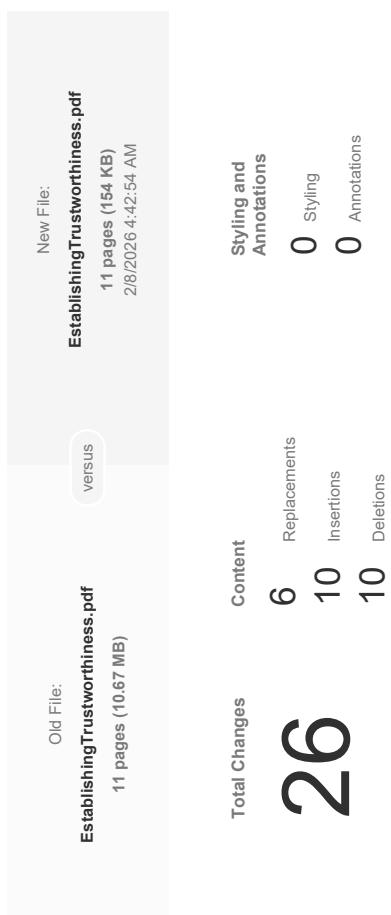


2/9/2026 12:47:58 PM

Compare Results



Summary of Comments on A91lnm2e2_7gd81v_td4.tmp

This page contains no comments

Page: 1

Image Replaced

Image Replaced

Page 1

1

גנום

Image Replaced

Establishing Trustworthiness: Rethinking Tasks and Model Evaluation

Robert Litschko^{1,3*} Max Müller-Eberstein^{2*} Rob van der Goot² Leon Weber^{1,3} Barbara Plank^{1,2,3†} [2mm] N

IMAGE NOT PROVIDED

as from knowledge of functional capacity.”

Trustworthiness – Working Definition

David G. Hays, 1979

Figure 1: 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.

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 (Schlangen, 2021). 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 (Wei et al., 2022a; Taunour et al., 2023b; Tsvetkov et al., 2023; OpenAI, 2023; Bubeck et al., 2023; *inter alia*), 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 (Rafel et al., 2020; Sanh et al., 2022; Wang et al., 2022b); which in turn enables extreme multi-task learning, and cross-task transfer learning.

This is stark contrast to the traditional compartmentalized NLP paradigm (visualized in Figure 2), wherein a human-motivated language task with an input expression and an output expectation is clearly formalized into a dataset with machine-evaluable 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 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?

¹As early as 44 years ago, Hays (1979) offers an attempt and provides a definition of trustworthiness (cf. quote).

²While not without deficiencies, evaluation protocols were arguably more heterogeneous and established than today w.r.t. quantitative/qualitative evaluation, human judgements etc.

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, a 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!”

*Equal contribution.

³Equal contribution.

Today, the topic of trustworthiness is an ongoing discussion deserving special attention (Baum et al., 2017; Eisenstein, 2022; Clarke et al., 2023). 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 (1979)'s definition centered around knowledge facets as a guiding principle for all aspects of the model development and evaluation cycle. We outline high-level desiderata (§2), and suggest directions on how to gain trust, by providing starting points of facets (§3) aimed to stipulate uptake and discussion. In §4 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 beginning 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 features engineering). Neural models have caused a shift towards learning representations, improving performance at the cost of interpretability. While analogy tasks (Milošević et al., 2013) have enabled analyses of how each word-level representation is grounded, contemporary representations have moved to the word 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 (Vulić et al., 2020, 2023).

This is simplified in today's instruction-based paradigm, in which tasks are no longer formalized by NLP researchers (De Cao et al., 2021; Malen et al., 2023), knowledge (De Cao et al., 2021; Malen et al., 2023),

and expert annotators but are formulated as natural language expressions by practitioners and end users (Ouyang et al., 2022). 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 (Manning et al., 2014), 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 (Ghoshnath et al., 2017), which makes it more difficult to expose intermediate 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 memory behaviour can be explained (Dziri et al., 2022; Malen et al., 2023), how models behave when trained on many languages (Conneau et al., 2020; Choeim et al., 2023), what internal features are overwintered when trained on different tasks sequentially (catastrophic forgetting; e.g., McCloskey and Cohen, 1989; French, 1999), how to improve models' ability to know when they do not know (model uncertainty; e.g., Li et al., 2022a), 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 (Sanh et al., 2022; Wang et al., 2022b; Seo et al., 2022; Wei et al., 2022a; Touvron et al., 2023a). 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 (Ouyang et al., 2022). 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-diagnostic user inputs and report an aggregate metric (Ouyang et al., 2022; Chung et al., 2022; Wang et al., 2023).

al., 2023b; Detmers et al., 2023). 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 functionality 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.³ 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; Pitkä et al., 2023), 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 (Veselovsky et al., 2023). LLMs trained on data generated by other LLMs can lead to “cense of recursion” where (im)probable events are over/underestimated (Shumailov et al., 2023).

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 abilities, which are studied, e.g., through probing tasks (Adi et al., 2017; Conneau et al., 2018; Amini and Cararanita, 2023), checklists (Belleiro et al., 2020) and linguistic profiling (Miaschi et al., 2020, 2021; Sarti et al., 2021). Model-driven approaches attribute regions in the model parameters space⁴ to skills (Anselli et al., 2022; Wang et al., 2022a; Pong et al., 2023; Ilharco et al., 2023). 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 explaining the knowledge⁵ 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 processes⁶. Regarding (i), existing work includes attributing training instances to model predictions (Prahl et al., 2020; Weller et al., 2023) and explaining predictions through the lens of white-box models (Frost and Hindebrand, 2017; Aytekin, 2022; Heidrich et al., 2022). 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 (Wei et al., 2022c; Wang et al., 2023a; Li et al., 2023), and pausing the generation process to call external tools (Schick et al., 2023; Shen et al., 2023; Paranjape et al., 2023; Mialon et al., 2023). 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

⁴Including acquired knowledge such as common sense and world knowledge (Li et al., 2022b; De Bruyn et al., 2022).

⁵For example, GPT-4 reportedly passed the bar exam and placed top at GRE exams, see (<https://openai.com/research/gpt-4>,

of least resistance. That is, neural networks are prone to predict the right thing for the wrong reasons (McCoy et al., 2019; Schramowski et al., 2020), which can be caused by spurious correlations (Eisenstein, 2022).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 skill/capabilities, they can be used as an indicator of progress.⁶ On the other hand, findings from negative conclusions 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 factors that explain how well sub-problems are solved remain hidden.

Complementary to holistic quantitative explanations, as proposed by HELM (Liang et al., 2022), 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 (Wang et al., 2018), which describe examples by their linguistic phenomena. (<https://gluebenchmark.com/diagnostics/>) 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.

It is therefore important to be explicit about the limitations of benchmarks (Rajai et al., 2021) and faithful in underscoring LLMs’ capabilities also obfuscates our knowledge of their functional capacity (Bowman, 2022). Additionally, in a recent study domain experts find the accuracy of LLMs to be mixed (Peskoff and Stevitt, 2023).

It is therefore important to be explicit about the limitations of benchmarks (Rajai et al., 2021) and faithful in communicating model capabilities. At the same time, it is an ongoing discussion whether reviewers should require (i.e., disentangle) the absence of closed-source baseline models such as ChatGPT and GPT-4, which do not meet our trust criterion (Rajai et al., 2023). 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.

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

⁵The sentiment of a movie should be invariant to the identity of the actors in the movie (Eisenstein, 2022).

⁶Note that baseline comparisons can still be obscured by unfair comparisons (Ruttmel et al., 2020).

⁷<https://github.com/microsoft/comdiagistics/>

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 es-

pecially important when drawing conclusions from benchmark results (D3). Tedeschi et al. (2023) question the notion of superhuman performance and claims of tasks being solved (i.e., overclaiming model capabilities) 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 (Bowman, 2022).

Additionally, in a recent study domain experts find the accuracy of LLMs to be mixed (Peskoff and Stevitt, 2023).

Closely related to the notion of trust is the notion of confidence in LLMs’ capabilities. Confidence in LLMs’ capabilities also obfuscates our knowledge of their functional capacity (Bowman, 2022). Additionally, in a recent study domain experts find the accuracy of LLMs to be mixed (Peskoff and Stevitt, 2023).

It is therefore important to be explicit about the

Limitations

pecially 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. (2021) 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-Reemer et al. (2020) 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 (Atanasova et al., 2023; Chiesanu et al., 2023). For a comprehensive overview on user trust we refer to the recent survey by Bach et al. (2022).

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 capability 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 workable definition of trustworthiness and establish desiderata required to improve our knowledge of LLMs (§2) followed by suggestions on how trust can be gained by outlining directions guided by what we call knowledge facets (§3). Finally, we draw a connection between trustworthiness as knowledge facets and user trust as means to evaluate their impact on human-AI collaboration (§4).

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This research is supported by the Independent Research Fund Denmark (DFF) Sapere Aude grant 9063-000778 and ERC Consolidator Grant DIALECT 101043325.

References

- [1] Yossi Adi, Einat Kermany, Yonatan Bellikov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In International Conference on Learning Representations.
- [2] Afra Amini and Massimiliano Cicarini. 2023. Probing in context: Toward building robust classifiers via probing large language models. arXiv preprint arXiv:2305.1471.
- [3] Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulic. 2022. Composable sparse fine-tuning for cross-lingual transfer. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages

- [1778–1796, Dublin, Ireland. Association for Computational Linguistics.
- [4] Pepa Atanasova, Oana-Maria Camburn, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. Faithfulness tests for natural language explanations. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 283–294, Toronto, Canada. Association for Computational Linguistics.
- [5] Caglar Avetkin. 2022. Neural networks are decision trees. arXiv preprint arXiv:2210.05189.
- [6] Tita Alissa Bach, Anna Khan, Hany Hallock, Gabriela Beltrao, and Sonia Souza. 2022. A systematic literature review of user trust in ai-enabled systems: An hc perspective. International Journal of Human-Computer Interaction, pages 1–16.
- [7] Kevin Baum, Maximilian A. K\"{o}chl, and Eva Schmidt. 2017. Two challenges for CI trustworthiness and how to address them. In Proceedings of the 1st Workshop on Explainable Computational Intelligence (XCI 2017), Dundee, United Kingdom. Association for Computational Linguistics.
- [8] Samuel Bowman. 2022. The dangers of underclaiming: Reasons for caution when reporting how NLP systems fail. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 7494–7499, Dublin, Ireland. Association for Computational Linguistics.
- [9] S\'{e}bastien Bubeck, Varun Chandrasekaran, Rishabh Iedan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-3. arXiv preprint arXiv:2303.12712.
- [10] Sabrina Chiesurin, Dimitris Dimakopoulos, Marco Antonio Sobrevilla Cabedo, Arash Eshghi, Ioannis Konstantas, Panagiota Papaiannou, Verena Rieser, and Ioannis Konstantas. 2023. The dangers of trusting stochastic partitions: Faithfulness and trust in open-domain conversational question answering. In Findings of the Association for Computational Linguistics: ACL 2023, pages 949–959, Toronto, Canada. Association for Computational Linguistics.
- [11] Rachelle Choenni, Dan Garrette, and Ekaterina Shitova. 2023. How do languages influence each other? Studying cross-lingual data sharing during fine-tuning. arXiv preprint arXiv:2305.13286.
- [12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirza Sajzani, Xinyang Chen, Adankanta Chowdhury, Sharar Narang, Gaurav Misra, Adams Yu, Vincent Y. Zhao, Yangqi Huang, Andrew M. Dai, Hongyan Yu, Slev Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. CoRR, abs/2210.11416.
- [13] Charles L. A. Clarke, Gianluca Demartini, Laura Dietz, Giuglielmo Faggio, Matthias Hagen, Claudia Hauff, Noriko Kando, Evangelos Kalousis, Marouf Pothast, Ian Soboroff, et al. 2023. 42zhmc: A spectrum of human-machine-collaborative relevance judgement frameworks. Frontiers of Information Access Experimentation for Research and Education, page 41.
- [14] Alexi Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzm\'{a}n, Zaidouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.
- [15] Alexi Conneau, German Kruszewski, Guillaume Lample, Loic Barrault, and Marco Baroni. 2018. What you can cram into a single * vector : Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 2136–2146, Melbourne, Australia. Association for Computational Linguistics.
- [16] Maxime De Bruyn, Ehsan Lorfi, Jeska Buhmann, and Walter Daelemans. 2022. Overfit-free word knowledge benchmark for language models. In Proceedings of the 2nd Workshop on Natural Language Generation Evaluation, and Metrics (GEN), pages 494–508, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- [17] Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. Edit-fact: Faithfulness and trust in open-domain conversational question answering. In Findings of the Association for Computational Linguistics: ACL 2021, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- [18] Tim Dettmers, Aridoro Pagani, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized lms. CoRR, abs/2305.14314.

- [19] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, and Satya Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 527–535, Seattle, United States. Association for Computational Linguistics.
- [20] Jacob Eisenstein. 2022. Informativeness and invariance: Two perspectives on spurious correlations in natural language. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 340–351, Seattle, United States. Association for Computational Linguistics.
- [21] Kavin Ethayarajh, Yejin Choi, and Swabha Swamyandipa. 2022. Understanding dataset difficulty with V-usable information. In Proceedings of the 39th International Conference on Machine Learning, volume 162 of Proceedings of Machine Learning Research, pages 5988–6008. PMLR.
- [22] Robert M. French. 1999. Catastrophic forgetting in connectionist networks. Trends in Cognitive Sciences, 3(4):128–135.
- [23] Nicholas Frost and Geoffrey Hinton. 2017. Distilling a neural network into a soft decision tree. arXiv preprint arXiv:1711.09784.
- [24] Michael Gira, Ruiwei Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.
- [25] Ana Valeria Gonçalves, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivas Tayyari. 2021. Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations. In Findings of the Association for Computational Linguistics (ACL-IJCNLP) 2021, pages 1103–1116, Online. Association for Computational Linguistics.
- [26] David G. Hays. 1979. Applications. In 17th Annual Meeting of the Association for Computational Linguistics, pages 89–89, La Jolla, California, USA. Association for Computational Linguistics.
- [27] Michael J. Heidrich, Jonas Fischer, Dietrich Klakow, Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 869–870. PMLR.
- [19] Nouha Dziri, Sivan Milton, Mo Yu, Osmar Zaiane, Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, and Satya Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 527–535, Seattle, United States. Association for Computational Linguistics.
- [20] Jacob Eisenstein. 2022. Informativeness and invariance: Two perspectives on spurious correlations in natural language. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 7901–7918, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- [21] Dongfang Li, Baotian Hu, and Qingcat Chen. 2022a. Calibration meets explanation: A simple and effective approach for model confidence estimates. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2775–2784, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- [22] Xiang Loraine Li, Adithguna Encoro, Jordan Hoffmann, Cyprien de Masson d’Autume, Phil Blunsom, and Alida Nematzadeh. 2022b. A systematic investigation of commonsense knowledge in large language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- [23] Nicholas Frost and Geoffrey Hinton. 2017. Distilling a neural network into a soft decision tree. arXiv preprint arXiv:1711.09784.
- [24] Michael Gira, Ruiwei Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.
- [25] Ana Valeria Gonçalves, Gagan Bansal, Angela Fan, Yashar Mehdad, Robin Jia, and Srinivas Tayyari. 2021. Do explanations help users detect errors in open-domain QA? An evaluation of spoken vs. visual explanations. In Findings of the Association for Computational Linguistics (ACL-IJCNLP) 2021, pages 1103–1116, Online. Association for Computational Linguistics.
- [26] David G. Hays. 1979. Applications. In 17th Annual Meeting of the Association for Computational Linguistics, pages 89–89, La Jolla, California, USA. Association for Computational Linguistics.
- [27] Michael J. Heidrich, Jonas Fischer, Dietrich Klakow, Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 869–870. PMLR.

- pages 55–60. Baltimore, Maryland. Association for Computational Linguistics.
- [36] Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of Learning and Motivation*, volume 24, pages 169–165. Elsevier.
- [37] Tom McCoy, Ellis Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.
- [38] Giuseppe Mislato, Roberto Densi, Maria Lomeli, Christoforos Nalplankis, Ram Pasunuru, Roberta Raffelani, Baptiste Roz’Fer, Tino Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. arXiv preprint arXiv:2302.07842.
- [39] Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venuti. 2020. Linguistic profiling of a neural language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 745–756, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- [40] Alessio Miaschi, Dominique Brunato, Felice Dell’Orletta, and Giulia Venuti. 2021. What makes my model perplexed? A linguistic investigation on neural language models perplexity. In *Proceedings of Deep Learning Inside Out (DeeLO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47, Online. Association for Computational Linguistics.
- [41] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.
- [42] OpenAI. 2023. GPT-4 technical report. arXiv preprint arXiv:2303.08774.
- [43] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Waiwright, Pamela Mishkin, Chong Zhang, Sandhini Aravam, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Trainfog: Language models to follow instructions with human feedback. In *NeurIPS*.
- [44] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hanan Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. ART: Automatic multi-step reasoning and tool-use for large language models. arXiv preprint arXiv:2303.09011.
- [45] Denis Peskoff and Brandon Stewart. 2023. Credible without credit: Domain experts assess generative language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 427–438, Toronto, Canada. Association for Computational Linguistics.
- [46] Aleksandra Pitkä, Christopher Akiki, Paulo Villegas, Hugo Laurenççon, Gréard Dupont, Alexandra Sifsha Luccioli, Yacine Jernite, and Anna Rogers. 2023. The roots search tool: Data transparency for *JMLR*. arXiv preprint arXiv:2302.14035.
- [47] Edoardo Maria Ponti, Alessandro Soranzo, Yoshua Bengio, and Siva Reddy. 2023. Combining parameter-efficient modules for task-level generalisation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 687–702, Dubrovnik, Croatia. Association for Computational Linguistics.
- [48] Karl Popper. 1954. *Karl Popper: Logik der Forschung*. Mohr Siebeck, Tübingen, Germany.
- [49] Lee, Sharon Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text+context transformer. *The Journal of Machine Learning Research*, 33:19920–19930.
- [50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Estimating training data influence by trying gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930.
- [51] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hana, and Anandolyne Paullada. 2021. AI and the everything-in-the-whole-wide-world benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran.
- [52] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with Checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

- should that change NLP leaderboards? In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4480–4503. Online: Association for Computational Linguistics.
- [54] Anna Rogers, Niranjana Balasubramanian, Leon Derczynski, Jesse Dodge, Alexander Koller, Sasha Lucion, Maarten Sap, Roy Schwartz, Noah A. Smith, and Emma Strubell. 2023. Closed AI models make bad baselines.
- [55] Daniel Ruffinelli, Samuel Broschit, and Rainer Gemulla. 2020. You can teach an old dog new tricks! on training knowledge graph embeddings. In International Conference on Learning Representations.
- [56] Victor Sanh, Albert Welson, Colin Raffel, Stephen H. Bach, Linhang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raju, Manan Dey, M. Saiful Bari, Canwen Xu, Ujjmish Thakker, Shanya Sharma, Sharma, Eliza Szczęsza, Taewoon Kim, Gunjan Chhablani, Md Alifjonayak, Debjyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Trishala Neeraj, Jos Rozem, Abhishek Sharma, Andreia Santilli, Thibault Fery, Jason Alan Fries, Ryan Tschab and Alexander M. Rush. 2022. Multitask prompted tr@p²: Roy Schwartz, Nicholas Lourie, Swabha Swayamidhia, Yizhong Yang, Hamnaneh Hajishirzi, Noah A. Smith, and and Leah Findlater. 2020. Dataset cartography: Training and dis- Yefim Yehieli. 2020. Dataset cartography: Mapping and di- appling datasets with training dynamics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9275–9293, Online: Association for Computational Linguistics.
- [57] Gabriele Sarti, Dominique Brunato, and Felice Dell’Orletta. 2021. That looks hard: Characterizing linguistic complexity in humans and language models. In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pages 48–50, Online. Association for Computational Linguistics.
- [58] Jevon Le Scio, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzanna Ilic, Daniel Hessov, Roman Castagnini, Alexandra Sasha Lucioni, Yannick Jernite, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. CorX abs/2211.09100.
- [59] Timo Schick, Jane Dyviced-Yu, Roberto Dossi, Roberta Raileanu, Maria Lopeli, Luke Zettlemoyer, Nicola Cancedda, and Theeras Scialom. 2023. Toolformer: Language models can teach themselves to use tools. arXiv preprint arXiv:2302.04761.
- [60] David Zholangen. 2021. Targeting the benchmark: Hugo Touvron, Thibaut Lavril, Gautier Izard, Xavier Martinet, Marie-Anne Lachaux, Timothy de Lacoste, Baptiste Roizere, Naman Goyal, Eric Hambo, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Association for Computational Linguistics and the 11th Conference on Computational Linguistics (Volume 2: Short Papers), pages 670–674. Online: Association for Computational Linguistics.

- Guillaume Lample, 2023a. Llama: Open and efficient⁷ foundation language models. CoRR, abs/2302.13911.
- [69] Hugo Touvron, Thibaut Lavié, Gautier Izard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacoste, Baptiste Rozière, Naman Goyal, Eric Hambo, Faïsal Azaïz, et al. 2023b. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.
- [70] Veniamin Veselovsky, Manoel Hora Ribeiro, and Robert West. 2023. Artificial artificial intelligence: Crowd workers widely use large language models for text production tasks. arXiv preprint arXiv:2306.07899.
- [71] Ivan Vulić, Goran Glavaš, Fangyu Liu, Nigel Collier, Edoardo Maria Ponti, and Anna Korhonen. 2023. Probing across-lingual lexical knowledge from multilingual sentence encoders. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2081–2097. Dubrovnik, Croatia.
- [72] Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7222–7240. Online. Association for Computational Linguistics.
- [73] Alex Wang, Anupreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 353–355. Brussels, Belgium. Association for Computational Linguistics.
- [74] Zhiyuan Liu, Kaiyue Wen, Zhengyan Zhang, Jia Hou, Zhiyuan Liu, and Juanzi Li. 2022a. Finding gaps in pre-trained transformer-based language models. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 1132–1152. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- [75] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Akashika Choudhary, and Denny Zhou. 2023a. Self-consistency improves citing of thought reasoning in language models. In The Eleventh International Conference on Learning Representations.
- [76] Yizhong Wang, Hanish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khayati Raghav Chandu, David Wadden, Kortney MacMillan, Noah A. Smith, Iz Belatagy, and Hananeh Hajishirzi. 2023b. How far can models go? exploring the state of instruction tuning on open resources. CoRR, abs/2306.04751.