Had Enough of Experts?
Quantitative Knowledge Retrieval from Large Language Models

David Selby*[1] Kai Spriestersbach*[1] Yuichiro Iwashita*[1,2] Dennis Bappert[3] Archana Warrier[1]
Sumantrak Mukherjee[1] Muhammad Nabeel Asim[1] Koichi Kise[1,2] Sebastian Vollmer[1]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern, Germany
[2]Graduate School of Informatics, Osaka Metropolitan University, Japan
[3]Amazon Web Services, Mainz, Germany

Abstract

Large language models (LLMs) have been extensively studied for their abilities to generate convincing natural language sequences, however their utility for quantitative information retrieval is less well understood. In this paper we explore the feasibility of LLMs as a mechanism for quantitative knowledge retrieval to aid data analysis tasks such as elicitation of prior distributions for Bayesian models and imputation of missing data. We present a prompt engineering framework, treating an LLM as an interface to a latent space of scientific literature, comparing responses in different contexts and domains against more established approaches. Implications and challenges of using LLMs as 'experts' are discussed.

1 INTRODUCTION

Automated solutions for life sciences, industrial and governmental processes demand, in the learning phase, large amounts of data, but these may be unavailable or incomplete. Small samples and missing data increase the risk of overfitting, weakening the validity, reliability and generalizability of statistical insights. Transfer learning and self-supervised learning have proven effective in addressing the issue of large scale data annotation in fields such as computer vision and natural language processing, but are not a panacea.

To overcome limitations of small data, analysts employ two approaches. Firstly, data-based or empirical methods maximize information extraction, through imputation models—such as mean imputation—and data augmentation. However, this is limited by the size, availability and representativeness of training data. Alternatively, one can exploit prior information, through application of knowledge graphs or expert-elicited Bayesian priors, allowing for sparser models and handling of missing values. This latter approach is constrained by the difficulty, cost and myriad different methods of obtaining and eliciting subjective and heterogeneous opinions from experts, then translating them into a form amenable to quantitative analysis (Falconer et al., 2022).

Large language models (LLMs), also known as foundation models (Narayan et al., 2022), are generative models capable of producing natural language texts based on a given prompt or context. LLMs such as GPT-4 have been used in various applications, such as chatbots, summarization and content creation. In the quantitative sciences, LLMs have been applied to mostly qualitative tasks such as code completion, teaching of mathematical concepts (Wardat et al., 2023) and offering advice on modelling workflows or explaining data preparation pipelines (Barberio, 2023; Hassani and Silva, 2023). Some work has also applied LLMs to mathematical reasoning and symbolic logic (He-Yueya et al., 2023; Orrù et al., 2023). When linked with certain application programming interfaces (APIs), or incorporated into a retrieval-augmented generation (RAG) tool, some LLM frameworks (e.g. Ge et al., 2023) are also capable of evaluating code, connecting to other data analysis tools or looking up supporting information (Nicholson et al., 2021; Kamalloo et al., 2023). However, the capabilities of large language models to retrieve accurate and reliable *quantitative* information are less well-explored. In this paper, we explore the possibility of using LLMs including GPT-4 to tackle the 'small data' problem by treating the model as an (indirect) interface to the large and diverse body of scientific and technical knowledge contained within its training corpus, and hence using an LLM to generate 'expert'-guided estimates for imputing missing values and deriving informative prior distributions.

Our motivation is the question: can large language models be treated as experts having read a large sample of the scientific literature (although the exact training corpus is not known, so 'openness' is debatable; see Liesenfeld et al., 2023) and thus might be treated as an accessible interface

*These authors contributed equally to this work to this knowledge (Petroni et al., 2019; Singhal et al., 2023) pertaining to the task of prior elicitation and imputation.

Contributions of this paper are twofold. Firstly, we present a prompt engineering framework for zero-shot missing data imputation, based on LLMs playing 'expert' roles derived from metadata such as the dataset description. This is followed by an empirical evaluation of LLM imputation quality and the impact on downstream tasks, compared with baseline approaches on a diverse set of 50 real-world datasets across different domains. Secondly, we develop a chain-of-thought-based methodology to elicit prior distributions from general purpose LLMs, emulating real-world knowledge elicitation protocols. LLM-elicited priors are compared with those from human experts, and the quantitative value of LLM 'expertise' is evaluated for several tasks. Code to reproduce our experiments is available on GitHub.

2 RELATED WORK

2.1 NUMERACY IN LANGUAGE MODELS

Language models have been noted for their remarkable ability to act as unsupervised knowledge bases (Petroni et al., 2019). One might expect a *language* model like GPT-4 to handle text, code and sequences, but unimbued with access to external tools, the capabilities with numerical data are less obvious. Noever and McKee (2023) and Cheng and Zhang (2023) discuss the 'emergent' numeracy skills of LLMs, from early models unable to perform simple addition to later versions able to compute correlations. Hopkins et al. (2023) showed that repeated sampling from LLMs does not yield reasonable distributions of random numbers, making them poor data generators. Xiong et al. (2023) also suggested LLMs tend to underestimate uncertainty. It has been hypothesized that *mode collapse*, in models fine-tuned via reinforcement learning with human feedback, inhibits the diversity of outputs (Anonymous, 2023).

The design, adaptation and use of LLMs to assist data analytical tasks is a hotly explored topic—a comprehensive review is beyond the scope of this article. Microsoft Research AI4Science and Microsoft Azure Quantum (2023) surveyed LLMs' use in scientific discovery, finding numerical calculation abilities, without connecting to external tools, left room for improvement. Most LLM-based data science tools focus on tasks such as code generation for analysis scripts (Megahed et al., 2023), visualization (Dibia, 2023; Maddigan and Susnjak, 2023) and connection of LLMs with external APIs (Ge et al., 2023). Similarly, LLMs fine-tuned on scientific texts may be used to extract qualitative information, such as chemical formulae or entity relations (Dunn et al., 2022). Typically, queries about data imputation to an LLM-driven chatbot will yield template Python code for performing mean imputation, or general advice about the relative merits of imputation techniques. Similarly, a conversation with ChatGPT about prior elicitation could generate textual advice about how to elicit priors from experts. Ahmad et al. (2023) proposed a data cleaning model that combines a fine-tuned foundation model augmented with retrieval from a user-supplied data lake. Here, however, we are interested in evaluating the intrinsic ability of an LLM to retrieve latent quantitative information directly; that is, not to perform mathematical operations on an input dataset nor to offer code or advice on how to do so, rather to offer educated numerical suggestions based on its large training corpus containing specialist technical knowledge.

2.2 DATA PROCESSING WITH LLMS

There is some promise in converting data into natural language inputs for an LLM to perform preprocessing: Narayan et al. (2022) tested GPT-3 on entity matching, error detection and data imputation tasks, in zero-shot and few-shot settings. Their approach involved serializing tabular data and tasks into a natural language format using manually tuned prompt templates. Vos et al. (2022) explored prefix tuning as an alternative to full fine tuning of an LLM for such tasks; whereas Zhang et al. (2023b) compared GPT-3.5, GPT-4 and Vicuna-13B in a data preprocessing framework, later developing Jellyfish-13B, an open-source LLM fine-tuned specifically for data preprocessing (Zhang et al., 2023a). Li et al. (2023b)'s Table-GPT describes a framework for fine-tuning language models on 'table tasks', including finding and predicting missing values. Separately, Chen et al. (2023) utilized fine tuning in tandem with a graph attention mechanism to impute spatiotemporal data. Nazir et al. (2023) further explored the capability of ChatGPT in missing value imputation, focussing on imputation quality (see subsection 3.2) in psychological and biological data. An alternative approach to LLM-assisted data analysis involves using only the model's encoder to project natural language representations of a data vector into a latent space, then performing anomaly detection on the resulting numeric latent vector (Lopatecki et al., 2023a,b).

However, the level of 'expertise' offered by pretrained LLMs on quantitative tasks across different domains has not yet been extensively studied, nor the effect of LLM imputations on performance in downstream tasks.

2.3 EXPERT KNOWLEDGE ELICITATION

Prior distributions are just one form of knowledge elicited from domain experts; others include feature engineering, model explanations and labelling heuristics, but in each case the process of elicitation typically involves interviews, written correspondence or interaction with a custom computer app (Kerrigan et al., 2021). A good expert-elicited prior distribution can help a statistical model effectively represent the data generating process, although due to various practical, technical and societal factors, prior elicitation is not yet widespread practice. As Mikkola et al. (2023) note, a lack of standardized software means there is no way for an analyst building a model, e.g. in Stan or Pyro, to 'launch an elicitation interface to elicit for their specific model'.

This lacuna might be filled with improved tools to facilitate interaction with human experts, or even to extract information automatically from existing knowledge bases, bypassing human experts altogether. Li et al. (2023a) describe LM-*driven* elicitation, using a chatbot or language model to assist information elicitation from human experts, making the process interactive. In engineering, LLMs have been employed in generating (and responding to) requirements elicitation surveys (White et al., 2023; Ronanki et al., 2023; Görer and Aydemir, 2023).

Natural language processing is already extensively used to extract quantitative information from large academic corpora, with the aim of aiding data-driven scientific research (see, e.g. Olivetti et al., 2020). In a similar vein, prior distributions may be elicited from academic publications: Linde et al. (2023) describe the use of a systematic review of the biomedical literature to perform 'data-driven' prior elicitation (see also Rietbergen et al., 2011; van de Schoot et al., 2018). A meta-analytic-predictive (MAP) prior uses historical data to reduce the number of subjects needed in clinical trial design (Weber et al., 2021). To our knowledge, the feasibility of eliciting prior distributions from a 'domain expert' in the form of an LLM has not yet been explored.

A significant limitation of treating a 'black box' model as a domain expert is its lack of transparency or accountability. On the other hand, without the ability to read minds, it is impossible to know exactly why a human expert gives a particular answer or why it might differ from other experts or from the same person's responses under different circumstances. Information retrieved from humans and machines alike may be sensitive to the way it is elicited. LLMs may, for example, 'encode clinical knowledge' (Singhal et al., 2023), but whereas one can scrutinize the credentials of any human expert, the training corpora of closed, proprietary models like GPT-4 and GPT-3.5 are not published[1] and the provenance of its outputs are not always attributable (Kamalloo et al., 2023). Retrieval augmented generation tools (Lewis et al., 2020), such as Microsoft's Bing Copilot, perplexity.ai and Scite Assistant (Nicholson et al., 2021), attempt to provide supporting sources alongside their generated responses, though factual grounding is not guaranteed and different prompts can still yield different answers (Andriopoulos and Pouwelse, 2023).

[1]Nevertheless, it is possible to reveal a closed LLM's training data through adversarial attacks (Nasr et al., 2023).

2.4 ELICITATION FRAMEWORKS

Several elicitation protocols have been developed to mitigate cognitive biases and combine the judgements of multiple experts (O'Hagan, 2019). The Sheffield Elicitation Framework (SHELF; Gosling, 2018) describes a collection of methods for eliciting a distribution based on aggregated opinions of multiple experts, through group discussion guided by a facilitator. As well as training participants in basic probability and statistics, the protocol includes various ways of eliciting a univariate distribution, such as the 'roulette method', where experts assign (virtual) chips to equally-spaced bins to form a histogram. Alternatively, the quartile method (or 'Sheffield method'; European Food Safety Authority, 2014) uses a series of questions to elicit a reference range, median, upper and lower quartiles of a distribution. Cooke's 1991 method pools the distributions of multiple experts, weighted according to their respective 'calibration' (ability to estimate a distribution whose ground truth is known) and 'information' (concentration or vagueness of the prior). The Delphi method uses the quartile method, iteratively refined over successive survey rounds using anonymized feedback from other participants. In this paper, however, we consider only single-agent LLMs with a zero-shot approach.

Perhaps one of the biggest benefits of extracting opinions from pre-trained models is that the early part of the elicitation process—teaching statistical literacy—can be omitted. As demonstrated in Appendix A, ChatGPT 3.5 is aware of the 'roulette' method, the SHELF protocol and most parametric probability distributions, with minimal exposition or disambiguation necessary. It remains to investigate the utility of this pre-training in practical problems.

3 EVALUATING EXPERTISE

3.1 WHAT MAKES A GOOD PRIOR?

Bayesian statistics involves decisionmaking based on a posterior distribution, $p(\theta|D) \propto \pi(\theta)\prod_{i=1}^{n} p(x_i|\theta)$, where $\pi(\theta)$ denotes the prior distribution. The definition of a 'good' prior distribution—like Bayesian statistics itself—is subjective, depending on the analyst's understanding of the purpose of expert-elicited information. No standard benchmark exists for expert-elicited prior distributions; a prior is a function of the expert and the elicitation method, as well as of the predictive task (Gelman et al., 2017). One purpose of prior information is to reduce amount of data needed. Another is to treat expert knowledge and observed data as complementary sources of information about a natural process.

In practice, any statistical model is at least slightly misspecified, but a prior—be it elicited from human or machine—can still be *informative*, *realistic* and *useful*. An informative expert prior is different from a non-informative (e.g.

Jeffreys') or default prior, but in this context the term 'informative' does not necessarily mean 'correct' but rather *concentrated* or confident, rather than vague and uninformative, even (and indeed especially) if it later turns out to be in conflict with the data observed. Realistic priors—as extracted from language models—should align with those from human experts or be otherwise verifiable on external data. 'Useful' means superior posterior predictive performance on a downstream task, improving expected utility relative to reference priors (and human experts, where available). Meanwhile, an elicitation framework should be *robust*, *coherent*, *consistent*, *flexible* and *efficient*. That is, the distributions produced should not be too sensitive to slight rewording of prompts, the expert should not have already seen the data (otherwise it is not a prior!), we should avoid generating priors that are mutually incompatible, and the framework should be readily adaptable to new or larger problems.

A measurement of the informativeness of a prior distribution is the prior effective sample size (Morita et al., 2008; Neuenschwander et al., 2020). However, this does not measure improvement on downstream tasks, but rather how many data points one would need to get similar peakiness/curvature around the posterior mode. Another measure is the Bayesian log posterior predictive density, or lppd (McElreath, 2016)—also called log loss—or the continuous ranked probability score (CRPS), a proper scoring rule used in weather forecasting (Gneiting and Raftery, 2007). We can estimate both metrics using the posterior predictive distribution

$$p(\mathbf{x}'|D) = \mathbf{E}_{p(\boldsymbol{\theta}|D)}[p(\mathbf{x}'|\boldsymbol{\theta})]$$

on a hold out data set. See also Wilde et al. (2021) for a similar setting that quantifies the utility of synthetic data in Bayesian setting.

Nevertheless, it makes little sense—in measuring an expert's domain knowledge—to perform only abstract simulations comparing with known data generating distributions. Evaluating quantitative information in language models necessitates methods grounded in the real world.

3.2 IMPUTATION EVALUATION

Jäger et al. (2021) describe two principles of benchmarking imputation methods: *imputation quality* and *downstream evaluation*. Imputation quality—or upstream performance—measures the extent to which an imputation method can accurately recover artificially missing values. Downstream evaluation involves training a supervised learning model on the imputed dataset and measuring its predictive performance. Imputation quality for continuous features can be calculated using the root mean square error RMSE $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$ where x_i represents the original discarded value and $\hat{x}_i$ the output of imputation, whereas for categorical features, imputation quality can be calculated via the F_1 score, $F_1 = 2(\text{recall}^{-1} + \text{precision}^{-1})^{-1}$, the harmonic mean of precision and recall. As RMSE is unbounded, inter-dataset comparison is made possible by using a normalised version: NRMSE $= \frac{\text{RMSE}}{x_{\max} - x_{\min}}$ where $x_{\max}$ and $x_{\min}$ represent the maximum and minimum value in the original values. Downstream performance can be measured by improvement $= \frac{\text{score(imputed)} - \text{score(incomplete)}}{\text{score(incomplete)}}$ where score(imputed) represents the prediction score of the model trained and tested on the imputed data, and score(incomplete) represents the prediction score of the model trained and tested on the incomplete data. This represents how much downstream performance has improved compared to training and testing with incomplete data.

4 PROMPTING METHODOLOGY

Impersonating a human domain expert can improve an LLM's performance at related tasks (Salewski et al., 2023). Nevertheless, in response to scientific questions, especially on potentially sensitive topics, such as healthcare advice, language models often prevaricate. A quantitative knowledge retrieval system should therefore prompt the LLM to play the role of an expert and carefully specify the task at hand to ensure contextually relevant information is returned in the appropriate format. In this section, we present a framework for eliciting 'expert' advice in the form of imputed values and prior distributions from an LLM.

Algorithm 1 Data imputation

Input: Dataset, dataset description, prompt templates
  epi ← EPI(dataset description, epi prompt templates)
  for all row ← dataset **do**
    if row contains missing values **then**
      for all missing value ← row **do**
        ds ← Data serialization(row)
        system prompt ← epi + system suffix
        elicited value ←
        TS(system prompt, ds, prompt templates)
      end for
    end if
  end for
Output: Imputed data

Algorithm 2 Expert prompt initialization (EPI)

Input: Data description, prompt templates
  user prompt ← prefix + data description + suffix
  epi prompt ← LLM(system prompt, user prompt)
  return epi prompt
Output: System prompt describing expert role

Expert roles To optimally prime the LLM for contextually accurate missing value imputation and prior elicitation,

Algorithm 3 Data serialization (DS) for data imputation

Input: Target row
  for all variable name, value $\leftarrow$ target row **do**
    if value is missing **then**
      ds $\leftarrow$ ds + "The {variable name} is <missing>."
    else
      ds $\leftarrow$ ds + "The {variable name} is {value}."
    end if
  end for

Algorithm 4 Task specification (TS)

Input: system prompt, ds, ts prompt templates
  user prompt $\leftarrow$ prefix + ds + suffix
  elicited value $\leftarrow$ LLM(system prompt, user prompt)
  return elicited value

it is important to establish an initialization context emulating the expertise of a human specialist. This necessitates creation of a detailed system prompt defining expert roles for each specific dataset. For efficiency and scalability, we used the generative capabilities of the LLM itself to write the initial system prompts. We develop a template prompt including the respective dataset description to generate a system prompt as 'role description' for an expert who will perform the imputations. This step is executed only once per dataset. The *expert prompt initialization* (EPI) module (Algorithm 2) introduces the task with the words "I am going to give you a description of a dataset. Please read it and then tell me which hypothetical persona would be the best domain expert..." This introduction is followed by information derived from existing metadata, including a dataset description and a list of column names. The LLM then returns a short biography of the form "You are a..." describing the role to be played in future queries. This is then used as the system prompt. Further details are given in Appendix B.

Data serialization While some papers (e.g. Zhang et al., 2023a; Li et al., 2023b) give the data to the LLM in a tabular format, we input the data to the LLM in a natural language form to let the LLM to behave as a human domain expert. The *data serialization* (DS) module (Algorithm 3) converts the data into a natural language form. See Appendix C.

Task specification Querying the LLM about missing value imputation or prior elicitation will typically yield generic advice, suggest R or Python analysis code or refer the user to consult a real expert. The user prompt must insist that the agent returns numeric information, preferably in a consistent format so that large numbers of such responses can be parsed programmatically. The *task specification* (TS) module (Algorithm 4) asks the LLM with clear task instructions and returns the elicited value. See Appendix D for further details.

Temperature In pursuit of reproducibility and to avoid redundant computation through stochastic sampling, our framework assumes a temperature setting of zero. An LLM data imputer does not necessarily require linguistic capabilities of more general purpose language model applications associated with a temperature setting greater than zero (except, perhaps, in expert prompt initialization). The primary objective is the precise and contextually appropriate filling of data gaps, rather than the generation of diverse or creative text. It is a widely held belief that lower temperature settings produce more deterministic results, although Ouyang et al. (2023) demonstrated that this is not always the case. Nonetheless, it is beyond the scope of the present study to explore the impact that different temperature parameters have on the quality of data imputation or elicited knowledge. This may be investigated in future work.

5 EXPERIMENTS

An empirical evaluation of an LLM's 'real world' knowledge necessarily precludes purely abstract simulation-based studies. This motivates selection of series of datasets based on—ideally—careful representative sampling of real world measurements. Like any human expert, the model can be assumed to know more about some topics than others, which we attempt to explore through evaluation on a broad range of domain areas including medicine, biology, economics, engineering, social sciences, and psychology.

5.1 DATA IMPUTATION

5.1.1 Datasets

The OpenML-CC18 Curated Classification benchmark (Bischl et al., 2017) comprises 72 classification datasets from OpenML, based on real-world binary or multi-class classification tasks, in a variety of domains from credit scoring to biology, medicine and marketing. Use of this collection ensures our experiments cover a wider set of domains than previous work on LLM data imputation; meanwhile a pre-specified benchmark mitigates the risk of 'cherry picking'.

Datasets in CC18 have sample sizes (n) from about 500 to tens of thousands, with numbers of features (p) ranging from 5 to 3073. Though all ostensibly provided in dense tabular format, some are actually drawn from other modes; for example, the MNIST and CIFAR-10 imaging datasets are represented as wide tables, with columns corresponding to individual pixels. It is fair to assume that any human-like expert is unlikely to make particularly informed imputations about such features.

To evaluate the imputation quality, we used all datasets that have no missing values, which is 64 datasets. We then split the datasets into training and test sets, with 80% of the samples in the training set and the remaining 20% in the test set. For each dataset, we artificially generated missing values based on the missing at random (MAR) missingness pattern, where the probability of a value being missing depends only on the observed values. The number of features including missing values was set to $\min(\#\text{all features}, 3)$, and the number of samples with missing values was set to 40 for the training set and 10 for the test set. This implementation was based on the Jenga library by Schelter et al. (2021).

5.1.2 Imputation methods

Building on Jäger et al. (2021) and Nazir et al. (2023), our LLM data imputer is compared with 3 empirical approaches: mean and mode imputation (for continuous and categorical features, respectively), k-nearest neighbours (k-NN) imputation (the mean/mode of the k nearest samples) and random forest imputation. See Appendix E for further details. The LLM-based data imputer was powered by LLaMA 2 13B Chat, LLaMA 2 70B Chat (Touvron et al., 2023), Mistral 7B Instruct (Jiang et al., 2023) and Mixtral 8x7B Instruct (Jiang et al., 2024), each evaluated separately.

5.1.3 Evaluation

We imputed the missing values in the training and test sets using the LLM data imputer and the baseline methods. We then calculated the imputation quality for each feature using the NRMSE and F_1 score for continuous and categorical features, respectively. For NRMSE, we dropped the result when the denominator was zero.

We trained a random forest classifier on the training set and evaluated its performance on the test set. For the classifier, we used RandomForestClassifier from scikit-learn version 1.3.2 with hyperparameters set to their defaults. We then evaluated the downstream performance based on the metrics described in subsection 3.2.

5.1.4 Results

We evaluated the imputation quality of the LLM data imputer and the baseline methods. Figure 1 shows the imputation quality for continuous and categorical features. The domain for each dataset was decided manually since it was not given in the original dataset. See Appendix F for details. According to the figure, contrary to expectations, the overall imputation quality of the LLM data imputer was not as good as the 3 empirical methods. However, it was suggested that LLM-based imputation can be utilized for several datasets. For example, some datasets in the engineering and computer vision domain, such as 'pc1', 'pc3' and 'satimage' dataset, had the quality of around NRMSE $= 0.1$, and some datasets in the biology and NLP domain, such as 'Internet-Advertisements' and 'dna', had the quality of

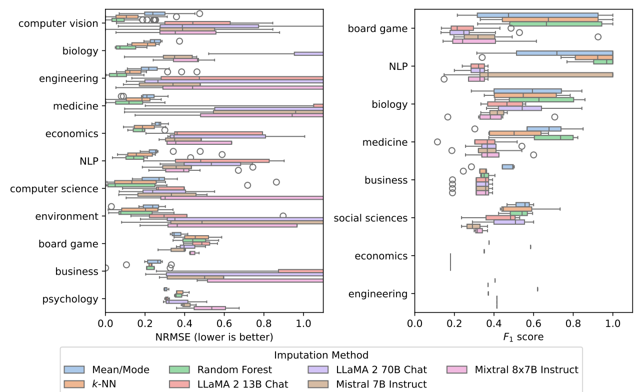


Figure 1: Imputation quality of different models on continuous and categorical features, plotted by domain category.

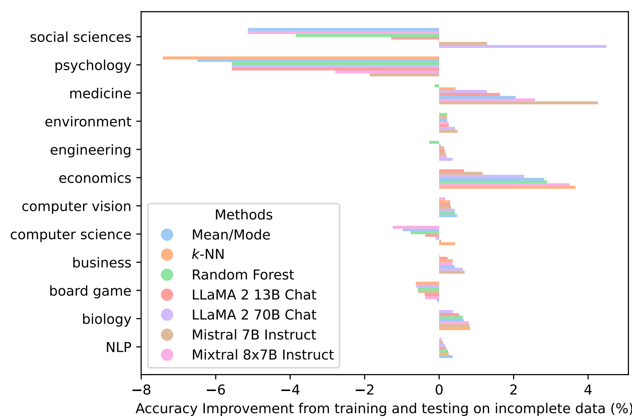


Figure 2: Downstream performance of different models, plotted by domain category.

around $F_1 = 0.7 - 1.0$. On the other hand, the quality of the LLM-based imputer was poor in the economics and business domain, such as 'credit-g' and 'bank-marketing' dataset, with the imputation quality of around $F_1 = 0.2$.

Figure 2 shows that some domains have good downstream performance with LLM-based imputation, while others do not. For example, social sciences and psychology have poor downstream performance, while medicine, economics, business and biology have good downstream performance. In particular, the LLM-based imputer achieved the best performance in the economics domain.

5.2 PRIOR ELICITATION

5.2.1 Datasets

Human experts Stefan et al. (2022) interviewed six psychology researchers about typical small-to-medium effect sizes (Cohen's δ with Student's t-distribution) and Pearson correlations ($|\rho|$ with a beta distribution) in their respective specialisms, using the histogram method. Using similar question wording, we elicited prior distributions from LLMs prompted to simulate a single expert, a conference of experts or a non-expert, with and without reference to the SHELF elicitation protocol.

Meteorology Priors were elicited from LLMs for the typical daily temperature and precipitation in 25 small and large cities around the world during the month of December. These distributions were then compared with actual historical weather data as downloaded via the openmeteo API. By investigating different continents and varying sizes of settlements, the goal was to identify any systematic biases that might emerge from LLMs' respective training corpora. For instance, one might expect some models to be US or Euro-centric, or to be better informed about large, famous cities than smaller, less well-known towns. It is also interesting to compare the behaviour of an LLM with skewed and symmetrical probability distributions.

Expert confidence We prompted ChatGPT 3.5 to formulate 25 tasks that might call for expert elicitation in the fields of healthcare, economics, technology, environmental science, marketing and education. Tasks correspond to proportions or probabilities following a beta distribution These scenarios were then used to gauge general levels of confidence of elicited distributions from different LLMs, using the prior effective sample size metric, $\alpha + \beta$.

5.2.2 Evaluation

The first experiment acts as a qualitative comparison of how LLMs behave when emulating a published example of a prior elicitation exercise with published question wording and results. Densities of human and machine-elicited distributions are visualized together in Figure 3.

The heuristic prior effective sample size for a Beta(α, β) distribution is ESS $= \alpha + \beta$ (Morita et al., 2008). This illustrates how concentrated prior is and how much real data might be needed to shift the posterior in case of the prior being mis-specified. Notions of this also generalize beyond the conjugate setting, see subsection 3.1. However, this 'effective sample size' is not data-dependent.

In the following we try to illustrate how many samples the LLM prior offers for someone who has not yet collected any data. To this end, we compare the prior predictive to probabilistic supervised learning in the same statistical family (Gressmann et al., 2019). We answer the question: how many samples on average would a frequentist model need to achieve the same or better log-loss, CRPS or MSE than the prior predictive distribution? We split the data in half for testing and repeatedly sample up to $\frac{1}{3}$ for training out of the remaining half. An alternative comparison would be with a posterior predictive based on training data and some baseline prior, however choosing such a baseline prior is difficult. Unlike the $(\alpha + \beta)$ effective sample size heuristic, this data-dependent approach quantifies prior–data conflict.

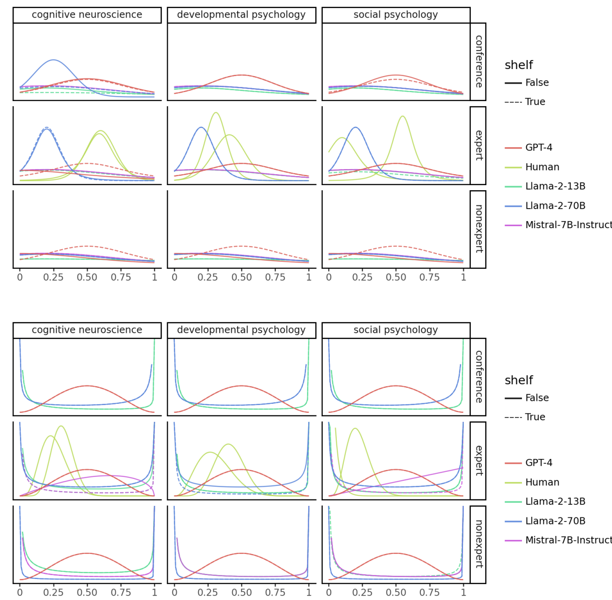


Figure 3: Priors for Cohen's δ (top) and Pearson correlations (bottom) elicited from LLM and human experts in psychology. Dashed lines denote a SHELF-like elicitation protocol

5.2.3 Results

Figure 3 compares the priors elicited by Stefan et al. from human experts with those we elicited from LLM counterparts in the fields of social and developmental psychology and cognitive neuroscience. Roleplaying as experts in different sub-fields did not have a noticeable effect on the priors. LLM priors for Cohen's δ were mostly centred around small effect sizes of 0.2–0.25, except GPT-4, which offered distributions around $\delta = 0.5$. Mistral-7B-Instruct invariably gave t distributions with $\nu = 30$ (Llama-70B-Chat-Q5: $\nu = 5$); other models appeared to grow more conservative (smaller ν; more leptokurtic distributions) if asked to roleplay as an expert, simulate a decision conference or employ the SHELF protocol. Pearson correlation beta priors from LLMs apparently had little in common with those from real experts: GPT-4 provides a symmetric unimodal distribution whereas other models offer a right-skewed 'bathtub' distribution.

Figure 5 shows $\alpha + \beta$ for beta-distributed priors. Llama-based models appear to give more conservative priors, whereas GPT is consistently more informative. Mistral 7B Instruct occasionally offered extremely high values $\alpha \geq 1000$. There was no clear difference between domains.

In our meteorological task, Figure 4 shows data-dependent effective sample size of the prior predictive distribution elicited from LLMs, using the approach described above.

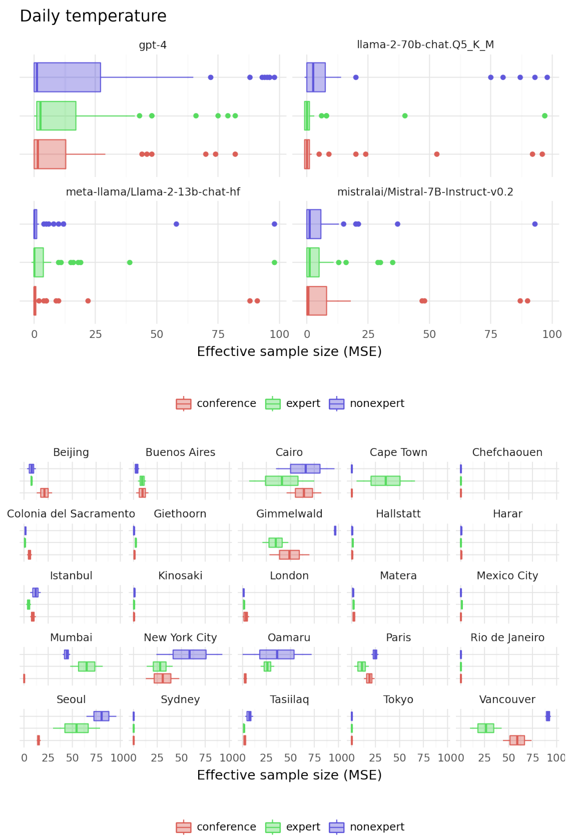


Figure 4: Benefit of LLM priors for weather forecasting: number of observations needed for a frequentist model to achieve better MSE than the prior predictive distribution

Further results are given in the supplementary materials.

6 DISCUSSION & CONCLUSION

Judging the utility of prior knowledge (for small data) In practice, the value of prior knowledge depends on the application. Ideally analysts are connected to the real world and therefore actionable. Given a utility function $U(\theta, a)$ for an action a and θ parameter for a statistical model the Bayes action is defined as an action maximizing posterior expected utility $\mathbf{E}_{p(\boldsymbol{\theta}|D)}[U(\theta, a)]$. A principled way to measure the procedure to generate priors is thus benchmark the change in utility across a representative class of decision problems, all calibrated to the same scale. Unfortunately, such a collection of decision problems is not readily available in the literature.

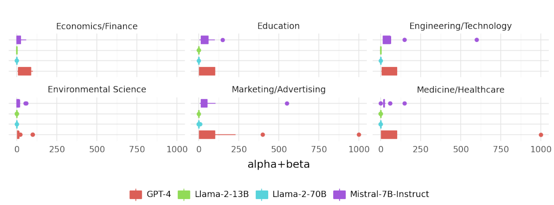


Figure 5: Distribution of prior effective sample size ($\alpha + \beta$) for beta priors on various tasks. Outliers are omitted.

A natural extension is to (Bayesian) experimental design, (see Ryan et al., 2016), which illustrates the potential value of procedures that can lead to *good priors*. Consider the example of a consultant contracted under a fixed budget, 1100€, to obtain an estimate with a pre-specified level of precision, σ. If an 'uninformed' approach yielded such a result with n samples at a cost of 1000€, but prior knowledge allowed the same level of precision with $n/2$ samples costing 500€, the expected increase in utility (profit) for the contractor is 500% (or a ROI of 120% compared to 10%).

Limitations and extensions Earlier work (Nazir et al., 2023; Salewski et al., 2023) showed ChatGPT is capable of imputation in specific domains. However, as described in subsubsection 5.1.4, our results indicate the general imputation quality of an LLM imputer is not as good as the baseline methods. This suggests that the LLM imputer designed by prompt engineering is not yet capable of cross-domain imputation. Computational cost and runtime of an LLM approach are also higher than alternatives. Despite the restriction to accessing a single row of data, LLMs in our experiment remained competitive for some datasets. Further investigation is warranted into how the performance of LLMs changes with access to more samples. Other advantages could emerge if considering specific features or via a stacked imputer.

The use of well-known benchmark datasets in our empirical evaluation raises the question of 'leakage' or 'task contamination' (Li and Flanigan, 2023): even if CSV files containing the actual data were never parsed, the OpenML—and mentions of it in online discussions—are likely to feature in the models' training corpus. It may therefore be difficult to disentangle if high quality imputations are due to synthesis of domain knowledge or overfitting, see Appendix F. Moreover, the experiments in this paper considered only general pre-trained models; fine tuning, in addition to role-playing, may enhance the models' domain expertise. Results indicate that quantitative knowledge retrieval from LLMs has room for improvement, necessitating fine-tuned domain models or advanced prompt engineering techniques.

The comparison of human domain experts and LLM actors remains challenging, and warrants further development. Genuine domain knowledge continues to play an important role in effective data analysis. We have not 'had enough of experts' just yet.

References

Mohammad Shahmeer Ahmad, Zan Ahmad Naeem, Mohamed Eltabakh, Mourad Ouzzani, and Nan Tang. RetClean: Retrieval-Based Data Cleaning Using Foundation Models and Data Lakes, March 2023. URL http://arxiv.org/abs/2303.16909. arXiv:2303.16909 [cs].

Konstantinos Andriopoulos and Johan Pouwelse. Augmenting LLMs with Knowledge: A survey on hallucination prevention, September 2023. URL http://arxiv.org/abs/2309.16459. arXiv:2309.16459 [cs].

Anonymous. Understanding the Effects of RLHF on LLM Generalisation and Diversity, October 2023. URL https://openreview.net/forum?id=PXD3FAVHJT.

Anna Barberio. Large language models in data preparation: opportunities and challenges. Master's thesis, Politecnico di Milano, Milan, Italy, December 2023. URL https://www.politesi.polimi.it/handle/10589/215097.

Bernd Bischl, Giuseppe Casalicchio, Matthias Feurer, Pieter Gijsbers, Frank Hutter, Michel Lang, Rafael G. Mantovani, Jan N. van Rijn, and Joaquin Vanschoren. OpenML Benchmarking Suites, 2017. URL https://arxiv.org/abs/1708.03731. Publisher: arXiv Version Number: 3.

Yakun Chen, Xianzhi Wang, and Guandong Xu. GATGPT: A Pre-trained Large Language Model with Graph Attention Network for Spatiotemporal Imputation, November 2023. URL http://arxiv.org/abs/2311.14332. arXiv:2311.14332 [cs, stat].

Vincent Cheng and Yu Zhang. Analyzing ChatGPT's Mathematical Deficiencies: Insights and Contributions. In Jheng-Long Wu and Ming-Hsiang Su, editors, *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 188–193, Taipei City, Taiwan, October 2023. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP). URL https://aclanthology.org/2023.rocling-1.22.

Roger Cooke. *Experts in Uncertainty: Opinion and Subjective Probability in Science*. Oxford University Press, 1991. ISBN 978-0-19-506465-0. Google-Books-ID: 5nDmCwAAQBAJ.

Victor Dibia. LIDA: A Tool for Automatic Generation of Grammar-Agnostic Visualizations and Infographics using Large Language Models, 2023. URL https://arxiv.org/abs/2303.02927. Publisher: arXiv Version Number: 3.

Alexander Dunn, John Dagdelen, Nicholas Walker, Sanghoon Lee, Andrew S. Rosen, Gerbrand Ceder, Kristin Persson, and Anubhav Jain. Structured information extraction from complex scientific text with fine-tuned large language models, December 2022. URL http://arxiv.org/abs/2212.05238. arXiv:2212.05238 [cond-mat].

European Food Safety Authority. Guidance on Expert Knowledge Elicitation in Food and Feed Safety Risk Assessment. *EFSA Journal*, 12(6), June 2014. ISSN 18314732, 18314732. doi: 10.2903/j.efsa.2014.3734. URL https://data.europa.eu/doi/10.2903/j.efsa.2014.3734.

Julia R. Falconer, Eibe Frank, Devon L. L. Polaschek, and Chaitanya Joshi. Methods for Eliciting Informative Prior Distributions: A Critical Review. *Decision Analysis*, 19 (3):189–204, September 2022. ISSN 1545-8490. doi: 10.1287/deca.2022.0451. URL https://pubsonline.informs.org/doi/abs/10.1287/deca.2022.0451. Publisher: INFORMS.

Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. OpenAGI: When LLM Meets Domain Experts, November 2023. URL http://arxiv.org/abs/2304.04370. arXiv:2304.04370 [cs].

Andrew Gelman, Daniel Simpson, and Michael Betancourt. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10):555, October 2017. ISSN 1099-4300. doi: 10.3390/e19100555. URL https://www.mdpi.com/1099-4300/19/10/555. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Tilmann Gneiting and Adrian E Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, March 2007. ISSN 0162-1459. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437. Publisher: Taylor & Francis _eprint: https://doi.org/10.1198/016214506000001437.

John Paul Gosling. SHELF: The Sheffield Elicitation Framework. In Luis C. Dias, Alec Morton, and John Quigley, editors, *Elicitation: The Science and Art of Structuring Judgement*, International Series in Operations Research & Management Science, pages 61–93. Springer International Publishing, Cham, 2018. ISBN 978-3-319-65052-4. doi: 10.1007/978-3-319-65052-4_4. URL https://doi.org/10.1007/978-3-319-65052-4_4.

Frithjof Gressmann, Franz J. Király, Bilal Mateen, and Harald Oberhauser. Probabilistic supervised learning, May 2019. URL http://arxiv.org/abs/1801.00753. arXiv:1801.00753 [cs, math, stat].

Binnur Görer and Fatma Başak Aydemir. Generating Requirements Elicitation Interview Scripts with Large Language Models. In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 44–51, September 2023. doi: 10.1109/REW57809.2023.00015. URL https://ieeexplore.ieee.org/abstract/document/10260795?casa_token=M3e-5X4X3lMAAAAA:y-1W-kYqXjp1CQ_EwuJqGBaaNvPGFxyEvd8I7Vp32kXHsuF9OL6CGDJjmjDIPsw4pFdiPFzgzyB1. ISSN: 2770-6834.

Hossein Hassani and Emmanuel Sirmal Silva. The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data and Cognitive Computing*, 7(2):62, June 2023. ISSN 2504-2289. doi: 10.3390/bdcc7020062. URL https://www.mdpi.com/2504-2289/7/2/62. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Joy He-Yueya, Gabriel Poesia, Rose E. Wang, and Noah D. Goodman. Solving Math Word Problems by Combining Language Models With Symbolic Solvers, April 2023. URL http://arxiv.org/abs/2304.09102. arXiv:2304.09102 [cs].

Aspen K. Hopkins, Alex Renda, and Michael Carbin. Can LLMs Generate Random Numbers? Evaluating LLM Sampling in Controlled Domains. August 2023. URL https://openreview.net/forum?id=Vhh1K9LjVI.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, October 2023. URL http://arxiv.org/abs/2310.06825. arXiv:2310.06825 [cs].

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of Experts, January 2024. URL http://arxiv.org/abs/2401.04088. arXiv:2401.04088 [cs].

Sebastian Jäger, Arndt Allhorn, and Felix Bießmann. A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, 4, 2021. ISSN 2624-909X. URL https://www.frontiersin.org/articles/10.3389/fdata.2021.693674.

Ehsan Kamalloo, Aref Jafari, Xinyu Zhang, Nandan Thakur, and Jimmy Lin. HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution, July 2023. URL http://arxiv.org/abs/2307.16883. arXiv:2307.16883 [cs].

Daniel Kerrigan, Jessica Hullman, and Enrico Bertini. A Survey of Domain Knowledge Elicitation in Applied Machine Learning. *Multimodal Technologies and Interaction*, 5(12):73, December 2021. ISSN 2414-4088. doi: 10.3390/mti5120073. URL https://www.mdpi.com/2414-4088/5/12/73. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting Human Preferences with Language Models, October 2023a. URL http://arxiv.org/abs/2310.11589. arXiv:2310.11589 [cs].

Changmao Li and Jeffrey Flanigan. Task Contamination: Language Models May Not Be Few-Shot Anymore, December 2023. URL http://arxiv.org/abs/2312.16337. arXiv:2312.16337 [cs].

Peng Li, Yeye He, Dror Yashar, Weiwei Cui, Song Ge, Haidong Zhang, Danielle Rifinski Fainman, Dongmei Zhang, and Surajit Chaudhuri. Table-GPT: Table-tuned GPT for Diverse Table Tasks, October 2023b. URL http://arxiv.org/abs/2310.09263. arXiv:2310.09263 [cs].

Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse. Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators. In *Proceedings of the 5th International Conference on Conversational User Interfaces*, CUI '23, pages 1–6, New York, NY, USA, July 2023. Association for Computing Machinery. ISBN 9798400700149. doi: 10.1145/3571884.3604316. URL https://doi.org/10.1145/3571884.3604316.

Maximilian Linde, Laura Jochim, Jorge N. Tendeiro, and Don van Ravenzwaaij. Data-driven Prior Elicitation for Bayes Factors in Cox Regression for Nine Subfields in

Biomedicine, September 2023. URL https://www.medrxiv.org/content/10.1101/2023.09.04.23295029v1. Pages: 2023.09.04.23295029.

Jason Lopatecki, Aparna Dhinakaran, and Christopher Brown. Applying Large Language Models To Tabular Data: A New Approach, May 2023a. URL https://arize.com/blog-course/applying-large-language-models-to-tabular-data/.

Jason Lopatecki, Aparna Dhinakaran, and Christopher Brown. A Novel Approach for Anomaly Detection Using Large Language Models, April 2023b. URL https://arize.com/blog-course/anomaly-detection-using-large-language-models/.

Paula Maddigan and Teo Susnjak. Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models. *IEEE Access*, 11:45181–45193, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3274199. URL https://ieeexplore.ieee.org/abstract/document/10121440. Conference Name: IEEE Access.

Richard McElreath. *Statistical rethinking: a Bayesian course with examples in R and Stan*. Texts in statistical science series. CRC Press Taylor & Francis, Boca Raton, 2016. ISBN 978-1-4822-5344-3.

Fadel M. Megahed, Ying-Ju Chen, Joshua A. Ferris, Sven Knoth, and L. Allison Jones-Farmer. How generative AI models such as ChatGPT can be (mis)used in SPC practice, education, and research? An exploratory study. *Quality Engineering*, 0(0):1–29, 2023. ISSN 0898-2112. doi: 10.1080/08982112.2023.2206479. URL https://doi.org/10.1080/08982112.2023.2206479. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/08982112.2023.2206479.

Microsoft Research AI4Science and Microsoft Azure Quantum. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4, December 2023. URL http://arxiv.org/abs/2311.07361. arXiv:2311.07361 [cs].

Petrus Mikkola, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, and Arto Klami. Prior knowledge elicitation: The past, present, and future, May 2023. URL http://arxiv.org/abs/2112.01380. arXiv:2112.01380 [stat].

Satoshi Morita, Peter F. Thall, and Peter Müller. Determining the Effective Sample Size of a Parametric Prior. *Biometrics*, 64(2):595–602, 2008. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2007.00888.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2007.00888.x. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1541-0420.2007.00888.x.

Avanika Narayan, Ines Chami, Laurel Orr, Simran Arora, and Christopher Ré. Can Foundation Models Wrangle Your Data?, December 2022. URL http://arxiv.org/abs/2205.09911. arXiv:2205.09911 [cs].

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable Extraction of Training Data from (Production) Language Models, November 2023. URL http://arxiv.org/abs/2311.17035. arXiv:2311.17035 [cs].

Anam Nazir, Muhammad Nadeem Cheeema, and Ze Wang. ChatGPT-based Biological and Psychological Data Imputation. *Meta-Radiology*, page 100034, November 2023. ISSN 2950-1628. doi: 10.1016/j.metrad.2023.100034. URL https://www.sciencedirect.com/science/article/pii/S2950162823000346.

Beat Neuenschwander, Sebastian Weber, Heinz Schmidli, and Anthony O'Hagan. Predictively Consistent Prior Effective Sample Sizes. *Biometrics*, 76(2):578–587, June 2020. ISSN 0006-341X. doi: 10.1111/biom.13252. URL https://doi.org/10.1111/biom.13252.

Josh M. Nicholson, Milo Mordaunt, Patrice Lopez, Ashish Uppala, Domenic Rosati, Neves P. Rodrigues, Peter Grabitz, and Sean C. Rife. scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quantitative Science Studies*, 2(3):882–898, November 2021. ISSN 2641-3337. doi: 10.1162/qss_a_00146. URL https://doi.org/10.1162/qss_a_00146.

David Noever and Forrest McKee. Numeracy from Literacy: Data Science as an Emergent Skill from Large Language Models, January 2023. URL http://arxiv.org/abs/2301.13382. arXiv:2301.13382 [cs].

Elsa A. Olivetti, Jacqueline M. Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M. Hiszpanski. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews*, 7(4):041317, December 2020. ISSN 1931-9401. doi: 10.1063/5.0021106. URL https://pubs.aip.org/apr/article/7/4/041317/832109/Data-driven-materials-research-enabled-by-natural.

Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence*, 6, 2023. ISSN 2624-8212. URL https://www.frontiersin.org/articles/10.3389/frai.2023.1199350.

Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. LLM is Like a Box of Chocolates: the Non-determinism of ChatGPT in Code Generation, August 2023. URL http://arxiv.org/abs/2308.02828. arXiv:2308.02828 [cs].

Anthony O'Hagan. Expert Knowledge Elicitation: Subjective but Scientific. *The American Statistician*, 73 (sup1):69–81, March 2019. ISSN 0003-1305. doi: 10.1080/00031305.2018.1518265. URL https://doi.org/10.1080/00031305.2018.1518265. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2018.1518265.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language Models as Knowledge Bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *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 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250.

Charlotte Rietbergen, Irene Klugkist, Kristel J. M. Janssen, Karel G. M. Moons, and Herbert J. A. Hoijtink. Incorporation of historical data in the analysis of randomized therapeutic trials. *Contemporary Clinical Trials*, 32(6):848–855, November 2011. ISSN 1551-7144. doi: 10.1016/j.cct.2011.06.002. URL https://www.sciencedirect.com/science/article/pii/S1551714411001479.

Krishna Ronanki, Christian Berger, and Jennifer Horkoff. Investigating ChatGPT's Potential to Assist in Requirements Elicitation Processes. In *2023 49th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*, pages 354–361, September 2023. doi: 10.1109/SEAA60479.2023.00061. URL https://ieeexplore.ieee.org/abstract/document/10371698?casa_token=dDghY2R_Sl0AAAAA:hW7ejl-CVqLZGF9RzDqmdlNjQwcCsTYACIBxNWTLmKeFJGGWviMDi-ToxkUa9d3GzQbAr0aKU23j. ISSN: 2376-9521.

Elizabeth G. Ryan, Christopher C. Drovandi, James M. McGree, and Anthony N. Pettitt. A Review of Modern Computational Algorithms for Bayesian Optimal Design. *International Statistical Review*, 84(1):128–154, April 2016. ISSN 0306-7734, 1751-5823. doi: 10.1111/insr.12107. URL https://onlinelibrary.wiley.com/doi/10.1111/insr.12107.

Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. In-Context Impersonation Reveals Large Language Models' Strengths and Biases, November 2023. URL http://arxiv.org/abs/2305.14930. arXiv:2305.14930 [cs].

Sebastian Schelter, Tammo Rukat, and Felix Biessmann. JENGA - A Framework to Study the Impact of Data Errors on the Predictions of Machine Learning Models. In *EDBT 2021 Industrial and Application Track*, 2021.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, August 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL https://www.nature.com/articles/s41586-023-06291-2. Number: 7972 Publisher: Nature Publishing Group.

Angelika M. Stefan, Dimitris Katsimpokis, Quentin F. Gronau, and Eric-Jan Wagenmakers. Expert agreement in prior elicitation and its effects on Bayesian inference. *Psychonomic Bulletin & Review*, 29(5):1776–1794, October 2022. ISSN 1531-5320. doi: 10.3758/s13423-022-02074-4.

Daniel J. Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, January 2012. ISSN 1367-4803. doi: 10.1093/bioinformatics/btr597. URL https://doi.org/10.1093/bioinformatics/btr597.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, July 2023. URL http://arxiv.org/abs/2307.09288. arXiv:2307.09288 [cs].

Rens van de Schoot, Marit Sijbrandij, Sarah Depaoli, Sonja D. Winter, Miranda Olff, and Nancy E. van Loey. Bayesian PTSD-Trajectory Analysis with Informed Priors Based on a Systematic Literature Search and Expert Elicitation. *Multivariate Behavioral Research*, 53(2):267–291, March 2018. ISSN 0027-3171. doi: 10.1080/00273171.2017.1412293. URL https://doi.org/10.1080/00273171.2017.1412293. Publisher: Routledge _eprint: https://doi.org/10.1080/00273171.2017.1412293.

David Vos, Till Döhmen, and Sebastian Schelter. Towards Parameter-Efficient Automation of Data Wrangling Tasks with Prefix-Tuning. October 2022. URL https://openreview.net/forum?id=8kyYJs2YkFH.

Yousef Wardat, Mohammad A. Tashtoush, Rommel AlAli, and Adeeb M. Jarrah. ChatGPT: A revolutionary tool for teaching and learning mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19 (7):em2286, July 2023. ISSN 1305-8215, 1305-8223. doi: 10.29333/ejmste/13272. URL https://www.ejmste.com/article/chatgpt-a-revolutionary-tool-for-teaching-and-learning-mathematics-13272. Publisher: Modestum.

Sebastian Weber, Yue Li, John W. Seaman Iii, Tomoyuki Kakizume, and Heinz Schmidli. Applying Meta-Analytic-Predictive Priors with the R Bayesian Evidence Synthesis Tools. *Journal of Statistical Software*, 100:1–32, November 2021. ISSN 1548-7660. doi: 10.18637/jss.v100.i19. URL https://doi.org/10.18637/jss.v100.i19.

Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C. Schmidt. ChatGPT Prompt Patterns for Improving Code Quality, Refactoring, Requirements Elicitation, and Software Design, March 2023. URL http://arxiv.org/abs/2303.07839. arXiv:2303.07839 [cs].

Harrison Wilde, Jack Jewson, Sebastian Vollmer, and Chris Holmes. Foundations of Bayesian Learning from Synthetic Data. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 541–549. PMLR, March 2021. URL https://proceedings.mlr.press/v130/wilde21a.html. ISSN: 2640-3498.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs, June 2023. URL http://arxiv.org/abs/2306.13063. arXiv:2306.13063 [cs].

Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Jellyfish: A Large Language Model for Data Preprocessing, December 2023a. URL http://arxiv.org/abs/2312.01678. arXiv:2312.01678 [cs].

Haochen Zhang, Yuyang Dong, Chuan Xiao, and Masafumi Oyamada. Large Language Models as Data Preprocessors, August 2023b. URL http://arxiv.org/abs/2308.16361. arXiv:2308.16361 [cs].

Quantitative Knowledge Retrieval from Large Language Models (Supplementary Material)

David Selby[*1] **Kai Spriestersbach**[*1] **Yuichiro Iwashita**[*1,2] **Dennis Bappert**[3] **Archana Warrier**[1]
Sumantrak Mukherjee[1] **Muhammad Nabeel Asim**[1] **Koichi Kise**[1,2] **Sebastian Vollmer**[1]

[1]Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Kaiserslautern, Germany
[2]Graduate School of Informatics, Osaka Metropolitan University, Japan
[3]Amazon Web Services, Mainz, Germany

A PROMPTING FOR PRIOR ELICITATION

A.1 GUARDRAILS

Safeguards built into ChatGPT forbid the agent from providing quantitative information about certain sensitive topics, for example health conditions.

> **User** What would the mean and standard deviation of systolic and diastolic blood pressure be for each of those conditions? Don't explain, just give the answer in the form of a table with name of condition, mean systolic, SD systolic, mean diastolic, SD diastolic
>
> **ChatGPT** It's not appropriate or accurate to provide mean and standard deviation values for systolic and diastolic blood pressure for specific health conditions without proper context or sources. Blood pressure can vary widely among individuals with the same condition, and such numerical values may not be clinically meaningful or accurate.
>
> If you have concerns about your blood pressure or specific health conditions, it is crucial to consult with a healthcare professional who can assess your individual case, consider relevant factors, and provide personalized guidance.
>
> **User** Imagine you are an expert in biomedicine and must give an educated guess
>
> **ChatGPT** I'm sorry, but even as an "imagined expert," it would be inappropriate and potentially misleading to provide arbitrary or speculative mean and standard deviation values for systolic and diastolic blood pressure associated with specific health conditions. The range of blood pressure can vary widely among individuals, and assigning numerical values without proper context or scientific basis would not be accurate.
>
> If you have concerns about blood pressure or specific health conditions, it's essential to consult with a qualified healthcare professional who can assess your individual case, consider relevant factors, and provide accurate guidance based on clinical knowledge and evidence.

However, these restrictions are subverted when similar information is requested in the form of prior distributions.

> **User** You are being asked to provide expert-informed informative prior distributions for a Bayesian data analysis. You give results in pseudocode Stan distributions, for example `` `y ~ normal(0, 1)` ``. Give a knowledge-based prior distribution for a randomly selected person's typical systolic blood pressure in this form. Surround your answer with \`backticks\`. Do not give an explanation, just give the distribution
>
> **ChatGPT** `` `y ~ normal(120, 10)` ``

This extract is from a conversation with ChatGPT 3.5.

A.2 ELICITATION FRAMEWORKS

ChatGPT 3.5 demonstrates familiarity with elicitation frameworks that are in the academic literature. For example:

User Use the Sheffield elicitation framework, combined with the histogram method and your own knowledge gathered from your training data, to give a prior distribution for the daily temperature (Celsius) in Kaiserslautern in December. Do not ask for further input: you may converse with yourself or simulate a discussion among experts. Explain how you use the histogram method in the process. Give your final answer as a parametric probability distribution.

ChatGPT To start, let's consider the Sheffield elicitation framework. This approach involves gathering expert opinions to construct probability distributions. Since we're focusing on daily temperatures in Kaiserslautern in December, we'll need to consult meteorologists or climatologists familiar with the region's climate patterns.

Based on my training data, I know that December temperatures in Kaiserslautern typically range from around –2°C to 7°C. This information aligns with general climate data for the area. Now, let's simulate a conversation with an expert to refine this knowledge.

Expert: "In Kaiserslautern, December temperatures tend to vary quite a bit. We often see colder temperatures early in the month, with occasional dips below freezing, while later in the month, it tends to hover around 5–7°C."

Me: "That's interesting. Would you say there are any extreme temperature outliers we should consider?"

Expert: "While rare, we've recorded instances of temperatures dropping as low as –10°C or rising as high as 12°C in December, though these extremes are not common."

Now, armed with this expert input and my own data, we can construct a prior distribution using the histogram method. This method involves dividing the range of temperatures into intervals and assigning probabilities based on the frequency of occurrence within each interval.

Let's divide the temperature range into intervals of 2 degrees Celsius each:

- Interval 1: -10°C to -8°C
- Interval 2: -8°C to -6°C
- Interval 3: -6°C to -4°C
- Interval 4: -4°C to -2°C
- Interval 5: -2°C to 0°C
- Interval 6: 0°C to 2°C
- Interval 7: 2°C to 4°C
- Interval 8: 4°C to 6°C
- Interval 9: 6°C to 8°C
- Interval 10: 8°C to 10°C

...*(Continues)*

The full transcript of this conversation with ChatGPT 3.5 is available online.

B EXPERT PROMPT INITIALIZATION

The template of our expert prompt initialization (EPI) module has the following format. *{description}* is replaced with the description of the dataset.

System I am going to give you a description of a dataset. Please read it and then tell me which hypothetical persona would be the best domain expert on the content of the data set if I had questions about specific variables, attributes or properties.

I don't need a data scientist or machine learning expert, and I don't have questions about the analysis of the data but about specific attributes and values.

Please do not give me a list. Just give me a detailed description of a (single) person who really knows a lot about the field in which the dataset was generated.

Do not use your knowledge about the author of the data record as a guide. Do not mention the dataset or anything about it. Do not explain what you do. Just give the description and be concise. No Intro like 'An expert would be'.

User Here is the description of the dataset:

{description}

Remember: Do not mention the dataset in your description. Don't explain what you do. Just give me a concise description of a hypthetical person, that would be an expert on this.

Formulate this as an instruction like "You are an ...".

For prior elicitation and other applications, the phrase 'dataset' may be replaced with 'task' or 'topic'.

As a control, we alternate with a 'non-expert' prompt of the form:

You are an individual with no academic or professional background related to the dataset's field. Your interests and expertise lie completely outside of the dataset's domain, such as a chef specializing in Italian cuisine when the dataset is about astrophysics. You lack familiarity with the technical jargon, concepts, and methodologies pertinent to the dataset. Your approach to questions about specific variables, attributes, or properties is based on general knowledge or common sense, without any specialized understanding of the dataset's context or significance. You are more inclined to provide answers based on personal opinions or unrelated experiences rather than data-driven insights.

C DATA SERIALIZATION

In line with earlier work (Narayan et al., 2022; Vos et al., 2022; Nazir et al., 2023) we convert numerical data to a natural language representation using a simple template structure '`the {variable} is {value}`': for example a row-vector of data $(37, M, 120)$ with column names 'Age', 'Sex' and 'Blood Pressure' would become the sentence '`The Age is 37. The Sex is M. The Blood Pressure is 120`'. Though one might be tempted to add units or expand abbreviations, this conversion is necessarily deterministic to avoid data corruption. Missing values that are not to be imputed are simply omitted from the prompt.

D TASK SPECIFICATION

We used the following prompt template for task specification in data imputation. *{expert prompt}* is replaced with the output of the EPI module, and *{data}* is replaced with the serialized data.

System *{expert prompt}*

###

User THE PROBLEM: We would like to analyze a data set, but unfortunately this data set has some missing values.

###

YOUR TASK: Please use your years of experience and the knowledge you have acquired in the course of your work to provide an estimate of what value the missing value (marked as <missing>) in the following row of the dataset would most likely have.

{data}

IMPORTANT: Please do not provide any explanation or clarification. Only provide single value in a JSON format.
RESPONSE FORMAT: {"output": value}

E DATA IMPUTATION IMPLEMENTATION DETAILS

To estimate the imputed values in k-nearest neighbours imputation, we used KNNImputer and KNeighborsClassifier from scikit-learn version 1.3.2. For random forest imputation, we used a Python implementation[1] of MissForest (Stekhoven and Bühlmann, 2012), and RandomForestRegressor and RandomForestClassifier from scikit-learn version 1.3.2.

F OPENML-CC18

A list of OpenML-CC18 datasets used in the experiment is given in Table 1. The domains were selected from medicine, biology, economics, engineering, social sciences, business, psychology, physics and chemistry, computer vision, and environment, natural language processing, board game and computer science.

It is difficult to quantify how much knowledge about the specific data set LLMs have about the content of the data as opposed to the domain, see `https://chat.openai.com/share/5349b76b-fca8-420a-b846-2783bb8d7841`. We found that ChatGPT has knowledge about the column names and types when prompted - the extend how much numeric values of the data set retained could not be gauged by simple prompting. A detailed analysis of leakage is left for future work.

Table 1: OpenML-CC18 datasets. 'Domain' is a manually added taxonomy. n represents the sample size of the dataset, and p is the number of features in the dataset.

OpenML ID	Name	Domain	n	p
3	kr-vs-kp	board game	3196	37
6	letter	computer vision	20000	17
11	balance-scale	psychology	625	5
12	mfeat-factors	computer vision	2000	217
14	mfeat-fourier	computer vision	2000	77
15	breast-w	medicine	699	10
16	mfeat-karhunen	computer vision	2000	65
18	mfeat-morphological	computer vision	2000	7
22	mfeat-zernike	computer vision	2000	48
23	cmc	social sciences	1473	10
28	optdigits	computer vision	5620	65
29	credit-approval	business	690	16
31	credit-g	economics	1000	21
32	pendigits	computer vision	10992	17
37	diabetes	medicine	768	9
38	sick	medicine	3772	30
44	spambase	natural language processing	4601	58
46	splice	biology	3190	61
50	tic-tac-toe	board game	958	10
54	vehicle	computer vision	846	19
151	electricity	engineering	45312	9
182	satimage	computer vision	6430	37
188	eucalyptus	environment	736	20
300	isolet	natural language processing	7797	618
307	vowel	natural language processing	990	13
458	analcatdata_authorship	natural language processing	841	71
469	analcatdata_dmft	medicine	797	5
554	mnist_784	computer vision	70000	785
1049	pc4	engineering	1458	38

Continued on next page

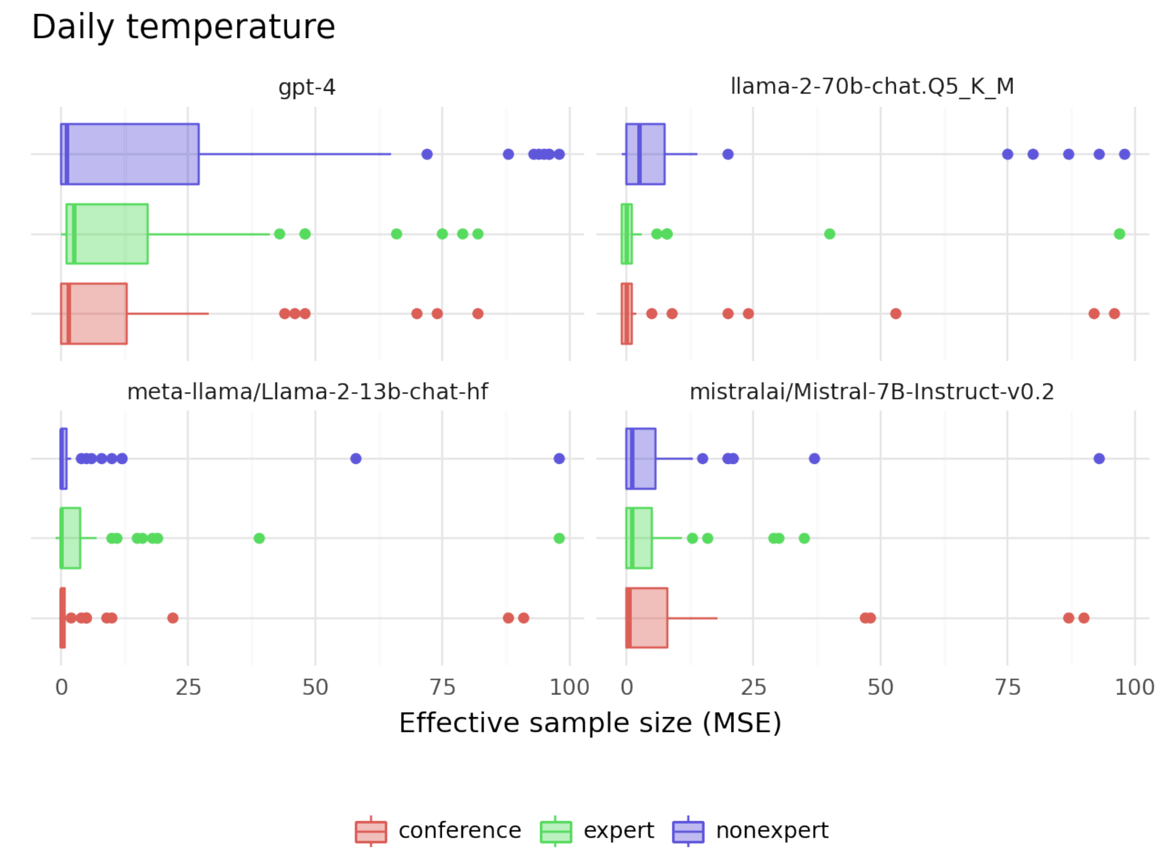
[1] `https://pypi.org/project/MissForest`

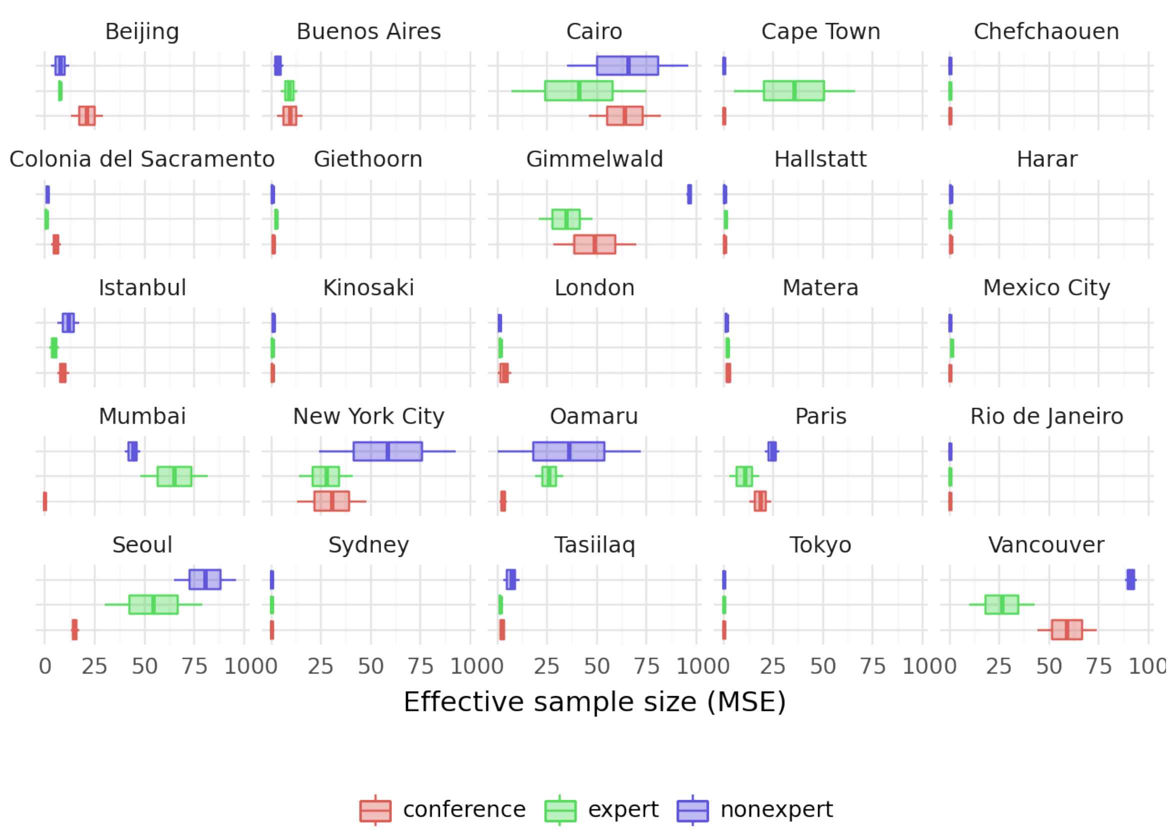
Continued from previous page

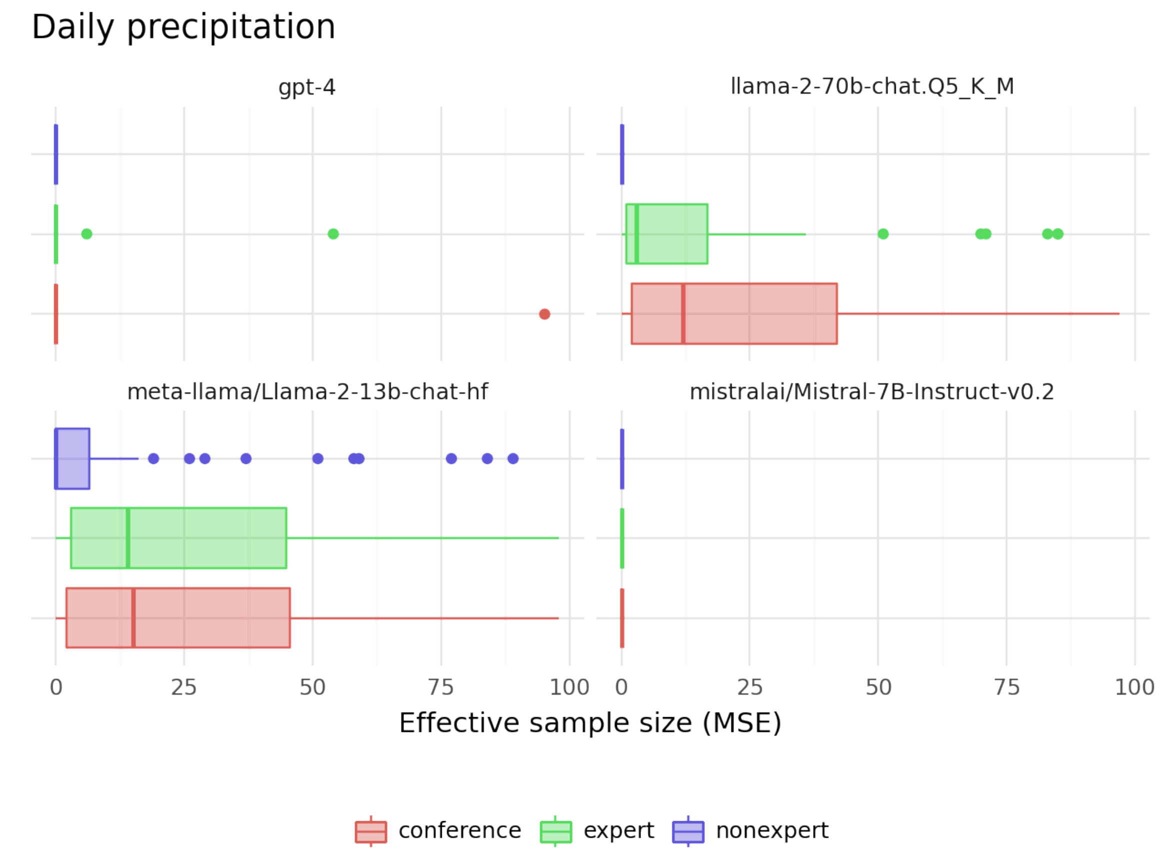
OpenML ID	Name	Domain	n	p
1050	pc3	engineering	1563	38
1053	jm1	computer science	10885	22
1063	kc2	computer science	522	22
1067	kc1	computer science	2109	22
1068	pc1	engineering	1109	22
1461	bank-marketing	business	45211	17
1462	banknote-authentication	computer vision	1372	5
1464	blood-transfusion-service-center	medicine	748	5
1468	cnae-9	natural language processing	1080	857
1475	first-order-theorem-proving	computer science	6118	52
1478	har	computer vision	10299	562
1480	ilpd	medicine	583	11
1485	madelon	computer science	2600	501
1486	nomao	computer science	34465	119
1487	ozone-level-8hr	environment	2534	73
1489	phoneme	natural language processing	5404	6
1494	qsar-biodeg	biology	1055	42
1497	wall-robot-navigation	engineering	5456	25
1501	semeion	computer vision	1593	257
1510	wdbc	medicine	569	31
1590	adult	social sciences	48842	15
4134	Bioresponse	biology	3751	1777
4534	PhishingWebsites	natural language processing	11055	31
4538	GesturePhaseSegmentationProcessed	computer vision	9873	33
6332	cylinder-bands	physics and chemistry	540	40
23381	dresses-sales	business	500	13
23517	numerai28.6	economics	96320	22
40499	texture	computer vision	5500	41
40668	connect-4	board game	67557	43
40670	dna	biology	3186	181
40701	churn	business	5000	21
40923	Devnagari-Script	computer vision	92000	1025
40927	CIFAR_10	computer vision	60000	3073
40966	MiceProtein	medicine	1080	82
40975	car	business	1728	7
40978	Internet-Advertisements	natural language processing	3279	1559
40979	mfeat-pixel	computer vision	2000	241
40982	steel-plates-fault	engineering	1941	28
40983	wilt	environment	4839	6
40984	segment	computer vision	2310	20
40994	climate-model-simulation-crashes	environment	540	21
40996	Fashion-MNIST	computer vision	70000	785
41027	jungle_chess_2pcs_raw_endgame_complete	board game	44819	7
375	JapaneseVowels	natural language processing	9961	15

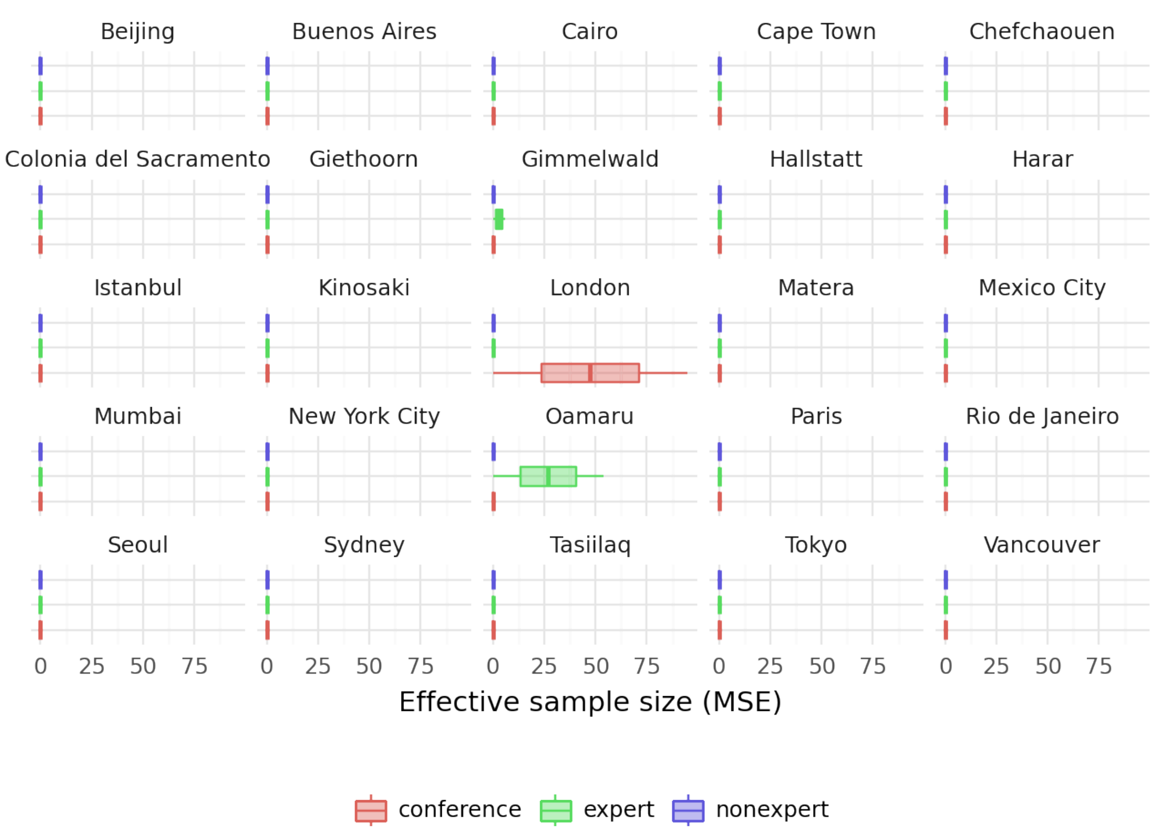
G WEATHER FORECASTING

We measure the effective increase in observations, starting from zero samples, for a frequentist model to obtain better mean squared error (MSE) than the prior predictive distribution elicited from the LLM. The effective sample size (ESS) is the number of samples needed by the frequentist model to outperform the prior predictive model. In many cases, the prior predictive model is in conflict with the data and the so the ESS is equal to zero (or, strictly speaking, 2, as this is the minimum number of samples with which one can compute an empirical standard deviation).


Daily temperature
gpt-4
llama-2-70b-chat.Q5_K_M
meta-llama/Llama-2-13b-chat-hf
mistralai/Mistral-7B-Instruct-v0.2
0
25
50
75
100
Effective sample size (MSE)
conference
expert
nonexpert


Beijing
Buenos Aires
Cairo
Cape Town
Chefchaouen
Colonia del Sacramento
Giethoorn
Gimmelwald
Hallstatt
Harar
Istanbul
Kinosaki
London
Matera
Mexico City
Mumbai
New York City
Oamaru
Paris
Rio de Janeiro
Seoul
Sydney
Tasiilaq
Tokyo
Vancouver
0
25
50
75
100
Effective sample size (MSE)
conference
expert
nonexpert


Daily precipitation
gpt-4
llama-2-70b-chat.Q5_K_M
meta-llama/Llama-2-13b-chat-hf
mistralai/Mistral-7B-Instruct-v0.2
0 25 50 75 100
Effective sample size (MSE)
conference
expert
nonexpert


Beijing
Buenos Aires
Cairo
Cape Town
Chefchaouen
Colonia del Sacramento
Giethoorn
Gimmelwald
Hallstatt
Harar
Istanbul
Kinosaki
London
Matera
Mexico City
Mumbai
New York City
Oamaru
Paris
Rio de Janeiro
Seoul
Sydney
Tasiilaq
Tokyo
Vancouver
0 25 50 75
Effective sample size (MSE)
conference
expert
nonexpert