Human Language Models (LLMs) have revolutionized natural language processing and demonstrated impressive capabilities in various tasks. Unfortunately, they are prone to hallucinations, where the model exposes incorrect or false information in its responses, which renders diligent evaluation approaches mandatory. While LLM performance in specific knowledge fields is often evaluated based on question and answer (Q&A) datasets, such evaluations usually report only a single accuracy number for the entire field, a procedure which is problematic with respect to transparency and model improvement. A stratified evaluation could instead reveal subfields, where hallucinations are more likely to occur and thus help to better assess LLMs’ risks and guide their further development. To support such stratified evaluations, we propose LLMMaps as a novel visualization technique that enables users to evaluate LLMs’ performance with respect to Q&A datasets. LLMMaps provide detailed insights into LLMs’ knowledge capabilities in different subfields, by transforming Q&A datasets as well as LLM responses into our internal knowledge structure. An extension for comparative visualization furthermore, allows for the detailed comparison of multiple LLMs. To assess LLMMaps we use them to conduct a comparative analysis of several state-of-the-art LLMs, such as BLOOM, GPT-2, GPT-3, ChatGPT and LLaMa-13B, as well as two qualitative user evaluations. All necessary source code and data for generating LLMMaps to be used in scientific publications and elsewhere will be available on GitHub: https://github.com/******

Index Terms—Large language models, explainable artificial intelligence

1 INTRODUCTION

- Patrik Puchert is with Ulm University. E-mail: patrik.puchert@uni-ulm.de
- Poonam Poonam is with Ulm University. E-mail: poonam.poonam@uni-ulm.de
- Christian van Onzenoodt is with Ulm University. E-mail: christian.van-onzenoodt@uni-ulm.de
- Timo Ropinski is with Ulm University. E-mail: timo.ropinski@uni-ulm.de

The advent of Large Language Models (LLMs) has been a significant breakthrough in the field of natural language processing (NLP) and has shown impressive capabilities in various tasks, such as language modeling, sentiment analysis, question-answering, and machine translation. LLMs are trained on massive text datasets that contain vast amounts of language data, which makes them perform well on tasks that require the understanding of the context and the ability to generate coherent and meaningful responses. Recently, researchers were able to show that state-of-the-art LLMs pass law exams [9] or even the United States
Medical Licensing Examination (USMLE) [33]. However, despite their impressive performance, LLMs also face several challenges and imply dangers [2], including the problem of hallucination [18], which is a common phenomenon where the model includes incorrect or false information in its responses [54, 55], while still providing an eloquent answer [27]. Due to this problem, the recently released GPT-4 is not fully reliable, which is considered the most important limitation by its authors [35]. Unfortunately, hallucinations are especially problematic in the most promising application domains, such as healthcare or legal settings, where incorrect information can have serious consequences. Therefore, it is mandatory to support a rigorous evaluation of LLMs, in order to enable risk transparency and to steer future development.

Evaluating the performance of LLMs on large question and answer (Q&A) datasets is a common practice in NLP and machine learning. In this evaluation method, the LLM is presented with a set of questions and is expected to generate correct responses, which are known via annotations but not revealed to the LLM via labels. While initially these question-based evaluations were tailored to investigate an LLM’s language capabilities [13], recently we can see a clear trend towards the investigation of knowledge capabilities [5, 6, 9, 21, 33, 35]. Based on the answered questions, the performance of the LLM is measured by computing the accuracy of its responses. While this method provides a quick and easy way to evaluate the accuracy of LLMs, it has limitations and does not provide a comprehensive understanding of the model’s capabilities and limitations. Rather a more stratified analysis of LLMs is necessary to identify knowledge fields where hallucinations are more likely to occur, and thus assess their performance and risks in different subfields. Such a stratified evaluation not only contributes to the transparency of LLMs, but also enables developers to further improve their performance. For instance, by identifying the subfields where hallucinations are more likely to occur, LLM developers could collect additional training data or steer human reinforcement learners towards these subfields in order to inspect LLM answers more critically or simply more often. Moreover, a stratified evaluation could also be used to compare the performance of several LLMs and thus simply allow for a comparative evaluation, or even to gain insights to orchestrate ensemble LLMs based on a specific prompt.

To support a stratified analysis of LLMs, we propose LLMMaps, a novel visualization technique that enables LLM users and developers alike to evaluate the performance of one or multiple LLMs with respect to one or multiple Q&A datasets (see Figure 1). Thus, LLMMaps rather than showing an aggregated accuracy, enable novel and detailed insights regarding the knowledge capabilities of LLMs in different knowledge subfields. Towards this end we make the following contributions within this paper:

- We present LLMMaps as a novel and carefully designed visualization technique to support a stratified evaluation of LLM knowledge capabilities on Q&A datasets.
- We suggest a novel knowledge stratification strategy for Q&A datasets, which allows for the derivation of a knowledge hierarchy as exploited by LLMMaps.
- We introduce an extension to LLMMaps, that enables comparative LLM visualization, and thus enables us to perform a first detailed comparison of SOTA LLMs.
- We conduct several qualitative evaluations highlighting the benefits and downsides of LLMMaps.
- We will release all source code and data upon acceptance, to enable both, NLP researchers to use, and VIS researchers to further improve LLMMaps.

2 Related Work

LLM evaluation. Evaluating the performance of LLMs is a crucial step in developing and improving their capabilities. To consider the state of the art on LLM evaluation, we have reviewed recent publications on LLMs, whereby we had a special focus on the evaluation of the models’ knowledge capabilities. While these publications consist of several works, which propose and evaluate a single LLM, there have also been frameworks proposed for the comparative evaluation of multiple LLMs. One common evaluation practice in NLP is to test LLMs on large Q&A datasets. In this method, the model is presented with a set of questions and is expected to generate correct responses without access to the annotated labels. Typically, obtained LLM responses are evaluated using metrics that aim to measure their performance in language-related tasks [24, 25]. However, for knowledge assessment, accuracy and F1-score are more relevant. To quantify results on Q&A datasets, accuracy is often simply considered as exact-match accuracy, i.e., the model generated response must exactly match one of the optional answers. This can be further expanded to a quasi-exact match score to focus on the knowledge contained in the answer rather than the exact formulation. For other tasks with an even broader output space, the accuracy becomes more difficult to quantify. For question answering tasks, the F1-score is commonly used to measure the overlap of words in desired and given responses. Furthermore, MRR and NDCG scores [17] can be applied to measure information retrieval in the LLMs generated output. For all these metrics, the accuracy of the model’s responses is then calculated as the ratio of the model’s correct answers to the total number of questions. While these metrics provide a quick and easy way to evaluate the accuracy of LLMs, they do not help to obtain a comprehensive understanding of their capabilities and limitations.

LLM visualization. Burkhard defines knowledge visualization as the use of visual representations to facilitate knowledge transfer between two or more persons [7]. Thus, it must not only visualize the available information, but boil down to what is important. While general approaches like knowledge graphs [16] exploit every relation in the data to generate a visual representation, this representation is overly extensive to be of help to non-experts. To this end, more stratified approaches can help to convey the desired knowledge in an easily understandable and efficient way. Popular visualizations of such stratified, and thus hierarchical, data can take the form of dendrograms, radial tree maps or mindmaps [41]. In the context of LLMs this has been primarily applied to the underlying data, by visualizing sentence structures [52]. Even
through this can be of great help for developers and prompt designers, it does not help in conveying knowledge capabilities.

Other applications of visualization on LLMs focus on aspects of explainable AI to make the decision making process of AI models understandable for humans. Usually, these approaches are addressed towards experts, as the understandability of what leads the network to a certain decision, requires also an understanding of its functioning. Some works focus on how the network actually achieves a certain result by exploring the textual and linguistic structure of the generated output [23], by visualizing the inner workings of the network utilizing saliency and attention maps [53], or by highlighting the contextualization of word embeddings [43, 44].

Although, these approaches are of great help to developers, they do not leverage the understandability of an LLM’s accuracy in the targeted knowledge fields.

3 DESIGN CONSIDERATIONS

To identify visualization design requirements for evaluating LLMs, it is necessary to first closely look into the current state of the art in LLM Evaluation, as we will do in Subsection 3.1. However, since the knowledge capabilities of LLMs are progressing fast, and to some extent towards those of a human, we also take inspiration from the current state of the art in knowledge assessment as done in education, which we will discuss in Section 3.2.

3.1 LLM Evaluation

In their report on risks of language models, Weidinger et al. outline six specific risk classes [55]. From these six classes, one is referred to as misinformation harms, which results from false or misleading information provided by language models. Since it could be shown by Martindale and Carpuat, in the context of machine translation, that fluency affects the perceived adequacy of generated text [28], it can probably be assumed that the risk of misinformation is reinforced by the eloquence of modern LLMs. Therefore, motivated by Tamkin et al.’s observation, that the assessment of LLM risk requires methodology which is beyond the methods traditionally used for language models [46], we have developed LLMMaps to target the risk of misinformation harms. This is in line with the fact, that the authors of GPT-4 underline the importance of its knowledge provision capabilities, by highlighting the fact that it passes a bar exam with a score around the top 10% of test takers, in the abstract of the 99 page GPT-4 white paper, while at the same time quoting misinformation as its major challenge [35]. Based on these observations together with the shift of LLM evaluations from language understanding towards knowledge capabilities (see Section 2), we see a clear need to support knowledge-based evaluations.

While the above observations have mainly been made by focusing on the evaluation of a single LLM, also the requirements of more centralized benchmark initiatives for the comparison of several LLMs must be met by LLMMaps. For such a purpose, large-scale initiatives integrating multiple Q&A datasets and providing leader-boards, exist. The most extensive framework in this area is Stanford’s Holistic Evaluation of Language Models (HELM), which is envisioned as a living benchmark, currently reporting 7 metrics for 30 language models [24]. More recently, OpenAI has released their Evals framework, which has the goal to measure the quality of completions provided by an OpenAI model as well as comparing performance across different datasets and models. Based on these publications and frameworks, we see two main usage scenarios for LLMMaps. First, the evaluation of a single LLM on one or multiple LLMs. And second, the comparative evaluation of multiple LLMs on one or multiple Q&A datasets. In the single LLM usage scenario, LLMMaps shall uncover an LLM’s limitations, which can give model developers useful clues on where improvements are necessary, which could in the simplest case just be selectively increasing the amount of training data. In the multiple LLM usage scenario, we see LLMMaps role more targeted towards AI users and policy makers, which require transparency to decide for or against a single model.

Finally, as the discussed LLM evaluations inspiring our visualization design can mainly be found in scientific publications or on web pages, LLMMaps should be usable in these contexts, and ideally beyond.

3.2 Knowledge Assessment

Knowledge assessment is a process for determining the nature and extent of human learning and development [32]. As modern LLMs, such as GPT-4 are quoted as exhibiting human-level performance on various professional and academic benchmarks [35], grounded knowledge assessment becomes more and more important in the context of LLMs.

When addressing knowledge assessment on a high level, we can differentiate between formative and summative assessment. Formative assessment is an ongoing, low-stakes evaluation of a student’s understanding throughout the learning process. It helps teachers to identify areas where students are struggling and adjust their teaching strategies accordingly. Summative assessment, on the other hand, is a final evaluation that measures the overall understanding of a subject and is used to assign grades or determine mastery of a topic. When considering these two assessment strategies in the context of LLMs, or even more broadly machine learning, formative assessment would most likely be associated with the actual learning process, while summative assessment would be the evaluation on the test set. Accordingly, we consider LLMMaps as a summative evaluation, since they are applied after model training.

When investigating summative evaluation, it is widely considered to be effective and useful when knowledge assessment is integrated across multiple settings such as targeting it towards individual courses of a curriculum [45]. For the individual courses again, it is often helpful to assess knowledge for different subfields, such as for instance the chapters of a lecture course. Humans might be doing well in some subfields, while performing subpar in others. We see this as a further motivation for supporting stratification in the context of LLM evaluation.

For effective assessment in LLMs, we should further focus on important learning goals and take into account what is considered evidence for human learning. From the main groups of evidence for human learning, such as written work, performance, and presentation [32], we see performance as measured based on correct test responses, as the most obvious direct evidence for LLM learning. While written work could certainly also be considered, it is more difficult to rate, as it mixes text generation and knowledge capabilities [27]. Therefore, we and many other researchers [5, 6, 21, 33, 35], focus on LLM knowledge assessment based on Q&A datasets.

According to Tamkin et al., one central question regarding LLMs, is whether they exhibit understanding [46]. Therefore, to assess the quality of LLMs or to find out the errors in responses or hallucinations within the responses, we should consider the dimensions of learning, which range from conceptual understanding, over creativity to problem solving [45]. As, these dimensions are also highly relevant for human learners, in the past, several taxonomies for classifying learning content, have been proposed. Bloom’s Taxonomy is perhaps the most widely known and used classification system for knowledge questions [1]. It is a hierarchical framework that classifies educational learning objectives into six different levels of complexity and specificity: Remembering, Understanding, Applying, Analyzing, Evaluating, and Creating. It has been used extensively in education, from K-12 classrooms to higher education and professional development. While several other such taxonomies exist (e.g., Structure of Observed Learning Outcomes (SOLO) Taxonomy, Webb’s Depth of Knowledge, Fink’s Taxonomy of Significant Learning, and Marzano’s Taxonomy), due to its widespread use, we have decided to choose Bloom’s Taxonomy to classify the type of knowledge gain for each question, contained in the processed Q&A datasets. This can be done by analyzing the question and identifying the specific cognitive processes involved, based on which the question can be categorized. For example, a question such as ‘What is the definition of photosynthesis?’ would be classified as a remembering question. A question such as ‘What are the causes of climate change?’ would be classified as an analyzing question. A question such as ‘How would you design an experiment to test the effects of different fertilizers on plant growth?’ would be classified as a creating question. Thus, we consider
the dimensions of learning by classifying all questions according to Bloom, and stratifying our evaluation also based on Bloom’s levels.

Besides learning dimension, we have identified several other potentially useful question qualifiers in our literature study (see Supplement). While many of these qualifiers are hard to assess, we leave them open for future work. Nevertheless, we focus on the difficulty level, which refers to how challenging it is for the learners to answer correctly. We believe that self-assessed difficulty levels could be an ingredient to uncover misinformation, one of the major risks in the context of LLMs [54].

4 Knowledge Processing

In this section, we will describe how we have generated the knowledge hierarchies used by LLMMaps (see Section 4.1), and we will propose a simple hallucination score (see Section 4.2), which we have developed to showcase how such a score could be communicated through LLMMaps. Readers mainly interested in the visualization aspects of LLMMaps might skip this section for now, and refer to it later when diving deeper into the visualized data.

4.1 Knowledge Field Stratification

Since knowledge field stratification is the basis for LLMMaps, we outline such a strategy within this section. Most Q&A datasets are either not or only very coarsely stratified. Some contain 1 out of only 5-10 different topic labels associated with each question [15,36]. In other cases, such labels are so fine-grained [19] that using them to define subfields would clutter the visualization. Sometimes a subfield related classification labels are available, e.g., a measured difficulty for the publicly available US Multistate Bar Examination questions, which do however not stratify the knowledge fields. While all question classification labels we found in the processed Q&A datasets make knowledge stratification difficult, the observed dataset differences further complicate matters, and we see a need for a robust and generalizable knowledge hierarchy derivation for Q&A datasets.

To address this challenge, we propose a top-down knowledge stratification strategy, which we developed to associate individual questions with the leaf nodes of the obtained knowledge hierarchy. Our approach starts by identifying the overarching subfields of the knowledge field represented by a Q&A dataset. Afterwards each of these overarching subfields, is further divide into finer-grained subfields, which we recursively divide again until the target depth of the resulting knowledge tree is reached. Finally, each question is associated with the leaf node best reflecting its subfield. In our experiments we found a number of 5-10 top-level subfields depending on Q&A data set combined with a tree depth of 3 most practical.

As the proposed knowledge stratification is a time intensive task, that relies on expert knowledge, we utilized ChatGPT as a state-of-the-art LLM to generate the knowledge hierarchy and to sort each question into the best fitting subfield. To do so, we facilitate the ChatGPT API to initially obtain the overarching subfields by prompting it to 'provide a list of the 5-10 main topics of' the knowledge field captured by the given Q&A data set. Each obtained topic then represents a first-level subfield. To further stratify these subfields, we have been inspired by the work of Welbl et al., who have chosen study textbooks as base for the SciQ natural sciences dataset, since such textbooks are relevant and linguistically tailored towards a student audience [56]. In our scenario, we focus on the outline of such textbooks, which meaningfully structure the covered subject area. However, in contrast to Welbl et al., who have reviewed particular books, we again make use of ChatGPT by prompting it to 'provide an outline for a textbook' about the given subfield. Since the generated textbook outline contained an introductory and a concluding chapter in all reviewed cases, we decided to automatically drop the first and the last chapter from the obtained list if it contained a respective keyword like 'Introduction' or 'Conclusion'. Thus, with carefully crafted prompts (see Supplement for details), we are able to obtain subfield hierarchies, from which we can build up or knowledge hierarchy data structure. All these steps are fully automatized, and the respective implementation, which is detailed in Section 6, will be open sourced upon publication of this paper. Nevertheless, in some cases users might want to change the obtained hierarchy. Accordingly, we allow for changes during any step during the hierarchy generation, or directly of the final hierarchy, before the questions are associated with subfields.

Since this LLM-based knowledge hierarchy generation approach could also be affected by hallucinations, they need to be carefully checked by a domain expert. However, as hallucinations are also relatively easy to detect in abstractive summarization [18], we do not see this as a hinderness, especially when considering the time it takes to curate a new Q&A dataset. In the hierarchies created for this paper, we could not spot any hallucination artifacts.

4.2 Hallucination Score

As we believe that LLMMaps can be a great tool to spot misinformation, we introduce a simple hallucination score which we use for visualization demonstration. Therefore, we consider the difficulty levels of the processed questions. So first, we ask the model to self-assess the difficulty level of each individual question on a 5 point Likert scale. We do this based on the assumption that wrong answers on self-assessed easy questions, are a strong indicator for hallucinations, as we expect the model to express uncertainty in cases, where it realizes that it faces a difficult question. This is also in line with Ji et al.’s definition of an AI hallucination as a confident response by an AI that cannot be grounded in any of its training data [18]. To then obtain a measure of hallucinations, we consider the ratio function \( f_i \) of the average accuracy over self-assessed difficulty level. While it is difficult to make general assumptions about \( f_i \), its function values can be expected to decrease monotonically with increasing difficulty. Since \( f_i \) is discrete, we use the monotonicity index for discrete functions [10] to compute the monotonicity of \( f_i \):

\[
\mathbf{S}_h = \frac{1}{N - 1} \sum_{i=0}^{N-1} \text{sign}(x_{i+1} - x_i),
\]

where in our case \( N = 5 \) is the number of observations and \( x_i \) the average accuracy for difficulty level \( i \). From this follows \( S_h \in [-1,1] \) with \( S_h = -1 \) being a strictly monotonically decreasing function. So, for our visualization purposes, we normalize \( S_h \) to lie in the range \([0,1]\), where 0 stands for no hallucination, while 1 stands for maximal hallucination.

We want to emphasize, that the main contribution of our paper is the visualization part of LLMMaps, and that the derived hallucination score is merely a tool to showcase the capabilities of our visualization, rather than a production ready LLM evaluation metric. We would leave the development of these open for NLP researchers, and would be happy to see them replacing this score in LLMMaps with other metrics derived in literature [18].

5 Visualization Design

In this section, we will detail our visualization design, carefully crafted according to the design considerations discussed in Section 3.

5.1 Visualization Layout

To support stratified knowledge-based visualization, our main visualization layout follows that of a mindmap. Mindmaps are not only used for problem-based learning, small-group teaching, as well as in examination tools [14], but it could also be shown that they engage learners [8,42] and positively affect learning outcomes in different domains [4,12]. With their explicit representation of the knowledge hierarchy, we also assume that the interpretation of LLMMaps does not require lengthy explanations. We further exploit an axis-parallel alignment of the mindmap, as opposed to a radial alignment, as we consider it a better fit for the usage of LLMMaps in scientific publications.

When using the knowledge stratification strategy suggested in Section 4.1 together with a horizontal axis-parallel layout, we have experienced figure height, as opposed to figure width, as the main constraining factor for all handled Q&A datasets. Therefore, we have decided to split the knowledge tree at the root node into halves and visualize one half oriented towards the left and one towards the right side. This
Fig. 2: Comparison of different values for leaf stacking parameter $h_{le}$. Larger values for $h_{le}$ result in more leaf nodes stacked (top row), which trades overall figure width for height (bottom row). Parent-child relations on leaf level are emphasized by parents encapsulating child nodes.

design decision is in line with Ondov et al.’s findings, which report cases with significant task performance improvements when arranging small multiples in a mirror-symmetric fashion [34]. While this layout reduces figure height roughly by 50%, figure height still remained the major limiting factor. Therefore, we have further decided to be able to horizontally stack leaf nodes in such a way, that we are able to exploit the full page width, while at the same time reducing the figure height. To realize this leaf node stacking, we introduce a user parameter $h_{le}$, which specifies the maximum number of leaves that are visualized on top of each other for a given parent. Figure 2 illustrates this concept for $h_{le} = 3, 4, 5$. As can be seen, the stacked leaf nodes are encased by their parent nodes, to better communicate parent-child relations. Since we found that $h_{le} = 3$ worked well for all handled Q&A datasets we decided to set this as the default value.

Another important consideration, which affects figure size, is whether to show empty leaf nodes, i.e., those subfields for which the Q&A dataset does not contain any questions. For the knowledge stratification strategy suggested in Section 4.1, we found that the handled Q&A datasets resulted in the following percentages of empty leaf nodes: PubMedQA [19] 51.00%, MedMCQA [36] 30.08%, SciQ [56] 21.13%, and USBAR [31] 19.12%. Especially for PubMedQA and MedMCQA the rather high number of empty leaf nodes, has a significant contribution to overall figure height, especially for low $h_{le}$. Empty leaf nodes play an important role, when analyzing Q&A datasets, as they indicate subfields which are not represented by the dataset. However, when the main goal is to see how a LLM performs on a given benchmark dataset, we consider them less relevant. Therefore, we have decided to omit empty leaf nodes per default from our visualization, while still giving users the opportunity to override this with a dedicated user parameter.

5.2 Quantity Encoding
While the visualization layout described in the previous subsection generates the main visualization canvas, it does not provide any quantitative values regarding an LLM’s performance. As the LLM performance per subfield is the most important information to be communicated by LLMMaps, our intention was to bring across this information as prominent and as clearly as possible. Thus, we have decided to exploit bar charts for this purpose, as we visualize accuracy per subfield by means of a horizontal bar. While this enables direct performance comparison within each half of the knowledge tree, comparison across halves is less direct. Furthermore, a bar does not provide exact quantitative values without the presence of appropriate axis labels. We therefore have decided to also directly display the accuracy as number. Since, we see this only as a secondary means of performance encoding, we have traded contrast for overall visualization design and have used the primary subfield color for this text label. Thus we are able to generate LLMMaps as for instance shown in Figure 1.

While Figure 1 supports the stratified evaluation of a single LLM, it does not allow for performance comparisons of multiple LLM’s based on one or several Q&A datasets. To also support this use case scenario, we have extended LLMMaps by being able to communicate the performance of multiple models. To do so, we horizontally split the performance bar into $k$ bars, where $k$ is the number of visualized LLMs. Furthermore, while for the single-model use case scenario, we have chosen hue, as the most discriminative color feature, to visually encode subfield, we have decided to use hue to visually encode model in our multi-model scenario. Figure 4 shows an example of the resulting visualization. As can be seen, it now becomes possible to compare models per subfield, while still being able to compare their overall performance through the conventional bar chart used in most LLM performance comparisons.

While the accuracies per subfield are important, their relevance for the overall model performance on a given Q&A dataset is only interpretable, when considering the actual number of questions associated with a specific subfield. Therefore, we also have to visualize this quantity, whereby we consider it of subordinate importance as opposed to the performance score. As the number of questions per subfield can be interpreted as a density value, we decided to choose commonly used density visualizations to display it. As point-based density plots are both widely used, and non-conflicting with the used bar chart, we decided to visualize the number of questions as dot plot. Since we have both dimensions within a subfield as free variables, we have decided to realize our dot plots as an extension to Blue Noise Plots [49]. By facilitating a blue noise distribution, we are able to evenly distribute the points within each subfield, without resulting in unwanted clustering effects. To do so, we have extended the blue noise plot algorithm to support point distribution within arbitrary shapes, as given by our subfield regions. During point relaxation, our approach computes on which side of a closed path consisting of multiple cubic Bezier curves a point lies. Once a point is outside of this path, we move it back along the direction normal onto the path. Doing so ensures that we obtain equally distributed points within each subfield. Figure 3 shows a comparison of our blue noise distribution as compared to a unconstrained random distribution, where overlap and clustering affects the perception of question density.

5.3 Visualization Augmentation
Besides the main visualization layout, and the two essential quantities, i.e., accuracy and number of questions, we believe that LLMMaps are also useful to visualize further information relevant for the evaluation of LLMs. In this section, we would like to outline three such examples, which can be visualized in an aggregated manner for each level in our knowledge hierarchy.

Textual legend. Most obvious, is the integration of global information by means of title and textual legend. So, for each LLMap we display a comprehensive title detailing which model(s) have been evaluated on which dataset(s), and we display the overall score of the visualized models.
Knowledge classification. Furthermore, since it is considered important to know, to which degree an LLM understands [46], we can also visualize Bloom’s taxonomy in the context of the evaluated LLM. To do so, we display the accuracy for each level in a pyramidal fashion, as it is usually chosen when displaying Bloom’s taxonomy (see Figure 1 (bottom)). Naturally to judge the reliability of this classification it is again important to estimate how many questions are associated with each Bloom level. To do so, we employ the same blue noise visualization as we have used to display the number of questions per subfield.

Difficulty rating. Finally, as hallucinations are considered a major risk for LLMs [54], we have also reserved some screen real estate to visualize additional information hinting towards these. We have decided to do so directly near the head of each subfield. In Figure 1, we also show how difficulty and response time can be visualized in this manner. To depict response time we suggest to use a speed gauge, which shows a value normalized over all displayed questions. To depict question difficulty, ranked on a 5 point Likert scale, we have instead opted for a discrete progress bar. Finally, to visualize our hallucination score $s_h$ we use a smiley icon with concentric circles, which fill state represents the degree of hallucination for the given subfield.

5.4 Color Coding
As in all visualization scenarios, an appropriate color coding is mandatory to obtain an interpretable visualization. Therefore, we have developed two alternative color coding strategies, which have been designed with the single- and multi-model evaluation in mind.

Data-centric color coding. Our data-centric color coding approach is meant for single-model evaluations. In this case, we have chosen a color scheme, which allows for differentiating the different subfields by employing Google’s material colors. For each subfield, we use a darker shade for the accuracy bar, whereby the subfield is visualized with a brighter shade (see Figure 1).

Model-centric color coding. The model-centric color coding is meant to be used for multi-model evaluations, where it is important to clearly associate each bar with a specific model. Therefore, we use hue, as the most dominant color parameter, to depict models, rather than subfields (see Figure 4). To also allow for differentiating the different subfields we use alternating shades for their background color.

6 IMPLEMENTATION DETAILS
Within this section, we briefly describe how we have implemented LLMMaps. Due to the widespread use of Python in the ML community, we have also used it to implement LLMMaps. We will first describe, how Q&A datasets can be imported and stored, before we describe how we obtain results for these from the LLMs. Finally, we will provide some details on implementing the visualization. All generated data and source code, for processing data and generating LLMMaps, is available on GitHub: https://github.com/sosoel.

6.1 Q&A Processing
To leverage automatic processing of different Q&A datasets in the visualization pipeline, they first need to be transformed into a common data structure. To define what information fields are necessary we combined the knowledge obtained from probing a large number of Q&A datasets, whereby we for instance considered the HELM dataset [24] and the datasets provided by the University of Freiburg’s Algorithms and Data Structures Group\(^1\). Based on these reviewed datasets, our data structure contains fields for the question text, short and long versions of the answer, context information, the question type, fields for Bloom classification and difficulty rating, the topic of the question, and if applicable the answer options. Depending on the raw unprocessed dataset, not every field must necessarily be filled. As such, the context content for short and long answers as well as for the topic and context differs depending on question type and available information. The short answer may be an index in case of multiple choice questions, a single word where questions are expected to be answered in such a way.

\(^1\)Data and source code will be provided in case of acceptance.

The long answer consequently can be the text of the correct choice in multiple choice datasets, a short answer with reasoning, or simply a long answer for questions expecting such an answer. The topic field is solely dependent on the available information of the dataset. If no more information is provided it is the general topic of the original dataset. Else it can contain a general topic of each question independently or a hierarchical list of topics and subtopics for each sample. The field for context information can either contain information which clarifies the setting of the question and may be required for a correct answer, or a reference text which contains the correct answer [56]. The fields for the Bloom classification and the difficulty rating are usually not available for a given dataset and contain the LLM-assessed values as discussed in Section 3.2. Question import is in most cases done by mapping variously named data fields of the original datasets to the corresponding fields in our internal data structure. In rare cases the original data needs to be processed with pattern matching techniques to split the data within a single field in the original data structure to obtain data for multiple fields in the target structure. With the described procedure we have processed the PubMedQA, SciQ, MedMCQA, US Bar-MBE, and the four knowledge-based Q&A data sets from the HELM library (MMLU [15], NaturalQA [22], OpenbookQA [29], WikiFact [37]), which we will release together with the automated processing scripts upon publication of this paper.

6.2 LLM Interface
To obtain correct and independent results in a visualizable and informative structure, the imported data needs to be processed in three steps. First we use the method proposed in Section 4.1 to stratify the knowledge field, resulting in a hierarchical, tree-like structure which forms the visualization basis of LLMMaps. Next we incorporate more information about each question by prompting the LLM to be evaluated to provide a self-assessed difficulty rating, and we further obtain a Bloom level classification [1] as described above. Finally, we need to obtain the actual information to be visualized with LLMMaps and thus need to obtain the answers for each question for each evaluated LLM. All of this happens also in an automated manner. The prompting for question answering follows some simple design rules. First of all, to keep the results as comparable as possible with respect to usability of the models, we kept the respective prompts for each model as similar as possible and did not engage in any finetuning of single models. The general setup of these prompt follows a few-shot approach, where each prompt consists of three samples with a known and given answer and one sample without an answer, which has to be generated by the model. The formulation is such, that the generated answer would naturally complete the prompt. For details and examples on the prompt setup, refer to the Supplement. The only exception to this rule is given by the size restrictions on the input prompt for different models. As such ChatGPT and GPT-3 [5] can process up to 4,096 tokens as input and response combined. BLOOM [40] can take only up to 1,000 input tokens when using the Huggingface API. As the smaller models GPT-2 and LLaMA-13B [48] are used offline, their input size is only restricted by GPU memory. Depending on the size and availability of context information and the possible size of the prompt, we then clip the context information of the few-shot examples in cases where it is not necessary to answer the question and if the restrictions are still not met, we lower the number of few-shot examples. For each dataset the examples used for few-shot prompting are constant and removed from the evaluated set of answers. For all of these tasks, we ensure a session reset of the respective LLM model, to remove possible bias stemming from past requests.

6.3 Visualization
As LLMMap’s main application is communicating LLM performance, our implementation goal was to be able to use LLMMaps in a wide range of contexts, reaching from scientific publications to webpages. Accordingly, we have opted for using SVG as a salable and universal realization for LLMMaps, such that they can be used in many contexts. As we want ML experts to easily generate and configure LLMMaps, we have realized a Python implementation that can generate LLMMaps.
While this makes the integration of interaction paradigms a bit more cumbersome, as compared to for instance a D3 implementation, we felt that the ease of use and adaptations outplay this downside.

7 USE CASES

In this section we demonstrate the applicability of LLMMaps by discussing different use cases.

Single-model use cases. To get a better insight on the performance of a single model, we show the single-model use case of LLMMaps in Figure 1. One thing to notice from the displayed knowledge stratification is, that ChatGPT performs significantly different for different knowledge fields in the same dataset. As such its accuracies on the first level of our knowledge hierarchy range from 59% to 78%, hinting that especially for genetics, regenerative medicine and immunities lower than on all other fields. We also observe a significantly different relative model performance between tasks. GPT-NeoX performs significantly better on MMLU than all other models and also second best on OpenbookQA, despite being moderately sized. At the same time, it only scores the second and third place on NaturalQA and WikiFact respectively. This is likely caused by MMLU and OpenbookQA being multiple choice questions, while the other two are free-form questions.

8 QUALITATIVE EVALUATION

To obtain feedback on LLMMaps from LLM users and experts, we have conducted two evaluation studies. First, an informal interview session with three medical researchers, in order to obtain feedback on the usage scenarios in application domains possibly using LLMs (see Subsection 8.1). Second, a structured interview session with four language model developers, where we asked them specific questions in order to assess how useful LLMMaps can be for LLM development and training (see Subsection 8.2).

8.1 LLM User Evaluation

Within an informal interview session, we have confronted three medical researchers from the areas of radiology, nuclear medicine and medical radiation physics (between 11 and 27 years of medical research experience) with our analysis of ChatGPT’s performance on PubMedQA (see Figure 1). While all of them knew about ChatGPT and to some extent about LLM in general, two already had made first medical experiments with ChatGPT and where in general impressed by its text processing capabilities, while being at the same time cautious regarding its factual knowledge. As none of the three researchers knew about the PubMedQA dataset, we provided a brief introduction, detailing the question types and available annotations. All three appreciated the availability of the dataset, and were curious to learn about ChatGPT’s performance on this dataset.

Confronting the three researchers with LLMMaps for ChatGPT on PubMedQA sparked engaging and explorative behavior. All three were interested to know, which medical fields were covered, and looked in particular into subfields representing their own area of research. After an initial exploration, we have explained the process, which we have applied to obtain the knowledge hierarchy, and asked them if they can spot subfields which are unexpected or missing. After giving it some thought, they were realizing, that some relevant basic medical subfields, such as orthopedics, dermatology, ophthalmology or internal medicine in general, were missing or underrepresented. While they could not spot any subfield, they would have considered unexpected, they pinpointed one example, where they would have expected the hierarchy to be different, i.e., ‘Computed tomography (CT) imaging’ being a subfield of ‘Radiographic Imaging’. After a small discussion they though acknowledged that this could also be a valid interpretation. Nevertheless, asked which alternative approaches they could imagine to obtain a knowledge hierarchy for the field of medicine, they suggested to scrape information from university hospital websites and use the underlying department structure. To follow up on this idea, we have prompted ChatGPT with “Which medical departments should a medical university hospital contain?”, and obtained the following list, which in fact contains the missing subdisciplines: Emergency Medicine, Internal Medicine, Surgery, Pediatrics, Obstetrics and Gynecology, Radiology and Imaging, Pathology, Neurology, Cardiology, Orthopedics, Dermatology, Ophthalmology, Oncology, Psychiatry, Rehabilitation Medicine, Infectious Diseases, Anesthesiology. We believe that this nicely shows, that modern LLMs have the capability to provide knowledge hierarchies while the output though highly depends on the prompting. For the future, this would mean that knowledge hierarchy prompting should further be investigated, and that the prompts should probably also be communicated with the knowledge hierarchy.

Regarding the visualization part of LLMMaps, after a brief (1 minute) introduction, all three researchers could directly understand how to interpret the visualization, and how to derive and compare the accuracy numbers for the different subfields. They were able to compare different subfields and in particular curious to know how ChatGPT performs on their own area of research. Since initially we discussed how
the amount of questions per subfield can be seen, they could success-
fully interpret our Q&A sample encoding as they immediately realized
and mentioned that overall ethical questions seem to be less frequent in
the data. So overall, we believe that this initial and informal interview
shows, that application experts see the need for better understanding
the knowledge capabilities of LLMs. Furthermore, we conclude that
LLMMaps do not only provide an engaging mechanism to visually
explore these capabilities, but that the visual encoding is also intuitive
and understandable. Finally, the separation of knowledge hierarchy
generation and visualization provides needed flexibility, which allows
for the adaptation of knowledge hierarchy generation for dedicated use-
cases. One of the medical researchers for instance stated, that he would
be interested in seeing a knowledge hierarchy which is focused on his
area of research and possibly also containing medical questions beyond
PubMedQA.

8.2 LLM Developer Evaluation

To further investigate the benefits of LLMMaps in the context of LLM
development, we conducted a more technical evaluation with four aca-
demic researchers in the form of structured one-on-one interviews. The
four participants were selected due to their familiarity with LLMs (be-
tween 1 and 2 years experience working with LLMs). Each participant
was shown three figures showing applications of LLMMaps on the
SciQ dataset. Two showing the performance of ChatGPT with and
without empty leaf nodes, and one multi-model case comparing the
performance of GPT-3, ChatGPT, and BLOOM on the same dataset.
The interviews comprised of open ended questions, and the participants
have not been compensated for the interviews.

We first asked the participants whether they could draw more informa-
tion from LLMMaps as compared to the conventional display of model
accuracies and if this information is helpful. To this we received
in general positive responses from every participant with some addi-
tional feedback regarding the visualization of the multi-model case, as
well as the number of samples in the leaf nodes, which we further
incorporated into our visualization designs as shown in this paper. The
participants mentioned the combined display of information in a single
figure, the stratification of the data domain as well as the containment
of additional quantifiers like the difficulty, response time and Bloom
taxonomy as valuable. Next, we asked the participants about the com-
prehensibility of LLMMaps. All participants found the visualization
easy to comprehend, apart from minor flaws perceived in the visual
design which we were able to incorporate into our final design. While
all found it especially helpful, that the multi-model LLMMaps quickly
convey the relative model performances for individual subfields, one
participant noted that he would like to see LLMMaps in addition to the
conventional model accuracy bar charts, as LLMMaps are harder
to digest than those due to the additional information. We took also
this suggestion into account, and added the overall performance in the
model legend as seen in Figure 4. We then asked the participants to
imagine an LLM application, and if LLMMaps would help them in
deciding when to trust and when not to trust the visualized models. All
participants noted that the stratified visualization helps in understand-
ing the model performance in the given subfields. Three of the four
participants also noted that on one hand the visualization of the number
of questions in a given node helps gaining trust to the model, while
also raising a feeling of caution towards such areas where only a few
samples are contained. The next two questions revolved around the
data domain. First we asked if LLMMaps help in understanding the
Q&A’s knowledge field. To this we received only affirmative answers
hinting on the visualized subfield information. On the question if this
also helps identifying content driven, i.e. knowledge based bias we
received mixed responses. While two answers were strictly affirmative
due to the stratified visualization of accuracies and the information of
sparsely or unpopulated subfields, one participant noted that it would be
helpful if a higher number of samples would be available, and another
one mentioned that this question is difficult to answer, as the training
data is not known. Lastly we asked if they prefer a visualization with
or without the empty leaf nodes. While all participants preferred the
display of empty leaf nodes due to the insights on missing data to
evaluate these areas, they also strictly found the visualization without
empty leaf nodes to be easier to comprehend and would thus prefer it
in cases where LLMMaps become larger.

Finally, we received some general feedback, where several partici-
pants pointed out that they found the visualization unique and novel,
and that they would like to use LLMMaps in their own research publi-
cations.

9 Conclusions and Future Work

In this paper, we have introduced LLMMaps as a visual metaphor for
the stratified evaluation of one or several LLMs on one or several Q&A
datasets. To our knowledge, LLMMaps provide the first approach for a
stratified evaluation of LLM performance on knowledge-based Q&A
datasets. We have developed LLMMaps in the hope, that the resulting transparency helps to cope with LLMs’ risk of misinformation [54]. The results obtained in our qualitative user studies indicate, that LLMMaps enable LLM users to better assess and compare the capabilities of existing LLMs, while at the same time supporting LLM developers with model improvements. These aspects are especially important in critical domains, where accuracy and reliability of LLMs are most crucial. Our work is well inline with current research initiatives on evaluating LLMs, while the HELM initiative [24] is only a few months old, during the time of writing of this paper. OpenAI has released their Evals benchmark repository, which is supposed to collect a diverse set of LLM datasets. In the future, we believe it would be interesting to derive appropriate hierarchies for this context, e.g., following the radial example by Wang et al. [52], and see how LLMMaps can contribute in this area.

For future work we see several possible endeavors. Most obvious would be the development of LLMMaps into an interactive application. While we have deliberately decided not to do so at this point, since we focus on the communication of evaluation results for scientific papers and elsewhere, we could imagine other scenarios which benefit from such interactions supporting an interactive exploration. Additionally, we believe that the area of knowledge stratification should get further attention. While the proposed approach works well for educational datasets, applying it to more trivia-like questions, incorporating subfields ranging from music over gossip to politics does not benefit from our textbook approach. Furthermore, in this paper we solely focus on the stratified evaluation of knowledge capabilities. Another large suite of benchmarks is available to investigate language understanding of LLMs. In the future, we believe it would be interesting to derive appropriate hierarchies for this context, e.g., following the radial example by Wang et al. [52], and see how LLMMaps can contribute in this area.

REFERENCES


