The Generalization and Robustness of Transformer-Based Language Models on Commonsense Reasoning

Ke Shen
Information Sciences Institute
USC Viterbi School of Engineering
4676 Admiralty Way 1001
Marina Del Rey, California 90292
keshen@isi.edu

Abstract

The advent of powerful transformer-based discriminative language models and, more recently, generative GPT-family models, has led to notable advancements in natural language processing (NLP), particularly in commonsense reasoning tasks. One such task is commonsense reasoning, where performance is usually evaluated through multiple-choice question-answering benchmarks. Till date, many such benchmarks have been proposed and 'leaderboards' tracking state-of-the-art performance on those benchmarks suggest that transformer-based models are approaching human-like performance. However, due to documented problems such as hallucination and bias, the research focus is shifting from merely quantifying accuracy on the task to an in-depth, context-sensitive probing of LLMs' generalization and robustness. To gain deeper insight into diagnosing these models' performance in commonsense reasoning scenarios, this thesis addresses three main studies: the generalization ability of transformer-based language models on commonsense reasoning, the trend in confidence distribution of these language models confronted with ambiguous inference tasks, and a proposed risk-centric evaluation framework for both discriminative and generative language models.

Generalizability Evaluation

As emphasized in a recent study (Davis 2023), the concept of *commonsense reasoning* implies that its involved commonsense knowledge is common. Thus, commonsense AI should be expected to generalize, that is, at least in aggregate, should not exhibit excessive performance loss across independent commonsense benchmarks, such as (Bhagavatula et al. 2020; Singh et al. 2021; Santos et al. 2022; Kejriwal et al. 2023), regardless of the specific benchmark on (the training set of) which it has been fine-tuned. In my first work (Shen and Kejriwal 2021), we evaluated this expectation by proposing a methodology and experimental study to measure the generalization ability of transformer-based language models using statistical significance analysis and a rigorous and intuitive metric (i.e., performance loss metric).

We conducted an in-depth evaluation of RoBERTa, a widely-used, transformer-based discriminative language model (LM), using five established commonsense reasoning benchmarks. The focus was to determine if there was a significant decrease in RoBERTa's performance when it was fine-tuned on one commonsense benchmark but tested on a different one. The experimental study shows that the models do not generalize well, and may be (potentially) susceptible to issues such as dataset bias. The results, therefore, suggest that current performance on benchmarks may be an overestimate, especially if we want to use such models on novel commonsense problems for which a 'training' dataset may not be available for the language representation model to fine-tune. Additionally, it suggests that sophisticated adversarial modifications are not necessary to conclude that generalization is a concern for transformer-based QA models. Regrettably, we haven't had the chance to verify whether the findings are applicable to the generative GPT-series models, which have gained considerable recognition recently. Examining if these conclusions hold for GPT models is crucial for their broader application in real-world scenarios. applications.

Risk Evaluation

Beyond the study of the generalizability of transformer-based language models, there is a growing body of research dedicated to diagnosing these models' performance in commonsense reasoning scenarios. (Wu et al. 2020) found that BERT-based LM does not fully 'understand' naturalistic concepts like negation by introducing a parameter-free probing technique to recover information from the token representation. Along with the impressive performance achieved by ChatGPT, the uncertainty or risk-related issues associated with these LLMs, including hallucination, bias, and overconfidence, have reignited concerns (Ferrara 2023).

Traditionally, in machine learning, the risk of a model's inference tended to be directly equated to its confidence score: lower confidence was assumed to signal increased risk, implicitly defined as the probability that the prediction made by the model is correct. Initial research on LLMs' 'self-understanding' of their own uncertainty has predominantly relied on interpreting raw softmax probabilities of the final output layer as 'confidence' scores (Vasudevan, Sethy, and Ghias 2019). (Kadavath et al. 2022) highlighted that these generative LLMs can accurately predict which questions they will be able to answer correctly on diverse NLI tasks based on their confidence scores.

In (Shen and Kejriwal 2022), we designed two prompt-

based and one choice-based perturb functions to examine to what extent the confidence behavior captured inherent risk in an ambiguous inference task with no, or a highly controversial, correct answer in the ground truth. The proposed confusion probes stimulated the *ambiguous* situations where (for example) no correct answer exists for a given prompt among the provided set of choices. A statistical analysis of the confidence behavior of a fine-tuned, high-performance language model was conducted across widely used commonsense multiple-choice QA benchmarks.

We found evidence that, when instances are perturbed using the prompt-based functions, the confidence distribution of (the originally) incorrect answers in most perturbed instances is close to random, and is similar to the distribution observed before perturbation. The model will still prefer the originally correct (and post-perturbation, 'pseudo-correct') answer even though it is now theoretically incorrect. For the choice-based perturbation, the model will choose the incorrect choice that appears superficially closest to the prompt, despite its incorrectness. Further analysis, including an analysis of potential 'irregularities' in the benchmarks, suggests that they cannot serve as causal explanations for the observed phenomena.

Risk-adjusted Calibrator

In the ongoing study, we move beyond a single and often implicit definition of risk by introducing a risk-centric framework that defines two different types of risk, and proposing four metrics for evaluating LLMs on these two risks. We also present a novel risk-adjusted calibrator (called DwD) for adjusting the raw confidence of an underlying LLM to navigate decision and composite risks better. Results show that DwD reduces in-domain and out-of-domain decision risks more for RoBERTa by margins of 28.3% and 18.9%, respectively. Similarly, using a risk sensitivity metric, DwD reduces additional composite risk of ChatGPT NLI by a margin of 17.9%, compared to the next best baseline.

Limitations and Future Work

Although many studies (Ferrara 2023; Wu et al. 2020), as well as our work, have brought significant attention to the evaluation of commonsense reasoning in LLMs, their robustness and reliability in real-world applications are still debatable. Some evidence already suggests that the evaluation of such models on common sense may be over-reliant on the multiple-choice question format, which is amenable to automated scoring, but does not reflect real-world conversational interactions. To obtain a more comprehensive understanding of LLMs' capabilities in machine commonsense reasoning and their robustness, it is essential to employ a wider range of evaluation methods, including those that involve human-in-the-loop evaluations.

Furthermore, although not exactly as expected, the advent of LLMs does present promising avenues for addressing machine commonsense reasoning. This advancement makes it more feasible to integrate LLMs into downstream applications that necessitate natural language understanding. An innovative application of this is the development of a knowledge-based analytical system. This system would allow domain experts, including those without technical backgrounds, to interact using natural language queries. Our forthcoming objective is to design a knowledge-based analytical system tailored for any specific domain, which can be operated through natural language questions instead of complex, formal queries.

References

Bhagavatula, C.; Bras, R. L.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; tau Yih, W.; and Choi, Y. 2020. Abductive Commonsense Reasoning. In *International Conference on Learning Representations*.

Davis, E. 2023. Benchmarks for automated commonsense reasoning: A survey. *arXiv preprint arXiv:2302.04752*.

Ferrara, E. 2023. Should chatgpt be biased? challenges and risks of bias in large language models. *arXiv preprint arXiv:2304.03738*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Dodds, Z. H.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Kejriwal, M.; Santos, H.; Shen, K.; Mulvehill, A. M.; and McGuinness, D. L. 2023. Context-Rich Evaluation of Machine Common Sense. In *International Conference on Artificial General Intelligence*, 167–176. Springer.

Santos, H.; Shen, K.; Mulvehill, A. M.; Razeghi, Y.; McGuinness, D. L.; and Kejriwal, M. 2022. A theoretically grounded benchmark for evaluating machine commonsense. *arXiv preprint arXiv:2203.12184*.

Shen, K.; and Kejriwal, M. 2021. On the generalization abilities of fine-tuned commonsense language representation models. In *Artificial Intelligence XXXVIII: 41st SGAI International Conference on Artificial Intelligence, AI 2021, Cambridge, UK, December 14–16, 2021, Proceedings 41*, 3–16. Springer.

Shen, K.; and Kejriwal, M. 2022. Understanding prior bias and choice paralysis in transformer-based language representation models through four experimental probes. *arXiv preprint arXiv:2210.01258*.

Singh, S.; Wen, N.; Hou, Y.; Alipoormolabashi, P.; Wu, T.-L.; Ma, X.; and Peng, N. 2021. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. *arXiv preprint arXiv:2106.00969*.

Vasudevan, V. T.; Sethy, A.; and Ghias, A. R. 2019. Towards better confidence estimation for neural models. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7335–7339. IEEE.

Wu, Z.; Chen, Y.; Kao, B.; and Liu, Q. 2020. Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4166–4176.