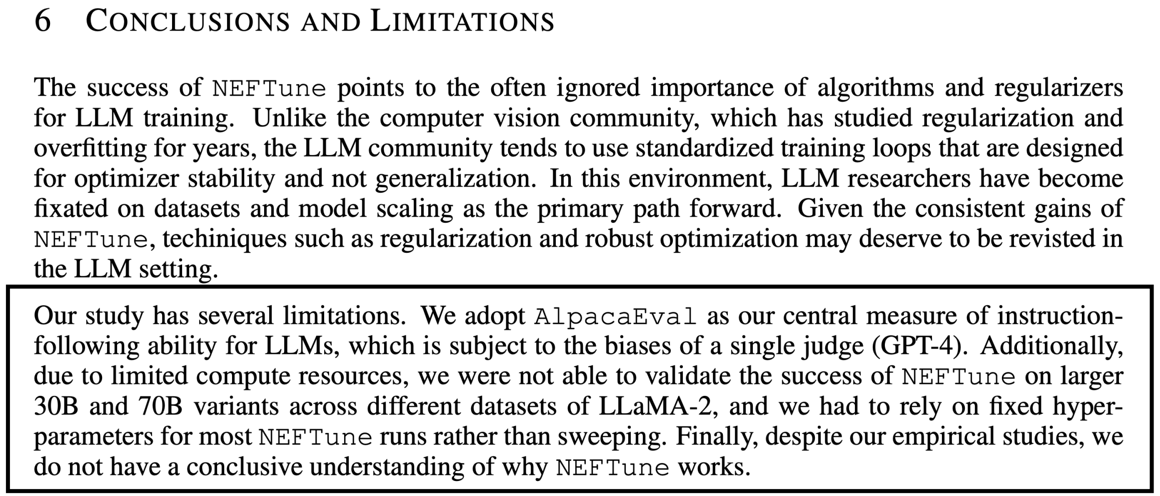
Authors' Proactively Disclosed Limitations



6 Conclusions and Limitations

The success of NEFTune points to the often ignored importance of algorithms and regularizers for LLM training. Unlike the computer vision community, which has studied regularization and overfitting for years, the LLM community tends to use standardized training loops that are designed for optimizer stability and not generalization. In this environment, LLM researchers have become fixated on datasets and model scaling as the primary path forward. Given the consistent gains of NEFTune, techiniques such as regularization and robust optimization may deserve to be revisted in the LLM setting.

Our study has several limitations. We adopt AlpacaEval as our central measure of instruction-following ability for LLMs, which is subject to the biases of a single judge (GPT-4). Additionally, due to limited compute resources, we were not able to validate the success of NEFTune on larger 30B and 70B variants across different datasets of LLaMA-2, and we had to rely on fixed hyper-parameters for most NEFTune runs rather than sweeping. Finally, despite our empirical studies, we do not have a conclusive understanding of why NEFTune works.

Our study has several limitations. We adopt AlpacaEval as our central measure of instruction following ability for LLMs, which is subject to the biases of a single judge (GPT-4). Additionally, due to limited compute resources, we were not able to validate the success of NEFTune on larger 30B and 70B variants across different datasets of LLaMA-2, and we had to rely on fixed hyperparameters for most NEFTune runs rather than sweeping. Finally, despite our empirical studies, we do not have a conclusive understanding of why NEFTune works.

LLM's Review of the Paper's Weaknesses

- Over-reliance on AlpacaEval as the primary evaluation metric.
    - There might be concerns about the bias inherent in a single evaluation metric (namely, GPT-4 evaluation).
    - Limited insight into other facets of instruction-following abilities or real-world applicability beyond model-generated metrics.
- Lack of comprehensive testing on larger model variants and datasets.
    - Paper admits the experiments exclude larger models like 30B and 70B, potentially leaving doubts about adaptability and scalability.
    - Possible untapped performance implications on larger datasets and more complex real-world tasks are not addressed.
- Uniform treatment of hyperparameters without finer adjustments or tuning.
    - Relying on a set of fixed hyperparameters might overlook potential peaks in performance, as model architectures and datasets vary significantly.
    - Limited exploration could lead reviewers to question the thoroughness and accuracy of reported improvements.
- Lack of explanation underlying the mechanism of improvement via noise.
    - Despite empirical success, the paper lacks theoretical explanation or understanding as to why NEFTune enhances conversation quality.
    - Practitioners may be hesitant to adopt without a clear understanding of underlying principles.