Establishing Trustworthiness: Rethinking Tasks and Model Evaluation

Robert Litschko[*1,3], Max Müller-Eberstein[†2], Rob van der Goot[2], Leon Weber[1,3], and Barbara Plank[1,2,3]

[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Department of Computer Science, IT University of Copenhagen, Denmark
[3]Munich Center for Machine Learning (MCML), Munich, Germany
[1]{rlitschk, leonweber, bplank}@cis.lmu.de
[2]{mamy, robv}@itu.dk

Abstract

Language understanding is a multi-faceted cognitive capability, which the Natural Language Processing (NLP) community has striven to model computationally for decades. Traditionally, facets of linguistic intelligence have been compartmentalized into tasks with specialized model architectures and corresponding evaluation protocols. With the advent of large language models (LLMs) the community has witnessed a dramatic shift towards general purpose, task-agnostic approaches powered by generative models. As a consequence, the traditional compartmentalized notion of language tasks is breaking down, followed by an increasing challenge for evaluation and analysis. At the same time, LLMs are being deployed in more real-world scenarios, including previously unforeseen zero-shot setups, increasing the need for trustworthy and reliable systems. Therefore, we argue that it is time to rethink what constitutes tasks and model evaluation in NLP, and pursue a more holistic view on language, placing trustworthiness at the center. Towards this goal, we review existing compartmentalized approaches for understanding the origins of a model's functional capacity, and provide recommendations for more multi-faceted evaluation protocols.

"Trust arises from knowledge of origin as well as from knowledge of functional capacity."

— Trustworthiness – Working Definition
David G. Hays, 1979

1 Introduction

Understanding natural language requires a multitude of cognitive capabilities which act holistically to form meaning. Modeling this ability computationally is extremely difficult, thereby necessitating a compartmentalization of the problem into isolated tasks which are solvable with available methods and resources [**?**]. Undoubtedly as of late 2022, we are witnessing a paradigm shift: Powerful LLMs, in the form of instruction-tuned, prompt-based generative models such as ChatGPT and GPT-4 [**?, ?, ?, ?, ?**], have found widespread adoption reaching far beyond the NLP community. Part of this success story is the casting of heterogeneous NLP tasks into sequence-to-sequence tasks [**?, ?, ?**]; which in turn enables extreme multi-task learning, and cross-task transfer learning.

This is in stark contrast to the traditional compartmentalized NLP paradigm (visualized in Figure **??**), wherein a human-motivated language task with an input expression and an output expectation is clearly formalized into a dataset with machine-readable inputs and outputs. Both feature design and model development are highly task-specific— often manually curated. Paired with evaluation protocols for comparing model predictions with human expectations via formalized metrics or qualitative judgement, this general methodology has been widely adopted and trusted.[1]

---

[*]Equal contribution.
[†]Equal contribution.
[1]While not without deficiencies, evaluation protocols were arguably more heterogeneous and established than today w.r.t. quantitative/qualitative evaluation, human judgements etc.

IMAGE NOT PROVIDED

Contemporary NLP Paradigm with language tasks formalized as datasets for which models produce predictions. Recent LLMs break down this compartmentalization (dashed lines), impacting all stages of the cycle. We argue that establishing trust requires rethinking every facet of this framework, as formalization and evaluation become increasingly difficult.

Figure 1: Contemporary NLP Paradigm

However, with contemporary LLMs this compartmentalization is breaking down—having severe impacts on all stages of the cycle. Therefore, a persistent and critical question regains importance: How can trust be established between the human and the model?

As early as 44 years ago, Hays [?] offers an attempt and provides a definition of trustworthiness (cf. quote). Today, the topic of trustworthiness is an ongoing discussion deserving special attention [?, ?, ?]. We argue that to establish trust, it is time to rethink how we deal with tasks and their evaluation.

Why now? It is getting increasingly hard to predict a priori when we can expect models trained on web-scale data to work well. Were we to live in a hypothetical world with full knowledge of origin and functional capacity, then each task instance could be routed to the right model(s) to not only tap into the LLMs' full potential, but to also enable trust in their predictions. Today, the absence of this knowledge is directly linked to our lack of trust in deploying models in real-world scenarios.

In this position paper, we synthesize contemporary work distributed throughout different subfields of NLP and ML into a conceptual framework for trust, guided by Hays [?]'s definition and centered around knowledge facets as a guiding principle for all aspects of the model development and evaluation cycle. We outline high-level desiderata (§??), and suggest directions on how to gain trust, by providing starting points of facets (§??) aimed to stipulate uptake and discussion. In §?? we discuss how trustworthiness relates to user trust.

2   Desiderata for Trustworthy LLMs

LLMs today pose a conundrum: They are seemingly universally applicable, having high functional capacity, however, the larger the model, the less we appear to know about the origins of its capabilities. How did we get here, which aspects contribute to trustworthiness, and what did we lose on the way?

In the following, we aim to provide a brief history of central trust desiderata (D1–4), discussing how our knowledge of functional capacity and its origins has changed over time.

D1. Knowledge about Model Input

In the beginnings of NLP, researchers followed strict, task-specific formalizations and had precise control over which "ingredients"[2] go into model training and inference (i.e., manual feature engineering). Neural models have caused a shift towards learning representations, improving performance at the cost of interpretability. While analogy tasks [?] have enabled analyses of how each word-level representation is grounded, contemporary representations have moved to the subword level, and are shared across words and different languages, obscuring our knowledge of the origin of their contents, and requiring more complex lexical semantic probing [?, ?]. This is amplified in today's instruction-based paradigm in which tasks are no longer formalized by NLP researchers and expert annotators but are formulated as natural language expressions by practitioners and end users [?]. The cognitive process of formalizing raw model inputs into ML features has been incrementally outsourced from the human to the representation learning algorithm, during which we lose knowledge over functional capacity.

D2. Knowledge about Model Behaviour

In the old compartmentalized view of NLP, higher-level tasks are typically broken down into pipelines of subtasks [?], where inspecting intermediate outputs improves our knowledge about model behaviour. Recently however, LLMs are usually trained on complex tasks in an end-to-end fashion [?], which makes it more difficult to expose intermediate

---

[2]We refer to ingredients as explicit inputs and LLM's parametric knowledge [?, ?].

outputs and analyze error propagation. Over time we have gained powerful black-box models, but have lost the ability to interpret intermediate states and decision boundaries, thus increasing uncertainty and complexity. Because as of today, we cannot build models that always provide factually correct, up-to-date information, we cannot trust to employ these models at a large scale, in real-world scenarios, where reliability and transparency are key. In this regard, pressing questions are e.g., how hallucination and memorization behaviour can be explained [**?**, **?**], how models behave when trained on many languages [**?**, **?**], what internal features are overwritten when trained on different tasks sequentially (catastrophic forgetting; e.g., [**?**, **?**]), how to improve models' ability to know when they do not know (model uncertainty; e.g., [**?**]), or how do LLMs utilize skills and knowledge distributed in their model parameters.

D3. Knowledge of Evaluation Protocols

The emergence of LLMs has raised the question of how to evaluate general-purpose models. Many recent efforts have followed the traditional NLP evaluation paradigm and summarized LLM performance into evaluation metrics across existing benchmark datasets [**?**, **?**, **?**, **?**, **?**]. This estimates LLM performance for tasks covered by the benchmark dataset and thus establishes trust when applying the model to the same task. However, the situation is different when LLMs are used to solve tasks outside of the benchmark, which is often the case for real-world usage of LLMs [**?**]. Then, the expected performance becomes unclear and benchmark results become insufficient to establish trust. One proposal to solve this issue is to evaluate on a wide variety of task-agnostic user inputs and report an aggregate metric [**?**, **?**, **?**, **?**]. This approach has the potential to cover a wider range of use cases, however, it relies mostly on manual preference annotations from human labelers or larger LLMs which is costly and has no accepted protocol yet.

D4. Knowledge of Data Origin

So far, we discussed trust desiderata from the viewpoint of knowledge of functional capacity. Next to this, a model's behaviour is also largely influenced by its training data. Knowledge about data provenance helps us make informed decisions about whether a given LLM is a good match for the intended use case. Therefore, open access to data must be prioritized. In compartmentalized NLP, models are trained and evaluated on well-known, manually curated, task-specific datasets. Today's models are instead trained on task-heterogeneous corpora at web scale, typically of unknown provenance. For novel tasks, this means we do not know how well relevant facets (e.g., language, domain) are represented in the training data. For existing tasks, it is unclear if the model has seen test instances in their large training corpora (i.e., test data leakage; [**?**]), blurring the lines between traditional train-dev-test splits and overestimating the capabilities of LLMs. To compound matters further, models are not only trained on natural, but also on generated data, and unknown data provenance is also becoming an issue as annotators start to use LLMs [**?**]. LLMs trained on data generated by other LLMs can lead to a "curse of recursion" where (im-)probable events are over/underestimated [**?**].

3 What Can We Do to Gain Trust Now and in Future?

In a world where generative LLMs seemingly dominate every benchmark and are claimed to have reached human-level performance on many tasks,[3] we advocate that now is the time to treat trust as a first-class citizen and place it at the center of model development and evaluation. To operationalize the concept of trust, we denote with *knowledge facets* (henceforth, facets) all factors that improve our knowledge of functional capacity and knowledge of origin. Facets can be local (instance) or global (datasets, tasks). They refer to 1) descriptive knowledge such as meta-data or data/task provenance, and 2) inferred knowledge; for example which skills are exploited. We next propose concrete suggestions on how facets can help us gain trust in LLMs based on the desiderata in §**??**.

Explain Skills Required versus Skills Employed

It is instructive to think of prompt-based generative LLMs as instance-level problem solvers and, as such, we need to understand a-priori the necessary skills for solving instances (local facets) as well as knowing what skills are actually employed during inference. Most prior work aims to improve our understanding of tasks and the skills acquired to solve them by studying models trained specifically for each task, and can be broadly classified into: (i) linguistically motivated approaches and (ii) model-driven approaches (D1). Linguistic approaches formalize skills as cognitive

---

[3]For example, GPT-4 reportedly passed the bar exam and placed top at GRE exams, see `https://openai.com/research/gpt-4`.

abilities, which are studied, e.g., through probing tasks [**?**, **?**, **?**], checklists [**?**] and linguistic profiling [**?**, **?**, **?**]. Model-driven approaches attribute regions in the model parameter space to skills [**?**, **?**, **?**, **?**]. The former can be seen as describing global facets (i.e., the overall functional capacity of black-box models), while the latter identifies local facets (i.e., skill regions in model parameters). To establish trust, we need to know what skills are required to solve instances, which is different from which skills are exercised by a model at inference time, as described next.

Besides knowledge about skills needed to solve a task, it is important to gain knowledge about what skills are actually being applied by an LLM. This is linked to explainability and transparency, corresponding to (i) understanding the knowledge[4] that goes into the inference process (D1), and (ii) the inference process itself in terms of applied skills (D2), e.g., examinations of LLMs' "thought processes". Regarding (i), existing work includes attributing training instances to model predictions [**?**, **?**] and explaining predictions through the lens of white-box models [**?**, **?**, **?**]. They are, however, often grounded in downstream task data and thus do not provide insights connected to the knowledge memorized by LLMs during pre-training (global facets). Regarding (ii), existing approaches include guiding the generation process through intermediate steps [**?**, **?**, **?**] and pausing the generation process to call external tools [**?**, **?**, **?**, **?**]. Their shortcoming is that they operate on the input level, and similarly do not capture cases where pre-existing, model-internal knowledge is applied. Furthermore, prior work has shown that LLMs follow the path of least resistance. That is, neural networks are prone to predict the right thing for the wrong reasons [**?**, **?**], which can be caused by spurious correlations [**?**].[5] On the path to gaining trust, we advocate for LLMs that are able to attribute their output to internal knowledge and the skills used to combine that knowledge. Alternatively, LLMs could be accompanied by white-box explanation models that (are at least a proxy) for explaining the inference process.

Facilitate Representative and Comparable Qualitative Analysis

Today, the standard target for NLP papers proposing a new model is to beat previous models on a certain quantitative benchmark. We argue that if datasets and metrics are well-designed and well-grounded in skills/capabilities, they can be used as an indicator of progress.[6] On the other hand, findings from negative results might be obscured without faceted quantitative analysis: even when obtaining lower scores on a benchmark, sub-parts of an NLP problem may be better solved compared to the baseline, but go unnoticed (D3). We therefore cannot trust reported SOTA results as long as the facets that explain how well sub-problems are solved remain hidden.

Complementary to holistic quantitative explanations, as proposed by HELM [**?**], we call for a holistic qualitative evaluation where benchmarks come with standardized qualitative evaluation protocols, which facilitates comparable qualitative meta-analysis. This proposal is inspired by the manually-curated GLUE diagnostics annotations [**?**], which describe examples by their linguistic phenomena.[7] Recycling existing tasks and augmenting them with diagnostic samples to study LLMs provides a very actionable direction for applying existing compartmentalization in a more targeted trustworthy way. Diagnostics samples should ideally represent the full spectrum of cognitive abilities required to solve a task. Designing these samples is however a complex task. We hypothesize that the set of required skills varies between tasks and should ideally be curated by expert annotators.

Be Explicit about Data Provenance

In ML, it is considered good practice to use stratified data splits to avoid overestimation of performance on dev/test splits based on contamination. Traditionally, this stratification was done based on, e.g., source, time, author, language (cross-lingual), or domain (cross-domain). Recent advances have hinted at LLMs' ability to solve new tasks, and even to obtain new, i.e., emergent abilities [**?**]. These are in fact similar cross-X settings, where X is no longer a property at the level of dataset sampling, but of the broader task setup. We call for always employing a cross-X setup (D4); whether it is based on data sampling, tasks, or capabilities—urging practitioners to make this choice explicit. Transparency about data provenance and test data leakage improve our trust in reported results. In practice, these data provenance facets are also valuable for identifying inferred knowledge such as estimated dataset/instance difficulty [**?**, **?**, **?**], especially when used in conjunction with the aforementioned diagnostic facets.

Data provenance is also important when drawing conclusions from benchmark results (D3). Tedeschi et al. [**?**] question the notion of superhuman performance and claims of tasks being solved (i.e., overclaiming model capabil-

---

[4]Including acquired knowledge such as common sense and world knowledge [**?**, **?**].

[5]"The sentiment of a movie should be invariant to the identity of the actors in the movie" [**?**]

[6]Note that baseline comparisons can still be obscured by unfair comparisons [**?**].

[7]https://gluebenchmark.com/diagnostics/

ities), and criticize how benchmark comparisons "do not incentivize a deeper understanding of the systems' performance". The authors discuss how external factors can cause variation in human-level performance (incl. annotation quality) and lead to unfair comparisons. Similarly, underclaiming LLMs' capabilities also obfuscates our knowledge of their functional capacity [**?**]. Additionally, in a recent study domain experts find the accuracy of LLMs to be mixed [**?**]. It is therefore important to be explicit about the limitations of benchmarks [**?**] and faithful in communicating model capabilities. At the same time, it is an ongoing discussion whether reviewers should require (i.e, disincentivize the absence of) closed-source baseline models such as ChatGPT and GPT-4, which do not meet our trust desiderata [**?**]. Closed-source models that sit behind APIs typically evolve over time and have unknown data provenance, thus lacking both knowledge of origin (D4), and the consistency of its functional capacity. Consequently, they make untrustworthy baselines and should not be used as an isolated measure of progress.

4   Trustworthiness and User Trust

So far we have discussed different avenues for improving our knowledge about LLM's functional capacity and origin, paving the way for establishing trustworthiness. From a user perspective it is essential to not only understand knowledge facets but also how they empirically impact user trust in a collaborative environment. This is especially important in high-risk scenarios such as in the medical and legal domain. One could argue, if LLMs such as ChatGPT are already widely adopted, do we already trust LLMs (too much)? To better understand user trust we need interdisciplinary research and user experience studies on human-AI collaboration. Specifically, we need to know what users do with the model output across multiple interactions (e.g., verify, fact check, revise, accept). For example, González et al. [**?**] investigate the connection between explanations (D2) and user trust in the context of question answering systems. In their study users are presented with explanations in different modalities and either accept (trust) or reject (don't trust) candidate answers. Similarly, Smith-Renner et al. [**?**] discuss how generated explanations can promote over-reliance or undermine user trust. A closely related question is how the faithfulness of explanations affect user trust [**?, ?**]. For a comprehensive overview on user trust we refer to the recent survey by Bach et al. [**?**].

While such controlled studies using human feedback are cost and time intensive, the minimum viable alternative for establishing trust may simply be the publication of a model's input-output history. In contrast to standalone metrics and cherry-picked qualitative examples, access to prior predictions enables post-hoc knowledge of model behaviour (D2), even without direct access to the model. This democratizes the ability to verify functional capacity and helps end users seeking to understand how well a model works for their task.

In summary, evaluating user trust is an integral part of trustworthiness and goes hand in hand with careful qualitative analyses and faceted quantitative evaluation. Towards this goal, we believe LLM development needs to be more human-centric.

5   Conclusions

In this position paper, we emphasize that the democratization of LLMs calls for the need to rethink tasks and model evaluation, placing trustworthiness at its center. We adopt a working definition of trustworthiness and establish desiderata required to improve our knowledge of LLMs (§**??**), followed by suggestions on how trust can be gained by outlining directions guided by what we call knowledge facets (§**??**). Finally, we draw a connection between trustworthiness as knowledge facets and user trust as means to evaluate their impact on human-AI collaboration (§**??**).

Limitations

To limit the scope of this work, we did not discuss the topics of social and demographic biases [**?**], discrimination of minority groups [**?**] and hate speech as factors influencing our trust in LLMs. Within our proposed desiderata, this facet would fall under 'Knowledge of Data Origin' (§**??**), in terms of understanding where model-internal knowledge and the associated biases originate from (D4).

Our proposed multi-faceted evaluation protocols rely strongly on human input—either via qualitative judgements and/or linguistically annotated diagnostic benchmarks (§**??**). We acknowledge that such analyses require more time and resources compared to evaluation using contemporary, automatic metrics, and may slow down the overall research cycle. While we believe that slower, yet more deliberate analyses are almost exclusively beneficial to establishing trust, our minimum effort alternative of publishing all model predictions can also be used to build user trust (§**??**).

This simple step closely mirrors the scientific method, where hypotheses must be falsifiable by anyone [**?**]. Identifying even a single incorrect prediction for a similar task in a model's prediction history, can already tell us plenty about the model's trustworthiness.

Acknowledgements