), the number of target classes increases (Gorishniy et al., 2021), and when more modern regularization techniques are used for deep nets (Kadra et al., 2021). Furthermore, none of the discussed tabular benchmarking papers consider settings with multiple or truly large and diverse datasets—even though the power of modern vision and text models may well be attributed to their training on billions of images (Schuhmann et al., 2022) or trillions of tokens (Touvron et al., 2023). These observations indicate that the scaling advantages observed in other modalities (Bommasani et al., 2022), may well hold in tabular data, yet these model sizes and data settings are simply not studied due to benchmarks consisting of primarily smaller datasets and models. This comes with another advantage: **developing SOTA LTMs is still within computational reach of many ML researchers**, cf. the economically-exclusionary cost of training modern LLMs.

Third, the impact of a foundation model for tabular data lies

not just in the ubiquity of the modality, but also in **humans' intrinsic inability to parse tabular data effectively *themselves***. Humans are incredibly good at understanding text and image, and foundation models in the text and vision space have aimed to match this: to encapsulate fundamental understanding of abstract concepts, resembling human-like skills. In the tabular domain the power of FMs may be just as large—foundation models being able to reason about real-world distributions over different variables, and generalize to new relationships. For example, a foundation model trained on tables with features A and B, and B and C, might well be able to reason about the relationship between A and C. Relatively speaking, however, this power would be much higher compared to humans, as table parsing, data analyses, and computations are not an intrinsic skill of ours.[1] We do not know what knowledge can be derived from combining wide, diverse tabular datasets, but the potential alone is exciting and worth exploring.

In a nutshell, designing the first generation of truly large, foundational tabular models could be *an immensely rewarding and exciting opportunity* for many researchers.

Overview. In this position piece, we start by contextualizing the term LTM with respect to other foundation models (Bommasani et al., 2022), and discussing model and data requirements (Section 2). In Section 3, we will discuss the current research related to LTMs, and how it is still rather limiting. We continue exploring the applications of LTMs (Section 4), as well as adaptation challenges (Section 5). At last, in Section 6 we return to our thesis and provide a head-to-head comparison of the impact of LTMs and LLMs along different dimensions, attempting to convince the ML research community to shift more attention to tabular.

2. Large Tabular Model

The term Foundation Model (FM) was first coined by (Bommasani et al., 2022), denoting "any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks; current examples include BERT (Devlin et al., 2019), GPT-3 (Brown et al., 2020), and CLIP (Radford et al., 2021)." The three mentioned examples being LLMs reflect the initial dominance of LLMs in FM research and discourses. Though (Bommasani et al., 2022) write about FMs for other modalities (including image and multimodal), tabular FMs are missing from the 200+ page report.

To draw a direct comparison with LLMs, we coin the term Large Tabular Model (LTM) for a tabular foundation model. In this paper, we focus on the simplest form of tabular data: single independent tables (cf. relational databases). Let us discuss how the foundation model definition moulds to tabular data, and what kind of model and data it necessitates.

[1] In fact, we expect many readers consider tables as abstract and "dry", and may well be the reason why it is so understudied.

2.1. Model Requirements

The two foundation model requirements, large-scale (self-supervised) training and adaptability, extent to LTMs by definition. However, let us propose four desiderata specific to LTM design, that are implied for "large-scale" and "adaptability" to be possible in tabular data. These are not strictly necessary, but desirable for adapting the model "to a wide range of downstream tasks" (see Section 4).

- **D1 Mixed-type Columns.** LTMs should be able to handle the different types of data that are common in tables, e.g. numerical, categorical, datetime, and missing data.
- **D2 Cross-dataset Modelling.** Tabular datasets only cover a specific topic and are often moderately sized. Consequently, to enable large-scale training and broad applicability we need one LTM to be trained on different datasets. This requires a capability to model heterogeneous feature spaces—i.e. feature spaces with different (numbers of) features that could carry very different meanings.
- **D3 Textual Context.** The meaning of tabular data is often dependent on contextual metadata, e.g. the dataset description, column names, and category names. An LTM should leverage this information.
- **D4 Invariance/Equivariance w.r.t. column order.** Column order is usually arbitrary for tables. We want an LTM to reflect this, and have output that is invariant or equivariant to input permutations. In other words, letting f denote the LTM, x a row, and T some permutation of the columns, we desire (invariance) $f(T(x)) = f(x)$ or (equivariance) $f(T(x)) = T(f(x))$.[2]

Model type and architecture. Similar to *FM* and *LLM*, we envision the term *LTM* to be very much open to interpretation. In particular, we set no restrictions on the type of models used for LTMs. In LLM literature, conditional generative models with a transformer backbone form a cornerstone for many applications. Yet, it would be false to state encoder-only LLMs that create embeddings are not foundation models. Similarly, we believe some type of conditional generative model (i.e. that allows generation based on a subset of the variables and metadata) could be an important model class for future LTMs: this would for example allow few-shot (probabilistic) prediction, imputation, and conditional generation. For representation learning,

[2] Note, equivariance only makes sense if the output is a sequence of the same length as the input. This may not be the case for some LTMs, e.g. LTMs that aim to embed rows as static embeddings.

encoder-only LTMs may be more suitable. Architecture-wise, attention-based architectures are dominant in related work (Section 3), for good reason: they can process a varying number of features and naturally satisfy **D4** (i.e. output is equivariant w.r.t. feature order when no positional encoding is used). Nonetheless, benchmarking papers (Gorishniy et al., 2021; Borisov et al., 2022; Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022) could motivate researchers to look beyond transformer architectures.

2.2. Data

Key to FMs, is that they are "trained on broad data" (Bommasani et al., 2022). In particular, many of the performance gains of recent LLMs and vision models are due to scaling (Brown et al., 2020; Kaplan et al., 2020; Rombach et al., 2022; Hoffmann et al., 2022). "Broad data" for LTMs implies having a large corpus of tables with context (e.g. descriptions). Fortunately, there has been a growing body of work composing these large datasets. WebTables (Lehmberg et al., 2016) consists of 10M HTML datasets from the web. Bhagavatula et al. (2015) proposed a dataset based on Wikipedia tables, which has the advantage of being curated and covering a wide range of topics. More recently, the size (and hence coverage) of these datasets has increased tremendously. Most recent is Eggert et al. (2023), who publish TabLib—a metadataset counting 627 million tables and 827 billion context tokens, covering a diverse range of topics. This is comparable to the size of large-scale image and text datasets, and hence should not prohibit research into LTMs. Nonetheless, there are some possible data challenges, which we will go into in Section 5.

2.3. Benchmarking tasks for LTMs

The broad definition of foundation model poses the question: even if we have build a supposed LTM, how do we benchmark it and deem it successful? Multiple tasks, settings, and datasets will likely be required, similar to LLM benchmarks—e.g. the massive text embedding benchmark (Muennighoff et al., 2023) consists of 8 tasks, 56 datasets, and 112 languages. In Appendix B we outline some tasks and settings that could be used for benchmarking LTMs. We hope to encourage others to focus on developing metrics and benchmarks further.

Takeaway: We use Large Tabular Model (LTM) to denote a foundation model for tabular data. This term is only loosely defined, but implies a large-scale tabular model that is adaptable to many downstream tasks.

3. Current State of LTMs

In the previous section we have discussed our position on what LTMs should be able to do, which included handling some of the inherent difficulties of tabular data (e.g. its lack of structural meaning, mixed-type variables, limited size of datasets). Let us discuss some of the works related to LTMs, how they already fit some of the LTM criteria, and which criteria they not yet fulfill.

Representation learning. Both Yin et al. (2020) and Deng et al. (2020) adapt BERT (Devlin et al., 2019) for table understanding and parsing (e.g. entity linking, column type annotation). Both set-ups are similar, to linearize tables into "sentences", add textual metadata, use masked reconstruction loss as training objective, and train on large tabular metadatasets. Their methods have numerous technical limitations that prohibit wide adaptation, e.g. a 128 token context window. (Zhu et al., 2023b) focuses on cross-table pretraining of larger transformer models. They tokenize numerical values more efficiently, as single tokens (Gorishniy et al., 2021), and for categoricals train an embedding for each unique category in all datasets. They pretrain on just 100 datasets, and their category look-up prevents large-scale cross-table training (**D2**). (Ye et al., 2023)'s CT-BERT present a similar approach, but instead of linearizing each row as tokens, they embed the tokens of each categorical feature using a pretrained LLM, pool these embeddings, and pass the feature embeddings to their model. This results in shorter sequences, and the pretrained LLM adds context efficiently (**D3**). They only consider prediction as downstream task, but do show good few-shot performance. Training is limited to 17k datasets and a maximum of 100 features per dataset, but we expect this could scale to larger models and be adapted beyond prediction.

Supervised learning Though supervised learning methods are usually disqualified from being foundation models, we find three works noteworthy. LIFT (Dinh et al., 2022) finetunes an LLM on linearized tabular data, and extensive experiments indicate similar performance to traditional baselines. TabFM (Zhang et al., 2023) takes a more general approach, representing rows differently, adding task-specific information, and showing good few-shot prediction using in-context examples. TabPFN (Hollmann et al., 2023a) takes a vastly different angle than the previous works: it is a transformer model that is trained on large quantities of tabular *synthetic* data, to provide few-shot priors for supervised learning. Despite it being trained on toy data, the authors claim that models trained with TabPFN priors outperform XGBoost (Chen & Guestrin, 2016) on real-world datasets. This generalization capability, from toy to real datasets, hints at there being more structure in the tabular domain than some previous works suggest (Grinsztajn et al., 2022).

Generative learning. Generative learning approaches are lacking behind significantly. Yoon et al. (2018) present a GAN approach that learns to generate samples across domains using a shared latent space, similar to Cycle-GAN

(Zhu et al., 2017) in the image domain. Their intention is to augment smaller domains with synthetic samples that have been "translated" from larger domains, so that this improves supervised learning on these small domains. The problem of this approach is that their method uses a separate encoder-decoder pair for each domain to translate to and from the shared latent space. This will not work well for very small domains, does not scale well to many domains, and prohibits few-shot translation. Additionally, their method does not satisfy **D1**, **D3** and **D4**.

Pretrained LLMs may provide a solution to generation, as they are generative, contain general knowledge, and can tokenize and embed rows into sequences of meaningful numerical vectors (**D3**). However, we believe there are significant disadvantages to adapting LLMs for tabular generation. Let us elaborate.

3.1. Tabular Generation with LLMs

Three recent works adapt LLMs for tabular generation. Borisov et al. (2023) use a finetuned LLM out-of-the-box for generating tabular data. (Solatorio & Dupriez, 2023) propose a similar model based on GPT-2, but which can also generate relational datasets and improve efficiency through modelling a reduced fixed-set vocabulary for each column (following (Padhi et al., 2021)). Last, (Zhao et al., 2023) show they achieve better performance by retraining a smaller LLM. Though these works finetune/train the LLM on just a handful of datasets—probably because their LLM backbone is expensive—they could relatively easily be applied to larger corpi of datasets. The advantage of LLM-based generation is its simplicity, it not requiring any manual preprocessing of data, and it allowing typical LLM techniques (e.g. prompting and in-context examples). However, LLMs have three major shortcomings for generating tabular data.

LLMs are not good at modelling continuous variables. Standard LLMs are inefficient for tabular data, due to tokenization of numerical values (Thawani et al., 2021). A single numerical variable is implicitly modelled as an autoregressive series of binned, categorical variables (e.g. 1.89 is modelled as $1 \rightarrow . \rightarrow 89$). This lacks any prior on the continuity of common data distributions, e.g. predicting 1.89 or 1.90 should have about the same probability, in contrast to 1.89 and 90.89, yet both differ in just one token. This has been shown to lead to unrealistic outputs in LLMs (Spithourakis & Riedel, 2018).

Some have aimed to improve numeracy through better tokenization of numerical values (Spithourakis & Riedel, 2018; Jiang et al., 2020; Golkar et al., 2023). Though we consider this promising, current solutions come with side effects. For example, some encode number r into fixed vector $r\mathbf{e}$ with $\mathbf{e}$ some constant vector, but this results in numerical variables being modelled as point estimates—removing the ability to generate using the LLM.

This is even more problematic for generative tasks in which we aim to sample from the LLM to mimic a continuous distribution. In Figure 2 we visualize the relatively complex process for an LLM to output just a single independent Gaussian variable. Evidently, modelling just a simple continuous variable requires a seemingly unnecessary amount of capacity, without even considering conditional dependencies on other variables. Other model classes (e.g. diffusion models (Song & Ermon, 2019; Ho et al., 2020)) inherently support continuous variables, which we believe makes them a better option.

LLMs are poorly calibrated. At the same time, Renda et al. (2023) found that LLMs are poor at sampling from theoretically trivial distributions, e.g. a uniform distribution. This is in line with the poor calibration of LLMs (Desai & Durrett, 2020; Jiang et al., 2021; 2023). This does not bode well for modelling more complex, continuous distributions autoregressively.

Standard LLMs are expensive. LLMs are designed to be highly flexible, and typically function autoregressively. This flexibility means we may learn difficult relationships (e.g. model continuous distributions, Figure 2). For tables, where a single row may contain hundreds of features that each contain several tokens (especially if numbers are tokenized naively), this would mean inefficient, slow, and expensive training and inference—lessening research and adaptation. For large datasets or inference with in-context examples, the context-window size could also be a hard limitation, e.g. GReaT (Borisov et al., 2023) cannot handle more than 100, low-dimensional in-context samples. But is this expense necessary? Ye et al. (2023) use a pretrained LLM for encoding fields, but use a smaller transformer on top—this approach reduces the sequence length to the number of features (cf. tokenized length of row and context) and speeds up training significantly through caching. Furthermore, smaller transformers may suffice: tabular data is highly structured and current state-of-the-art methods, e.g. XGBoost for prediction, perform well despite their small size. Consequently, we believe it would be fruitful to explore other architectures (including transformer-based) beyond the standard LLM.

Takeaway: A recent surge of papers use cross-table training for unsupervised or supervised learning. Many adapt LLMs, but we argue why LLMs are not efficient and performant in the presence of numerical columns. Building a generative LTM is the largest challenge.

4. Real-World Impact

4.1. Categorizing Adaptation

The foundation of machine learning (ML) is data. Unfortunately, data can be limiting for some domains (e.g. healthcare, where data is costly and privacy-sensitive), subpopulations (e.g. historically underrepresented minorities), and regions (e.g. developing countries). A lack of good data hinders the development, training, and testing of ML models, and science in general. In particular, we have discussed how ML for tabular data is lagging behind, which may be explained by modern ML methods requiring large training sets that the tabular domain does not always offer (Section 1). Additionally, even if data is available, it may not always be easy to derive meaning or knowledge from it. LTMs can help: LTMs may exhibit the same few-shot abilities as LLMs (Brown et al., 2020), allowing us to adapt them to a wide variety of downstream tasks with fewer data. We highlight some applications for this adaptation.

Direct vs indirect adaptation. We consider two types of adaptation: direct adaptation to the target task (e.g. fine-tuning the LTM for some prediction task) or indirect adaptation through synthetic data (e.g. create a "fake" dataset that augments the real data). To illustrate the advantages of each, let us consider classification on a small imbalanced dataset as downstream task.[3] For direct adaptation, one can fine-tune on this small dataset or supply the data as in-context examples, such that the LTM itself provides predictions. This directness is convenient, could be cheap, and allows us to measure the LTM's success directly through its downstream test performance. The indirect approach is different: it uses the LTM to generate a synthetic dataset that augments the real data, which is then used by a downstream model for training. This might be more complex, partly because it is hard to determine how well the LTM did at generating the data until we have run the downstream model (see Section 5). The advantages, however, are that we can (i) explicitly *improve* the real data, e.g. by sampling additional points for the underrepresented group, (ii) can share the synthetic data with others, and (iii) downstream researchers can perform their usual task (e.g. train a model, conduct data analysis), without any need of the LTM. To ensure downstream results are trustworthy, evaluation of the LTM and synthetic data are essential—we discuss this further in Section 5. The choice for direct or indirect adaptation will likely be application-dependent, so let us look into some of these next.[4]

[3]There are strong connections to the debate around discriminative versus generative machine learning models. However, FMs are usually self-supervised (and many FMs like LLMs are generative by nature), hence adapting an LTM to be generative may not necessarily be harder than making it discriminative.

[4]The "indirect" approach is rather unique to tabular data. In the text domain, creating synthetic data*sets* for downstream models often makes little sense—we would still need an LLM to use this data. In the vision generative model literature this is slightly more explored, e.g. (Wang et al., 2019) use generative models to turn CGI-generated images into photorealistic training images for better downstream prediction.

4.2. LTMs for Responsible AI

Inclusiveness and representation. Historically, it has been hard to model underrepresented groups well due to the inherent data scarcity of these groups. Studying *subgroup robustness* is thus an active area of research, e.g. see (Gardner et al., 2022). Large language models have few-shot reasoning abilities (Brown et al., 2020), hence we can expect similar behaviour for LTMs. This would allow adapting these models better to small datasets or underrepresented subgroups within large datasets, hence improving performance on these subgroups. Additionally, an LTM could be used to generate synthetic data for these groups—augmenting the real data to improve representation of marginalized groups—in turn enabling better downstream science and ML development for these groups (van Breugel & van der Schaar, 2023). This is similar to other tabular augmentation approaches, e.g. SMOTE (Chawla et al., 2002) or the deep generative (van Breugel et al., 2023a), but may provide more realistic examples due to the LTM's prior knowledge.

Robustness through out-of-domain simulation. ML models can perform unpredictably bad on out-of-distribution and distributionally shifted test data (Kolesnikov, 2023; Liu et al., 2023; Gardner et al., 2024). Synthetic data generated by generative models has been used in the past to simulate unseen or scarcely seen scenarios for testing. For example, van Breugel et al. (2023a) show how simulated data can be used for estimating the change in performance of trained ML predictive models (due to an environment change), and Tucker et al. (2020) generate synthetic data to assess healthcare software. These methods are heavily limited by their inability to generate out-of-distribution data, and their generative models' data-intensive requirement for training. The generalization and few-shot potential of LTMs could resolve both problems, thus LTMs could play a central role in simulating data for ML model development, probing, and post-deployment monitoring.

Privacy, data democratization, and reproducibility. Training ML models on private data requires a trade-off between privacy leakage and performance (Dwork & Roth, 2014). LTMs that have been pretrained on open, non-private data can be adapted to private target data, thereby reducing the privacy budget compared to standard private ML models—we need less information from the target dataset to model it well. Similarly, an LTM can generate synthetic data with a better privacy-utility trade-off compared to normal generative models. Both direct and indirect adaptation may prove especially beneficial for underrepresented groups, as

these groups have inherently a worse privacy-performance trade-off: there are less points to learn from, hence more information per sample is needed for accurate modelling (van Breugel et al., 2023b).

Generating synthetic data that accurately mimics sensitive data without revealing it, could also allow the publication of more datasets (e.g. data from healthcare providers). This could promote better reproducibility, democratise data access, and enable more powerful meta-analyses (see below).

4.3. LTMs for Science

Meta-analyses. Meta-analyses are essential for consolidating and analyzing research results across studies. Individual studies may only focus on small subpopulations (e.g. one country), resulting in sometimes conflicting, biased, or insignificant results. At the same time, obtaining more data is often difficult and expensive. Meta-analyses re-use data from multiple, existing studies to create more powerful and cost-efficient analyses, removing local bias and noise. They are influential in informing policy—e.g. meta-analyses are at the top of evidence-based practice (Haidich, 2010).

In meta-analyses, different datasets need to be consolidated into one, which is complicated by possible heterogeneity (e.g. different features and contexts). LTMs could reduce manual effort and cost by automating the integration of datasets from different studies by harmonizing formats, inconsistent column names and categories, and data types.

Bridging datasets. Going even further, LTMs could help find and combine relevant tables that may not match directly, and may come from very different scientific fields. This can be compared to performing seamless, automated SQL joins using the LTM's full training database.This could enable multidisciplinary science that is currently intractable.

Data scientist's assistant. LTMs' native understanding of data compared to LLMs, could make them a suitable assistant tool for data scientists. Abilities could include automatic cleaning, exploratory data analyses, finding related datasets, applying SQL queries on complex databases, running automated statistics, and helping visualize and interpret results. This could improve data scientist' productivity, enjoyment of their work (by removing repetitive and often frustrating cleaning), and aid in outwards communication—-e.g. in improving public understanding of data meaning and reliability.

Knowledge base. By learning from a wide range of datasets, the LTM itself becomes a repository of knowledge. LTMs trained on large, high-quality data (e.g. WikiTables) could facilitate data-driven question-answering systems, and statisticians could distill Bayesian priors from LTMs.

4.4. LTMs and Non-Tabular Data

Time-series and relational data. We have focused on the simplest form of tabular data: single tables with independent entries. This could lay the foundation for time-series and relational databases, which are important data types in practice (e.g. electronic healthcare records), yet contain dependent samples. In time-series, entries are usually timestamped and multiple samples are linked by an ID (e.g. a patient, where entries correspond to measurements), and in relational databases there can be multiple tables where one ID is linked to (possibly many) entries in different tables. Some work has already aimed

Multimodality. This paper's supposed dichotomy between tabular and image/text is false of course. Many domains are multimodal. LTMs can be aligned or used to inform models in other modalities, and vice versa. For example, desideratum **D3** states LTMs should use textual context—it would be sensible to acquire this context using a pretrained LLM encoder. Similarly, some of the applications above could still benefit from an LLM intermediary to communicate between the LTM and human. Research into LTMs can also be encapsulated into multimodal FMs, which so far largely ignore tabular data (Li et al., 2023a).

Takeaway: We envision an important role of LTM in promoting responsible AI (including privacy, representation, reproducibility, data democratization, robustness testing and as a powerful tool for scientists (including data preprocessing and analysis, for bridging datasets, and as a source of knowledge itself). LTMs will also play a role for ML research into other modalities, or multimodal settings.

5. Challenges

Building generative LTMs. We have seen how generative LTMs are lagging behind. Modelling *distributions* across datasets (**D2**), types (**D1**) with context (**D3**) is conceptually and architecturally complicated, and poses unique and exciting challenges for researchers.

Scale. As discussed in Section 3, current models are still limited in scale and applicability. Training is usually restricted to relatively small number of datasets (e.g. 100), and often evaluation is restricted to one (supervised) task. As a result, it is very uncertain whether these models generalize well to new datasets and tasks, rendering them unsatisfactory foundation models. Upscaling to larger datasets may be harder than in other modalities, however, because data is "dirtier" (e.g. consist of different file formats, contain missing and noisy fields, or be preprocessed).

Data diversity and quality. The size of datasets like TabLib (Eggert et al., 2023) is impressive, but it is yet to be determined whether their diversity is sufficient for wide appli-

cability of future LTMs trained on this data. In the image and text domain, large-scale datasets composed of web data are likely to cover a large part of the "world-distribution", e.g. include photographs from places all over Earth, and text from different languages. Even then, medical images or small languages may be underrepresented. In the tabular domain, underrepresentation is more likely to be problematic for two reasons. First, because tables have heterogeneous feature spaces and context, most tables are irrelevant to most other tables (cf. images, where each image can be scaled to the same space and where common shapes compose objects, which compose scenes, which compose meaning). Secondly, many forms of tabular data are proprietary, unpublished, or private, and will thus not appear in datasets like TabLib. As a result, the extent to which LTMs will generalize to the medical domain is yet to be seen. On the other hand, recent work (Gunasekar et al., 2023) have shown that small LLMs trained on small, but highly curated data (e.g. textbooks), can perform remarkably well. This is promising for LTMs, especially LTMs specialized to smaller domains for which data banks exist (e.g. Wikipedia tables (Bhagavatula et al., 2015) or gene expressions (Clough & Barrett, 2016)).

Evaluation. Foundation models make mistakes, e.g. hallucinate false information (Ji et al., 2023), hence careful evaluation is essential. Bommasani et al. (2022) divide evaluation into intrinsic and extrinsic. Extrinsic evaluation refers to measuring the performance of the adapted FM on a downstream task, whereas intrinsic evaluation refers to directly measuring the FM's quality. Extrinsic evaluation may not be representative of how "good" an FM is, as it may not capture the performance on other possible downstream tasks. On the other hand, intrinsic evaluation is difficult as FMs are defined in terms of their adaptability to downstream tasks, which is hard to measure without performing (or even knowing what are) those tasks. Intrinsic metrics can also be biased towards one FM—e.g. using test loss of FMs with different loss functions is not possible. LTM evaluation encounters the same difficulties, but is made even more difficult by *visual inspection being unhelpful*: whereas the realism of image and text output can be relatively easily evaluated by humans (and forms a central part in most text and image papers), this is hard for tables. Inspiration may be drawn from the synthetic data literature, where similar evaluation challenges make metrics an active area of research (van Breugel & van der Schaar, 2023). Until intrinsic metrics are improved, we expect most LTM performance evaluation to be mostly extrinsic, using multiple downstream tasks. In Appendix B we include different tasks and settings that could be considered for extrinsic evaluation.

Evaluation goes further than measuring performance. LTMs need to avoid leaking private or copyrighted material, which requires measuring a model's memorization. Ideas from generative model generalization metrics (Theis et al., 2016; Arora et al., 2017; Alaa et al., 2022) and privacy-focused ML (Dwork & Roth, 2014) could inspire LTM metrics. Evaluating bias is another challenge, which we discuss separately.

Bias. Similar to LLMs, LTMs may copy or even exacerbate bias in their training data. Not much is known about the bias in large tabular datasets, e.g. (Eggert et al., 2023). One may hope that tables are usually created to represent information and facts, and are less opinionated than text. This would be naive: datasets can contain discriminative features (e.g. the much-used Boston Housing dataset (Harrison & Rubinfeld, 1978)), under- or misrepresent some groups, and display other forms of bias, e.g. publication bias (Thornton & Lee, 2000). Research into LTMs should go hand-in-hand with research into this bias, and put safeguards in place when publishing LTMs.

Takeaway: The most important challenges for LTMs relate to (i) the challenges of building a model that can handle the unique challenges of tabular data (**D1**-**D4**); (ii) uncertainty around current datasets' quality, bias and diversity; (iii) reliable evaluation of LTMs; and (iv) how bias can be tested and safeguarded against in trained LTMs.

6. Comparing LTM and LLM Impact

Let us end this position piece with a comparison of LTMs and LLMs along five key dimensions of impact.

1. Public and commercial use. The surge in LLM use is a testament to their adaptability, and the fundamental role they may play in productivity and facilitating human-ML interactions (e.g. through tools called through LLMs). LTMs may play an important role in data-driven industries, e.g. finance, education, government, but this will likely be smaller than LLMs.

2. Scientific use. Tabular data is likely the largest data modality in science (Borisov et al., 2022). We have discussed how LTMs may revolutionize how tabular data is processed, analyzed, and re-used across domains. Though LLM applications in the scientific domain are plentiful, current LLMs encounter inefficiency and numerical issues with tabular data (Section 3.1).

3. Potential for public good. LLMs may play a central role in education, personalized healthcare (Gates, 2023), and law (Bommasani et al., 2022). LTMs impact is much more specific to ML and data science, e.g. more inclusive data, the development of more robust models, improving privacy and reproducibility, and scientific knowledge derived from using LTMs (e.g. that use multidisciplinary data). Consequently, both exhibit great potential for public good, yet cover vastly different areas.

4. Risk of misuse. Generative FMs come with a risk of misuse. Highly-realistic content generated by LLMs or vi-

sion models can be used to impersonate individuals (i.e. deepfakes) (Rana et al., 2022), spread misinformation at an unprecedented scale (Marcus, 2023), and raise plagiarism concerns for creators and academia (Kasneci et al., 2023). This risk is smaller for LTMs compared to text and vision, because contributing tabular data to the community often requires identity verification (i.e. data can be traced back to an individual and institution) and is prone to scrutiny (i.e. data scientists are strained to verify the data source). Additionally, malicious data fabrication to fit a false narrative can already be done easily manually without the need of an LTM (cf. image generators, where the alternative is arduous and skillful photoediting). FM risks can also be unintentional however, e.g. it is difficult to guarantee models are unbiased, reliable, and do not reproduce copyrighted or private information. For evaluation of these unintentional consequences, LTMs and LLMs face similar challenges.

5. Impact per researcher. The amount of research in the LLM space vastly outnumbers work related to LTMs, yet there are many interesting, high-potential open questions and applications for LTMs. Furthermore, the scale of current LLM research requires vast computational resources, whereas LTMs have been underexplored at even moderate scale. Consequently, the potential for impact is arguably more plausible and more widely attainable in LTM research.

Takeaway: The above is *not* to give a verdict on which modality is best—inherently we are comparing apples to pears. However, we do hope to make the point that each model class is very much deserving of researchers' attention—attention that LTMs are currently not getting.

7. Conclusion

We believe large tabular models are significantly understudied and within reach. LTMs provide exciting and unique research challenges, impactful applications, and promises a world where data usability extends across single datasets. We hope some readers will reconsider the modality they prioritize in their research, and help build, evaluate, and responsibly apply the first generation of true LTMs.

Impact Statement

In this work we have argued for the importance of more research into tabular foundation models. In Section 4 we have elaborated on how these methods can aid ML inclusiveness, representation, privacy, data democratization, and robustness. On the other hand, we have also discussed how LTMs, like LLMs, can be biased, misused, and require proper evaluation (Section 5). We acknowledge that research on and adaptation of LTMs need to go hand-in-hand with bias and misuse prevention.

Acknowledgements

This research has been supported by the Office of Naval Research UK. Additionally, we would like to thank Julianna Piskorz for some very useful feedback on the initial draft of this paper, and we thank the four ICML reviewers for their constructive comments—all of this has greatly improved the paper. Lastly, BvB is grateful for the brainstorming input and continuing support of Sophie Mary.

References

Ajay, A., Han, S., Du, Y., Li, S., Gupta, A., Jaakkola, T. S., Tenenbaum, J. B., Kaelbling, L. P., Srivastava, A., and Agrawal, P. Compositional Foundation Models for Hierarchical Planning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=dyXNh5HLq3.

Alaa, A. M., van Breugel, B., Saveliev, E., and van der Schaar, M. How Faithful is your Synthetic Data? Sample-level Metrics for Evaluating and Auditing Generative Models. In *International Conference on Machine Learning*, pp. 290–306, February 2022. URL https://arxiv.org/abs/2102.08921. arXiv: 2102.08921.

Arora, S., Ge, R., Liang, Y., Ma, T., and Zhang, Y. Generalization and Equilibrium in Generative Adversarial Nets (GANs). *34th International Conference on Machine Learning, ICML 2017*, 1:322–349, March 2017. doi: 10.48550/arxiv.1703.00573. arXiv: 1703.00573 Publisher: International Machine Learning Society (IMLS) ISBN: 9781510855144.

Bao, G., Zhao, Y., Teng, Z., Yang, L., and Zhang, Y. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Bpcgcr8E8Z.

Bhagavatula, C. S., Noraset, T., and Downey, D. TabEL: Entity Linking in Web Tables. In Arenas, M., Corcho, O., Simperl, E., Strohmaier, M., d'Aquin, M., Srinivas, K., Groth, P., Dumontier, M., Heflin, J., Thirunarayan, K., and Staab, S. (eds.), *International Semantic Web Conference*, volume 9366, pp. 425–441, Cham, 2015. Springer International Publishing. doi: 10.1007/978-3-319-25007-6_25. URL http://link.springer.com/10.1007/978-3-319-25007-6_25. Book Title: The Semantic Web - ISWC 2015 Series Title: Lecture Notes in Computer Science.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D.,

Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. On the Opportunities and Risks of Foundation Models, July 2022. URL http://arxiv.org/abs/2108.07258. arXiv:2108.07258 [cs].

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. Publisher: IEEE.

Borisov, V., Seßler, K., Leemann, T., Pawelczyk, M., and Kasneci, G. Language Models are Realistic Tabular Data Generators. In *The Eleventh International Conference on Learning Representations*, October 2023. ISBN 2210.06280v2. URL https://arxiv.org/abs/2210.06280v2. arXiv: 2210.06280.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, May 2020. doi: 10.48550/arxiv.2005.14165. arXiv: 2005.14165 Publisher: Neural information processing systems foundation.

Buczak, A. L. and Guven, E. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176, 2016. ISSN 1553-877X. doi: 10.1109/COMST.2015.2494502. URL https://ieeexplore.ieee.org/document/7307098. Conference Name: IEEE Communications Surveys & Tutorials.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357, 2002.

Chen, T. and Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. doi: 10.1145/2939672. URL http://dx.doi.org/10.1145/2939672.2939785. Publisher: ACM Place: New York, NY, USA ISBN: 9781450342322.

Chong, M. J. and Forsyth, D. Effectively Unbiased FID and Inception Score and where to find them. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6069–6078, November 2019. ISSN 10636919. doi: 10.1109/CVPR42600.2020.00611. URL https://arxiv.org/abs/1911.07023v3. arXiv: 1911.07023 Publisher: IEEE Computer Society.

Clough, E. and Barrett, T. The Gene Expression Omnibus Database. In Mathé, E. and Davis, S. (eds.), *Statistical Genomics: Methods and Protocols*, Methods in Molecular Biology, pp. 93–110. Springer, New York, NY, 2016. ISBN 978-1-4939-3578-9. doi: 10.1007/978-1-4939-3578-9_5. URL https://doi.org/10.1007/978-1-4939-3578-9_5.

Dastile, X., Celik, T., and Potsane, M. Statistical and machine learning models in credit scoring: A systematic literature survey. *Applied Soft Computing*, 91:106263, June 2020. ISSN 1568-4946. doi: 10.1016/j.asoc.2020.106263. URL https://www.sciencedirect.com/science/article/pii/S1568494620302039.

Deng, X., Sun, H., Lees, A., Wu, Y., and Yu, C. TURL: Table Understanding through Representation Learning. *SIGMOD Record*, 51(1):33–40, June 2020. ISSN 01635808. doi: 10.1145/3542700.3542709. URL https://arxiv.org/abs/2006.14806v2. arXiv: 2006.14806 Publisher: Association for Computing Machinery.

Desai, S. and Durrett, G. Calibration of Pre-trained Transformers, October 2020. URL http://arxiv.org/abs/2003.07892. arXiv:2003.07892 [cs].

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Dinh, T., Zeng, Y., Zhang, R., Lin, Z., Gira, M., Rajput, S., Sohn, J.-y., Papailiopoulos, D., and Lee, K. LIFT: Language-Interfaced Fine-Tuning for Non-Language Machine Learning Tasks, October 2022. URL http://arxiv.org/abs/2206.06565. arXiv:2206.06565 [cs].

Dwork, C. and Roth, A. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–487, August 2014. ISSN 15513068. doi: 10.1561/0400000042. URL https://dl.acm.org/doi/abs/10.1561/0400000042. Publisher: Now Publishers Inc. PUB4850 Hanover, MA, USA.

Eggert, G., Huo, K., Biven, M., and Waugh, J. TabLib: A Dataset of 627M Tables with Context, October 2023. URL https://arxiv.org/abs/2310.07875v1. arXiv: 2310.07875 ISBN: 2310.07875v1.

Fatima, M. and Pasha, M. Survey of Machine Learning Algorithms for Disease Diagnostic. *Journal of Intelligent Learning Systems and Applications*, 09(01):1, 2017. doi: 10.4236/jilsa.2017.91001. URL http://www.scirp.org/journal/PaperInformation.aspx?PaperID=73781&#abstract. Number: 01 Publisher: Scientific Research Publishing.

Gardner, J., Popović, Z., Schmidt, L., and Allen, P. G. Subgroup Robustness Grows On Trees: An Empirical Baseline Investigation. *Advances in Neural Information Processing Systems*, 35:9939–9954, November 2022. doi: 10.48550/arxiv.2211.12703. URL https://arxiv.org/abs/2211.12703v1. arXiv: 2211.12703.

Gardner, J., Popovic, Z., and Schmidt, L. Benchmarking Distribution Shift in Tabular Data with TableShift, February 2024. URL http://arxiv.org/abs/2312.07577. arXiv:2312.07577 [cs].

Gates, B. The Age of AI has begun, March 2023. URL https://www.gatesnotes.com/The-Age-of-AI-Has-Begun.

Golkar, S., Pettee, M., Eickenberg, M., Bietti, A., Cranmer, M., Krawezik, G., Lanusse, F., McCabe, M., Ohana, R., Parker, L., Blancard, B. R.-S., Tesileanu, T., Cho, K., and Ho, S. xVal: A Continuous Number Encoding for Large Language Models, October 2023. URL http://arxiv.org/abs/2310.02989. arXiv:2310.02989 [cs, stat].

Gorishniy, Y., Rubachev, I., Khrulkov, V., and Babenko, A. Revisiting Deep Learning Models for Tabular Data. *Advances in Neural Information Processing Systems*, 34: 18932–18943, 2021.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. Why do tree-based models still outperform deep learning on tabular data? In *Advances in Neural Information Processing*, July 2022. URL https://arxiv.org/abs/2207.08815v1. arXiv: 2207.08815.

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C. C. T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O., Salim, A., Shah, S., Behl, H. S., Wang, X., Bubeck, S., Eldan, R., Kalai, A. T., Lee, Y. T., and Li, Y. Textbooks Are All You Need, October 2023. URL http://arxiv.org/abs/2306.11644. arXiv:2306.11644 [cs].

Haidich, A. B. Meta-analysis in medical research. *Hippokratia*, 14(Suppl 1):29–37, December 2010. ISSN 1108-4189. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049418/.

Harrison, D. and Rubinfeld, D. L. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, March 1978. ISSN 0095-0696. doi: 10.1016/0095-0696(78)90006-2. URL https://www.sciencedirect.com/science/article/pii/0095069678900062.

Ho, J., Jain, A., and Abbeel, P. Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems*, 2020-December, June 2020. ISSN 10495258. doi: 10.48550/arxiv.2006.11239. arXiv: 2006.11239 Publisher: Neural information processing systems foundation.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J., and Sifre, L. An empirical analysis of compute-optimal large language model training. *Advances in Neural Information Processing Systems*, 35:30016–30030, December 2022.

Hollmann, N., Müller, S., Eggensperger, K., and Hutter, F. TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second. In *The Eleventh International Conference on Learning Representations*, September 2023a. doi: 10.48550/arXiv.2207.01848. URL http://arxiv.org/abs/2207.01848. arXiv:2207.01848 [cs, stat].

Hollmann, N., Müller, S., and Hutter, F. Large Language Models for Automated Data Science: Introducing CAAFE for Context-Aware Automated Feature Engineering. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=9WSxQZ9mG7.

Huang, X., Khetan, A., Cvitkovic, M., and Karnin, Z. TabTransformer: Tabular Data Modeling Using Contextual Embeddings, December 2020. URL https://arxiv.org/abs/2012.06678v1. arXiv: 2012.06678.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12), March 2023. ISSN 15577341. doi: 10.1145/3571730. URL https://dl.acm.org/doi/10.1145/3571730. arXiv: 2202.03629 Publisher: ACM PUB27 New York, NY.

Jiang, C., Nian, Z., Guo, K., Chu, S., Zhao, Y., Shen, L., and Tu, K. Learning Numeral Embedding. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2586–2599, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.235. URL https://aclanthology.org/2020.findings-emnlp.235.

Jiang, Z., Araki, J., Ding, H., and Neubig, G. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, September 2021. ISSN 2307-387X. doi: 10.1162/tacl_a_00407. URL https://doi.org/10.1162/tacl_a_00407.

Jiang, Z., Zhang, Y., Liu, C., Zhao, J., and Liu, K. Generative Calibration for In-context Learning. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2312–2333, Singapore, December 2023. Association for Computational Linguistics.

Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S. N., and Weller, A. Synthetic Data – what, why and how?, May 2022. URL https://arxiv.org/abs/2205.03257v1. arXiv: 2205.03257.

Kadra, A., Lindauer, M., Hutter, F., and Grabocka, J. Well-tuned Simple Nets Excel on Tabular Datasets, November 2021. URL http://arxiv.org/abs/2106.11189. arXiv:2106.11189 [cs].

Kaggle. 2017 Kaggle Machine Learning & Data Science Survey, 2017. URL https://www.kaggle.com/datasets/kaggle/kaggle-survey-2017.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs, stat].

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., Stadler, M., Weller, J., Kuhn, J., and Kasneci, G. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103: 102274, April 2023. ISSN 1041-6080. doi: 10.1016/J.LINDIF.2023.102274. Publisher: JAI.

Kolesnikov, S. Wild-Tab: A Benchmark For Out-Of-Distribution Generalization In Tabular Regression, December 2023. URL http://arxiv.org/abs/2312.01792. arXiv:2312.01792 [cs].

Kong, K., Zhang, J., Shen, Z., Srinivasan, B., Lei, C., Faloutsos, C., Rangwala, H., and Karypis, G. OpenTab: Advancing Large Language Models as Open-domain Table Reasoners. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Qa0ULgosc9.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://github.com/kynkaat/improved-precision-and-recall-metric.

Lehmberg, O., Ritze, D., Meusel, R., and Bizer, C. A Large Public Corpus of Web Tables containing Time and Context Metadata. In *Proceedings of the 25th International Conference Companion on World Wide Web - WWW '16 Companion*, pp. 75–76, Montréal, Québec, Canada, 2016. ACM Press. ISBN 978-1-4503-4144-8. doi: 10.1145/2872518.2889386. URL http://dl.acm.org/citation.cfm?doid=2872518.2889386.

Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., and Gao, J. Multimodal Foundation Models: From Specialists to General-Purpose Assistants, September 2023a. URL http://arxiv.org/abs/2309.10020. arXiv:2309.10020 [cs].

Li, H., Su, J., Chen, Y., Li, Q., and Zhang, Z. SheetCopilot: Bringing Software Productivity to the Next Level through Large Language Models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b. URL https://openreview.net/forum?id=tfyr2zRVoK.

Li, X., Zhao, R., Chia, Y. K., Ding, B., Joty, S., Poria, S., and Bing, L. Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. In *The*

Twelfth International Conference on Learning Representations, 2024. URL https://openreview.net/forum?id=cPgh4gWZlz.

Liu, J., Wang, T., Cui, P., and Namkoong, H. On the Need for a Language Describing Distribution Shifts: Illustrations on Tabular Datasets, July 2023. URL http://arxiv.org/abs/2307.05284. arXiv:2307.05284 [cs].

Liu, S., Wei, Y., Lu, J., and Zhou, J. An Improved Evaluation Framework for Generative Adversarial Networks, July 2018. URL http://arxiv.org/abs/1803.07474. arXiv:1803.07474 [cs].

Manikandan, H., Jiang, Y., and Kolter, J. Z. Language Models are Weak Learners. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=559NJBfN20.

Marcus, G. Why Are We Letting the AI Crisis Just Happen?, March 2023. URL https://www.theatlantic.com/technology/archive/2023/03/ai-chatbots-large-language-model-misinformation/673376/. Section: Technology.

McElfresh, D., Khandagale, S., Valverde, J., C, V. P., Feuer, B., Hegde, C., Ramakrishnan, G., Goldblum, M., and White, C. When Do Neural Nets Outperform Boosted Trees on Tabular Data?, October 2023. URL http://arxiv.org/abs/2305.02997. arXiv:2305.02997 [cs, stat].

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. MTEB: Massive Text Embedding Benchmark, March 2023. URL http://arxiv.org/abs/2210.07316. arXiv:2210.07316 [cs].

Naeem, M. F., Oh, S. J., Uh, Y., Choi, Y., and Yoo, J. Reliable Fidelity and Diversity Metrics for Generative Models. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 7176–7185. PMLR, June 2020. URL http://proceedings.mlr.press/v119/naeem20a.html. Series Title: Proceedings of Machine Learning Research.

Ni, A., Iyer, S., Radev, D., Stoyanov, V., Yih, W.-t., Wang, S. I., and Lin, X. V. LEVER: Learning to Verify Language-to-Code Generation with Execution. In *International Conference on Machine Learning*, September 2023. doi: 10.48550/arXiv.2302.08468. URL http://arxiv.org/abs/2302.08468. arXiv:2302.08468 [cs].

Office for National Statistics. 2021 Census, 2021. URL https://www.ons.gov.uk/census.

Padhi, I., Schiff, Y., Melnyk, I., Rigotti, M., Mroueh, Y., Dognin, P., Ross, J., Nair, R., and Altman, E. Tabular Transformers for Modeling Multivariate Time Series. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3565–3569, June 2021. doi: 10.1109/ICASSP39728.2021.9414142. URL https://ieeexplore.ieee.org/abstract/document/9414142. ISSN: 2379-190X.

Patnaik, S., Changwal, H., Aggarwal, M., Bhatia, S., Kumar, Y., and Krishnamurthy, B. CABINET: Content Relevance-based Noise Reduction for Table Question Answering. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=SQrHpTllXa.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, February 2021. doi: 10.48550/arxiv.2103.00020. arXiv: 2103.00020.

Rana, M. S., Nobi, M. N., Murali, B., and Sung, A. H. Deepfake Detection: A Systematic Literature Review. *IEEE Access*, 10:25494–25513, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3154404. URL https://ieeexplore.ieee.org/abstract/document/9721302. Conference Name: IEEE Access.

Renda, A., Hopkins, A. K., and Carbin, M. Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains. In *ICML 2023 Workshop: Sampling and Optimization in Discrete Space*, 2023. URL http://people.csail.mit.edu/renda/llm-sampling-paper.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, December 2022. doi: 10.48550/arxiv.2112.10752. arXiv: 2112.10752.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J., and Norouzi, M. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *Advances*

in Neural Information Processing Systems, 35:36479–36494, May 2022. URL https://arxiv.org/abs/2205.11487v1. arXiv: 2205.11487.

Sajjadi, M. S. M., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing Generative Models via Precision and Recall. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., and Jitsev, J. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, December 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

Shwartz-Ziv, R. and Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion*, 81: 84–90, May 2022. ISSN 1566-2535. doi: 10.1016/J.INFFUS.2021.11.011. arXiv: 2106.03253 Publisher: Elsevier.

Solatorio, A. V. and Dupriez, O. REaLTabFormer: Generating Realistic Relational and Tabular Data using Transformers, February 2023. URL https://arxiv.org/abs/2302.02041v1. arXiv: 2302.02041.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C. B., and Goldstein, T. SAINT: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

Song, Y. and Ermon, S. Generative Modeling by Estimating Gradients of the Data Distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Spithourakis, G. P. and Riedel, S. Numeracy for Language Models: Evaluating and Improving their Ability to Predict Numbers. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2104–2115, 2018. doi: 10.18653/v1/P18-1196. URL http://arxiv.org/abs/1805.08154. arXiv:1805.08154 [cs, stat].

Thawani, A., Pujara, J., Szekely, P. A., and Ilievski, F. Representing Numbers in NLP: a Survey and a Vision, March 2021. URL http://arxiv.org/abs/2103.13136. arXiv:2103.13136 [cs].

Theis, L., Van Den Oord, A., and Bethge, M. A note on the evaluation of generative models. *4th International Conference on Learning Representations*, 2016.

Thornton, A. and Lee, P. Publication bias in meta-analysis: its causes and consequences. *Journal of Clinical Epidemiology*, 53(2):207–216, February 2000. ISSN 0895-4356. doi: 10.1016/S0895-4356(99)00161-4. URL https://www.sciencedirect.com/science/article/pii/S0895435699001614.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C. C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P. S., Lachaux, M.-A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E. M., Subramanian, R., Tan, X. E., Tang, B., Taylor, R., Williams, A., Kuan, J. X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., and Scialom, T. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Tucker, A., Wang, Z., Rotalinti, Y., and Myles, P. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digital Medicine 2020 3:1*, 3(1):1–13, November 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00353-9. URL https://www.nature.com/articles/s41746-020-00353-9. Publisher: Nature Publishing Group.

van Breugel, B. and van der Schaar, M. Beyond Privacy: Navigating the Opportunities and Challenges of Synthetic Data, April 2023. URL http://arxiv.org/abs/2304.03722. arXiv:2304.03722 [cs].

van Breugel, B., Seedat, N., Imrie, F., and van der Schaar, M. Can You Rely on Your Model Evaluation? Improving Model Evaluation with Synthetic Test Data. In *Advances in Neural Information Processing (NeurIPS 2023)*, October 2023a. arXiv: 2310.16524.

van Breugel, B., Sun, H., Qian, Z., and van der Schaar, M. Membership Inference Attacks against Synthetic Data through Overfitting Detection. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023b.

Wang, Q., Gao, J., Lin, W., and Yuan, Y. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8198–8207, 2019.

Wang, Z., Cai, S., Chen, G., Liu, A., Ma, X., and Liang, Y. Describe, Explain, Plan and Select: Interactive Planning with LLMs Enables Open-World Multi-Task Agents. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KtvPdGb31Z.

Wang, Z., Zhang, H., Li, C.-L., Eisenschlos, J. M., Perot, V., Wang, Z., Miculicich, L., Fujii, Y., Shang, J., Lee, C.-Y., and Pfister, T. Chain-of-Table: Evolving Tables in the Reasoning Chain for Table Understanding. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=4L0xnS4GQM.

Yang, Y., Wang, Y., Liu, G., Wu, L., and Liu, Q. UniTabE: A Universal Pretraining Protocol for Tabular Foundation Model in Data Science. In *The Twelfth International Conference on Learning Representations*, October 2023. URL https://openreview.net/forum?id=6LLho5X6xV.

Ye, C., Lu, G., Wang, H., Li, L., Wu, S., Chen, G., and Zhao, J. Towards Cross-Table Masked Pretraining for Web Data Mining. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*. ACM, July 2023. arXiv: 2307.04308.

Yin, P., Neubig, G., Yih, W. T., and Riedel, S. TaBERT: Pretraining for Joint Understanding of Textual and Tabular Data. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 8413–8426, May 2020. ISSN 0736587X. doi: 10.18653/v1/2020.acl-main.745. URL https://arxiv.org/abs/2005.08314v1. arXiv: 2005.08314 Publisher: Association for Computational Linguistics (ACL) ISBN: 9781952148255.

Yoon, J., Jordon, J., and Van Der Schaar, M. RadialGAN: Leveraging multiple datasets to improve target-specific predictive models using Generative Adversarial Networks. *35th International Conference on Machine Learning, ICML 2018*, 13:9060–9068, February 2018. arXiv: 1802.06403 Publisher: International Machine Learning Society (IMLS) ISBN: 9781510867963.

Yoon, J., Drumright, L. N., and Van Der Schaar, M. Anonymization through data synthesis using generative adversarial networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, August 2020. ISSN 21682208. doi: 10.1109/JBHI.2020.2980262. Publisher: Institute of Electrical and Electronics Engineers Inc.

Yuan, L., Chen, Y., Wang, X., Fung, Y., Peng, H., and Ji, H. CRAFT: Customizing LLMs by Creating and Retrieving from Specialized Toolsets. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=G0vdDSt9XM.

Zhang, H., Wen, X., Zheng, S., Xu, W., and Bian, J. Towards Foundation Models for Learning on Tabular Data, October 2023. URL http://arxiv.org/abs/2310.07338. arXiv:2310.07338 [cs].

Zhao, Z., Birke, R., and Chen, L. Y. TabuLa: Harnessing Language Models for Tabular Data Synthesis. *Proceedings of ACM Conference (Conference'17)*, 1, October 2023. URL http://arxiv.org/abs/2310.12746. arXiv: 2310.12746.

Zhu, B., Sheng, Y., Zheng, L., Barrett, C., Jordan, M., and Jiao, J. Towards Optimal Caching and Model Selection for Large Model Inference. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL https://openreview.net/forum?id=gd20oaZqqF.

Zhu, B., Shi, X., Erickson, N., Li, M., Karypis, G., and Shoaran, M. XTab: Cross-table Pretraining for Tabular Transformers. In *International Conference on Machine Learning*, May 2023b. URL https://arxiv.org/abs/2305.06090v1. arXiv: 2305.06090.

Zhu, J. Y., Park, T., Isola, P., and Efros, A. A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October: 2242–2251, March 2017. ISSN 15505499. doi: 10.1109/ICCV.2017.244. URL https://arxiv.org/abs/1703.10593v7. arXiv: 1703.10593 Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9781538610329.

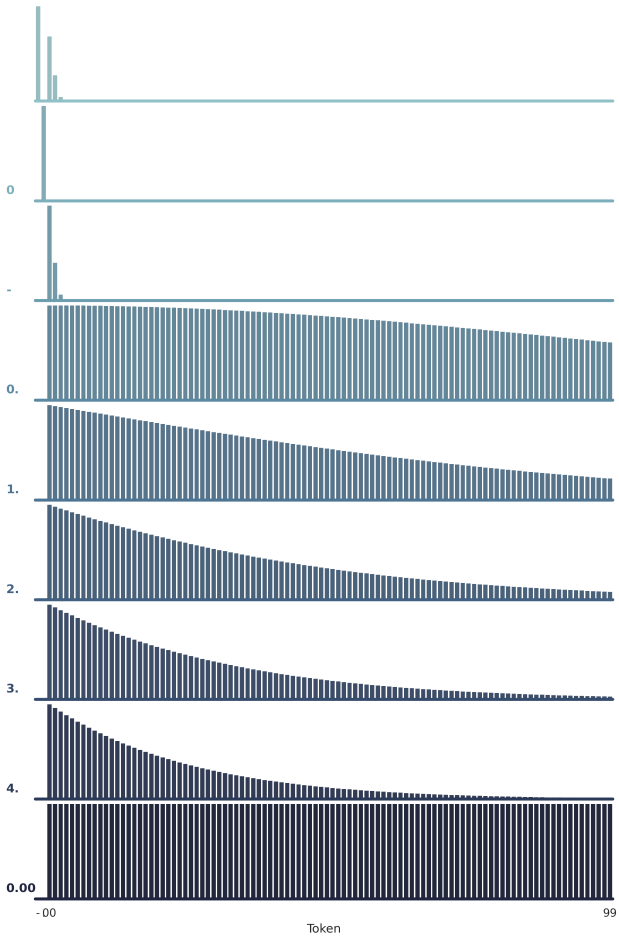

Figure 2. **Sampling continuous distributions using LLMs autoregressively is inefficient**. Assume we autoregressively sample tokens aiming to generate numbers that follow a standard Gaussian. What token probabilities should the LLM output at each sampling step? Let us consider a total vocabulary of just 102 tokens, [" − ", ".", "00", ..., "99"]. For different direct histories of generated text (examples given by each row, already generated digits on left), the output probabilities need to be very different. For example, a single "draw" is generated by sampling from probabilities in the first row (e.g. giving"0"), then second row (conditional on "0", giving "."), fourth row ("0."→ e.g. "00"), and last row ("0.00"→ e.g. "00").

A. Details Figure 1

The rough estimates in Figure 1 are created by checking the titles and abstracts of all accepted papers on the presence of certain keywords, and plotting the number of counts per modality which describe "foundation model" in some way. The keywords for "foundation model" and each modality are given in Table 1. Papers with multiple modalities are counted towards each modality. The total number of accepted papers per conference are 1828 (ICML 2023), 3218 (NeurIPS 2023), and 2250 (ICLR 2024).

Table 1. Key words for each search term. Abstracts are made lower-case. We intentionally do not include "text" and "time" as keywords for respectively Language and Time Series applications, as we found this lead to many false positives. * For "table" and "llm", we use a word boundary ($\backslash b$) at the start of the keyword, to avoid common false positives (e.g. "portable", "mutable", "bellman").

Research Topic	Keywords
Foundation Model	foundation model, llm*, cross-dataset
Graph	graph
Image	image, vision
Language	language, llm
Table	tabular, table*, cross-table
Time Series	time-series, time series, temporal

Papers that satisfy both "foundation model" and "table" and thus form "Table" group in Figure 1 are the following per conference: ICML 2023 (Zhu et al., 2023b; Ni et al., 2023), NeurIPS 2023 (Li et al., 2023b; Zhu et al., 2023a; Ajay et al., 2023; Wang et al., 2023; Hollmann et al., 2023b; Manikandan et al., 2023), ICLR 2024 (Li et al., 2024; Wang et al., 2024; Patnaik et al., 2024; Kong et al., 2024; Yuan et al., 2024; Bao et al., 2024; Yang et al., 2023). We note these include LLM works that are evaluated on tabular tasks, e.g. Table QA (Ni et al., 2023).

B. Benchmarking LTMs

For LTMs, benchmarks will be very important, yet the breadth of their applications will make it hard to fairly compare models. More research into LTM benchmarks is essential. For now, we would like to indicate a number of relevant benchmarks. We split this up in ML tasks and experimental set-ups, where the latter aims to measure adaptability of the LTM for different data settings.

Tasks could include:

1. **Supervised learning**. Predictive performance can be measured either through using the LTM for retrieving row representations that can be used by a small downstream predictor, or by using the LTM directly (i.e. by generating the target conditional on features).
2. **Synthetic data generation**. For generative LTMs, conditional and unconditional generation quality can be measured similar to more traditional generative models. Any synthetic data metric can be used, e.g. train-on-synthetic-test-on-real performance (for downstream utility) (Jordon et al., 2022), fidelity and diversity metrics (Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Naeem et al., 2020), and ϵ-identifiability (for privacy) (Yoon et al., 2020).
3. **Imputation**. Generative LTMs can be used for imputation, and benchmarked on this. This can be further split up in single imputation tasks (where the aim is to model $\mathbb{E}(X_{unobserved}|X_{observed})$) or multiple imputation (where the aim is to sample from $p(X_{unobserved}|X_{observed})$). The latter is only relevant for generative LTMs (as representation learning-based LTMs only provide point estimates). Another axes could be the missingness mechanism (MCAR, MAR, MNAR).
4. **LTMs for science.** This includes a range of tasks that require more research and adaptation beyond the current line of LTM work, but which could help describe success in some of the applications described in Section 4.3. For example, subtasks could include dimensionality reduction, data cleaning (e.g. outlier detection), clustering, and cross-dataset tasks (e.g. "which of the following datasets is most relevant to [some other dataset or question]").

Experimental set-ups may include:

1. **Few-shot.** Performance on new datasets for which relatively few samples are available. This could be split up further into approaches that allow for finetuning of the LTM, and methods that do not (e.g. in-context learning for LLMs). For this setting, both performance and robustness of the LTM (compared to baseline methods trained on the same few samples) could be a deciding factor in measuring the LTMs success. This closely relates to out-of-distribution and distributional shift benchmarking, for which good benchmarks exist (Kolesnikov, 2023; Gardner et al., 2024).

2. **Zero-shot.** This is the same as few-shot, but without any samples from the target dataset. This requires true generalization of the LTM, and is likely not achieved by the current LTMs.

3. **In-distribution.** The performance of an LTM on hold-out test sets from the datasets on which the LTM was trained. This is less interesting for foundation models, as we generally want adaptability—i.e. genereralizability *beyond* the training data. Nonetheless, in-distribution evaluation would be useful to quantify generalization gaps in the previous two settings (e.g. compare few-shot performance w.r.t. in-distribution experiments).

One difficulty in these experimental setups is that pretrained and published foundation models do not always come with descriptions of the precise data (or splits thereof) on which they were trained. This is a ubiquitous problem in foundation model evaluation, and especially problematic when one assumes they are measuring few-shot or zero-shot performance, but actually data leakage has occurred.