Bibliographic Study of Robustness in Foundational Models

1 Introduction

Foundation models in natural language processing (NLP) began with the introduction of the Transformer architecture, which revolutionized sequence processing with self-attention [29]. Built on this, BERT [11] demonstrated the rise of pre-trained bidirectional transformers for deep contextual understanding with masked language modelling and next-sentence prediction. GPT [25] introduced generative pre-training, showing that large models could perform well on diverse tasks. However, fine-tuning the model for every downstream task was still a problem. GPT-2 [26] and GPT-3 [3] expanded on this by showcasing multitasking and few-shot learning giving rise to in-context learning where models could perform tasks with minimal or no task-specific fine-tuning simply by conditioning on a few examples. This demonstrated the rise of emerging capabilities in LLMs happening due to extensive pre-training on a large corpus. Foundation models comprise models pre-trained on transformer architecture or using contrastive learning techniques. They can used for NLP tasks or can be extended to vision [13, 27] or even graph-based tasks [18, 9].

2 Challenges

Robustness refers to the ability of a model to maintain consistent and reliable performance across diverse inputs, including those that are adversarial, out-of-distribution (OOD), or contain noise.

The major concern for foundational models is the robustness achieved during pre-training does not always transfer to downstream tasks unless robust fine-tuning is also performed [5]. For example, a vision transformer (ViT) pre-trained on ImageNet may not inherently be robust to adversarial attacks when fine-tuned for a specific task like medical image analysis [4]. Existing solutions proposed to improve robustness transfer from pre-training to fine-tuning include adversarial contrastive learning (AdvCL) [14] and robust pre-training with model sparsification [6].

2.1 Distribution shift

Detecting OOD inputs in foundation models is crucial because it helps prevent the model from generating unreliable and hallucinated content. There has been extensive work in evaluating how well a model can distinguish between IID and OOD inputs[28]. Using distributionally robust optimization (DRO) with BERT for sentiment analysis [20] has demonstrated that the DRO model improved performance on the test set with a distributional shift from the training set.

2.2 Knowledge editing

Knowledge editing refers to the process of updating a pre-trained model with new information or correcting existing knowledge without the need for retraining the entire model. knowledge-based model editing is considered a problem with a constrained optimization objective that simultaneously ensures the accuracy and retention of editing [31]. In the context of robustness, the challenge is ensuring that after the knowledge editing process, the model remains stable, and the changes do not lead to degradation in the model's generalization ability or accuracy on previously learned tasks. Hence a successful edit should ensure reliability, generalization and localization. Existing editing techniques mainly involve local modification [22], global optimization [10] and external memorization [17]. Local modification involves identifying the specific parameters associated with the updated knowledge and fixing only those specific parameters. Global optimization is learning

the updates (changes in W) on a proxy model in this case it's a hypernetwork. And then the updates (ΔW) are added to the originally pre-trained model. External memorization does not edit the model but maintains a separate database for the updates.

2.3 Adversarial attacks

LLMs are vulnerable to adversarial attacks [16] that can bypass their safety guardrails and adversarial training has proven to be one of the most promising methods to reliably improve robustness against such attacks [32]. In the context of adversarial training for LLMs, recent work has demonstrated significant progress in addressing the computational challenges associated with discrete adversarial attacks. Zero-shot robustness evaluation directly assesses the performance of LLMs on test datasets without further fine-tuning or optimization [30].

3 Evaluating robustness

Evaluating LLMs against multi-dimensional metrics including accuracy, calibration, robustness, fairness, bias, toxicity and efficiency is crucial [2, 21]. They are usually evaluated on scenarios including question answering, information retrieval, summarization, sentiment analysis, toxicity detection and other text classification tasks.

Invariance and equivariance are two important properties when evaluating LLMs for robustness[21]. These properties describe how a model's outputs change or do not change in response to specific transformations of its inputs. Invariance refers to a model's ability to produce consistent outputs despite certain input transformations, such as synonym substitution, paraphrasing or input corruption, ensuring that semantically equivalent inputs yield the same result. Whereas, equivariance refers to the change in the model's output if the input semantically changes. Framework to evaluate the robustness of LLMs during test time involved Out-Of-Context (OOC) prompting, a zero-shot method that simulates stratified counterfactual data augmentations in LLM predictions [8]. Instead of merely using counterfactual data, which is slightly infeasible to generate for all possible cases, stratified counterfactual data was generated from existing prompts. This augmented data was used to evaluate the model against perturbations and bias. There exist comprehensive evaluation frameworks for LLMs, such as HELM [21], Google BIG-Bench, and OpenAI Evals, which provide structured and multi-dimensional approaches to assess LLM's performance across accuracy, robustness, fairness, and efficiency. Robustness Gym [15], an evaluation toolkit that supports a broad set of evaluations for NLP tasks. They designed a single interface tool to evaluate the model across multiple robustness dimensions, including adversarial attacks, data shifts and subgroup analysis to assess fairness and performance across demographic subgroups.

3.1 Datasets

Natural Language Augmenter is used to check invariance, which is a framework for generating noisy textual augmentations.

The selection of the dataset depends on the downstream tasks such as question answering, sentiment analysis or information retrieval.

- BoolQ (Boolean Questions) [7] is a question-answering scenario that focuses on yes/no questions. The input consists of a question and a passage from a Wikipedia article that potentially contains the answer. The model must predict whether the answer to the question is yes or no.

- NaturalQuestions [19] is a question-answering scenario featuring factual, naturally-occurring questions derived from Google search queries that have a Wikipedia page in the top 5 results.
- IMDB Movie Reviews is a sentiment analysis dataset consisting of movie reviews from IMDB users. Each review is labelled as either positive or negative based on the user's rating. Used to evaluate the robustness of sentiment analysis models on IMDB using both synthetic perturbations and a contrast set.
- MS MARCO (Microsoft Machine Reading Comprehension) [23, 1] evaluates robustness by applying perturbations to the input queries and assessing how these changes affect the model's ability to rank the truly relevant passages highly.
- BOLD (Bias in Open-Ended Language Generation Dataset) [12] to evaluate toxicity in the generated output given a toxic or neural input. This dataset provides more toxic prompts sourced from Wikipedia articles related to professions, gender, race, religion, and political ideology. The goal is to assess if models generate toxic content even with neutral prompts, and how this varies across different demographic groups.
- BBQ (Bias Benchmark for Question Answering) [24] dataset assesses social biases in language models through question-answering. Different prompts in this dataset are designed to probe for biases against specific social groups.

4 Conclusion

Limitations of existing methods

- For knowledge editing, existing techniques like local parameter modification and hypernetwork-based global optimization can introduce unintended effects, such as catastrophic forgetting or interference with previously learned information. Balancing edit reliability, generalization, and localization remains challenging.
- Methods such as adversarial training, contrastive learning, and model sparsification, add significant computational cost. This makes them impractical for large-scale models.

References

[1] Payal Bajaj et al. "Ms marco: A human generated machine reading comprehension dataset". In: *arXiv preprint arXiv:1611.09268* (2016).

[2] Rishi Bommasani et al. "On the opportunities and risks of foundation models". In: *arXiv preprint arXiv:2108.07258* (2021).

[3] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[4] Pin-Yu Chen et al. "Foundational Robustness of Foundation Models". In: *Annual Conference on Neural Information Processing Systems*. 2022.

[5] Tianlong Chen et al. "Adversarial robustness: From self-supervised pre-training to fine-tuning". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 699–708.

[6] Tianlong Chen et al. "Data-efficient double-win lottery tickets from robust pre-training". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 3747–3759.

[7] Christopher Clark et al. "Boolq: Exploring the surprising difficulty of natural yes/no questions". In: *arXiv preprint arXiv:1905.10044* (2019).

[8] Leonardo Cotta and Chris J Maddison. "Test-Time Fairness and Robustness in Large Language Models". In: *arXiv preprint arXiv:2406.07685* (2024).

[9] Haotian Cui et al. "scGPT: toward building a foundation model for single-cell multi-omics using generative AI". In: *Nature Methods* 21.8 (2024), pp. 1470–1480.

[10] Nicola De Cao, Wilker Aziz, and Ivan Titov. "Editing factual knowledge in language models". In: *arXiv preprint arXiv:2104.08164* (2021).

[11] Jacob Devlin et al. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.

[12] Jwala Dhamala et al. "Bold: Dataset and metrics for measuring biases in open-ended language generation". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 862–872.

[13] Alexey Dosovitskiy et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[14] Lijie Fan et al. "When does contrastive learning preserve adversarial robustness from pretraining to finetuning?" In: *Advances in neural information processing systems* 34 (2021), pp. 21480–21492.

[15] Karan Goel et al. "Robustness gym: Unifying the NLP evaluation landscape". In: *arXiv preprint arXiv:2101.04840* (2021).

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).

[17] Tom Hartvigsen et al. "Aging with grace: Lifelong model editing with discrete key-value adaptors". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 47934–47959.

[18] John Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *nature* 596.7873 (2021), pp. 583–589.

[19] Tom Kwiatkowski et al. "Natural Questions: a Benchmark for Question Answering Research". In: *Transactions of the Association of Computational Linguistics* (2019).

[20] Shilun Li, Renee Li, and Carina Zhang. "Distributionally Robust Classifiers in Sentiment Analysis". In: *arXiv preprint arXiv:2110.10372* (2021).

[21] Percy Liang et al. "Holistic evaluation of language models". In: *arXiv preprint arXiv:2211.09110* (2022).

[22] Kevin Meng et al. "Locating and editing factual associations in gpt". In: *Advances in neural information processing systems* 35 (2022), pp. 17359–17372.

[23] Tri Nguyen et al. "Ms marco: A human-generated machine reading comprehension dataset". In: (2016).

[24] Alicia Parrish et al. “BBQ: A hand-built bias benchmark for question answering”. In: *arXiv preprint arXiv:2110.08193* (2021).

[25] Alec Radford et al. “Improving language understanding by generative pre-training”. In: (2018).

[26] Alec Radford et al. “Language models are unsupervised multitask learners”. In: *OpenAI blog* 1.8 (2019), p. 9.

[27] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PmLR. 2021, pp. 8748–8763.

[28] Jie Ren et al. “Out-of-distribution detection and selective generation for conditional language models”. In: *arXiv preprint arXiv:2209.15558* (2022).

[29] Ashish Vaswani et al. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[30] Jindong Wang et al. “On the robustness of chatgpt: An adversarial and out-of-distribution perspective”. In: *arXiv preprint arXiv:2302.12095* (2023).

[31] Song Wang et al. “Knowledge editing for large language models: A survey”. In: *ACM Computing Surveys* 57.3 (2024), pp. 1–37.

[32] Sophie Xhonneux et al. “Efficient adversarial training in llms with continuous attacks”. In: *arXiv preprint arXiv:2405.15589* (2024).