Generalizing Trust: Weak-to-Strong Trustworthiness in Language Models

Anonymous Authors[1]

Abstract

As large language models continue to advance, ensuring their trustworthiness is critical. However, inaccessible real-world ground truth labels pose a significant challenge in high-stakes domains. Recent studies have highlighted weak-to-strong generalization, where a strong model trained only on a weak model's labels surpasses the weak model in task performance. Yet, whether critical trustworthiness properties such as robustness, fairness, and privacy can generalize similarly remains an open question. This is the first work to study this question by examining if a stronger model can enhance trustworthiness when fine-tuned on a weaker model's labels, a paradigm we term *weak-to-strong trustworthiness*. To address this, we introduce two fundamental fine-tuning strategies that leverage trustworthiness regularization during the fine-tuning of the weak and weak-to-strong models. Our experimental evaluation on real-world datasets reveals that while some trustworthiness properties, such as fairness, adversarial, and OOD robustness, show significant improvement in trustworthiness generalization when both models were regularized, others like privacy do not exhibit signs of weak-to-strong trustworthiness. Our results highlight the potential of weak-to-strong trustworthiness as a practical pathway for enhancing the trustworthiness of increasingly capable AI systems, even under imperfect real-world conditions.

1. Introduction

In recent years, developments in large language models (LLMs) have demonstrated breakthroughs in capability and scale (Radford et al., 2019; Bubeck et al., 2023). As models continue to improve, trustworthiness has emerged as a critical aspect of AI systems, especially as LLMs are increasingly deployed in high-stakes domains like healthcare, finance, and criminal justice (Wang et al., 2023).

A fundamental challenge in developing trustworthy models is that real-world supervision is often imperfect. The lack of ground-truth labeled data is a bottleneck for training capable models, particularly in the domains where trustworthiness matters most. Consider the case of judicial bail: training data comes from judges' decisions about pretrial release, but these human judgments may carry inherent biases (Lakkaraju et al., 2017). Moreover, obtaining complete ground truth labels is practically infeasible; defendants cannot all be released ethically simply to observe outcomes. Similarly, in loan approval systems, we only observe repayment outcomes for approved applications, rendering any training data incomplete and unreliable due to the inherent selection bias. The challenge of imperfect supervision parallels a question in AI alignment: if we only have access to potentially biased supervision (like human supervision), how can we control more capable AI systems to be more aligned with human values and trustworthiness?

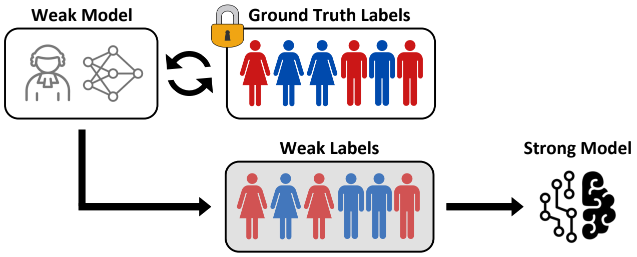


Figure 1. **Weak-to-strong framework for when ground truth labels are unavailable.** The weak model (e.g. human supervision or small LLM) has been trained to predict an inaccessible set of complete ground truth labels. The weak labels (weak model's predictions) are then used to fine-tune the strong model.

A recent study demonstrated the phenomenon of weak-to-strong (WTS) generalization, where a strong model outperforms a

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

weak model by fine-tuning on only the weak model's labels (Burns et al., 2024). Weak-to-strong learning is particularly promising for studying superalignment, where ground truth labels are unknown by humans, addressing the real-world inaccessibility of ground truth data (Bach et al., 2017; Ratner et al., 2017) (Figure 1). A few follow-up studies have focused on applying weak-to-strong learning to improve performance in various settings, yet none have investigated trustworthiness (Chen et al., 2024; Yang et al., 2024).

In this work, we introduce the *weak-to-strong trustworthiness* paradigm. We investigate the unexplored question: *Can trustworthiness properties be generalized to a strong model from fine-tuning on a weak model's labels?*

While previous work mainly use weak-to-strong learning to enhance raw predictive accuracy, our objective is to show that trustworthiness can also be improved when fully ground truth labels remain unavailable (Chen et al., 2024; Yang et al., 2024). In the context of the superalignment scenario, our approach examines if superintelligent strong models trained on human weak labels can overcome human biases to become more trustworthy.

To enable a systematic study of this phenomenon, we find that two fundamental fine-tuning strategies serve as strong baselines: Weak Trustworthiness Fine-tuning (Weak TFT), which applies trustworthiness regularization during weak model training, and Weak and Weak-to-Strong Trustworthiness Fine-tuning (Weak+WTS TFT), which adds regularization during both weak model training and weak-to-strong learning.

We perform rigorous empirical experiments using the Pythia model suite (Biderman et al., 2023) to analyze our fine-tuning strategies on standard trustworthiness datasets. Our main contributions are:

- **Weak-to-strong trustworthiness is feasible:** We present the novel conceptual framework of weak-to-strong trustworthiness. As the first study examining whether trustworthiness properties generalize through WTS learning, our results indicate that *WTS trustworthiness is indeed feasible*.
- **Standard weak-to-strong learning is insufficient:** Simply fine-tuning a weak-to-strong model on a weak model's labels yields inconsistent generalization of trustworthiness across properties (fairness, OOD robustness, adversarial robustness, privacy).
- **Fundamental fine-tuning strategies improve weak-to-strong trustworthiness**: We introduce the Weak TFT and Weak+WTS TFT strategies by incorporating trustworthiness regularization within the weak-to-strong process. After regularizing the weak model and weak-to-strong learning, our Weak+WTS TFT strategy consistently improves trustworthiness generalization, significantly enhancing fairness and robustness.
- **Comprehensive empirical evaluation**: We evaluate our strategies across 4 properties, 20 datasets, 14 definitions and tasks, and 5 model sizes ranging from 14M to 6.9B parameters. In addition, our sensitivity analysis demonstrates consistent weak-to-strong trustworthiness across a wide range of hyperparameter values.

Our study is critical for understanding the promising potential and limitations of weak-to-strong trustworthiness. Our findings have broad implications for the future of AI development: by demonstrating that trustworthiness properties can be systematically enhanced as models scale, we provide a pathway for ensuring that increasingly powerful AI systems remain aligned with human values even when perfect supervision is unavailable.

2. Methodology

In this section, we present our methodology for investigating weak-to-strong trustworthiness. Our approach systematically explores whether and how fairness, robustness, and privacy can be effectively generalized from weak to strong models. We begin by outlining the weak-to-strong learning process, followed by techniques for eliciting specific trustworthiness properties in language models. Finally, we introduce a multi-strategy approach to investigate weak-to-strong trustworthiness, proposing the fundamental fine-tuning strategies: Weak TFT and Weak+WTS TFT. See Appendix B for Preliminaries.

2.1. Fine-tuning Strategies for Studying Weak-to-Strong Trustworthiness

We systematically study how trustworthiness can be generalized through three fine-tuning strategies: No TFT, Weak TFT, and Weak+WTS TFT. While No TFT is described in Burns et al. (2024), we propose the later two fundamental strategies, applying trustworthiness regularization during weak model training (Weak TFT, Weak+WTS TFT) and weak-to-strong learning (Weak+WTS TFT).

No trustworthiness fine-tuning (No TFT). This fine-tuning strategy establishes baseline performance by conducting weak model training and weak-to-strong learning without applying trustworthiness regularization, as outlined in Burns et al. (2024).

- **Weak model:** A small pretrained LLM is fine-tuned on ground truth labels ($\lambda = 0$, no regularization). Weak labels are collected from the fine-tuned weak model $f_w(\cdot, \lambda)$ on a held-out validation set.
- **Weak-to-strong learning:** We fine-tune a weak-to-strong model through standard weak-to-strong learning on the weak labels.

Weak trustworthiness fine-tuning (Weak TFT). We propose this fine-tuning strategy to investigate whether applying regularization to a weak model can lead to trustworthiness generalization through standard weak-to-strong learning.

- **Trustworthy weak model:** A small pretrained LLM is fine-tuned using equations from Section B, with $\lambda > 0$ for trustworthiness regularization. Trustworthy weak labels are collected from the fine-tuned weak model $f_w(\cdot, \lambda)$ on a held-out validation set.
- **Weak-to-strong learning:** We fine-tune the weak-to-strong model through standard weak-to-strong learning on the trustworthy weak labels.

Weak and weak-to-strong trustworthiness fine-tuning (Weak+WTS TFT). We propose this fine-tuning strategy to investigate whether applying regularization to both a weak model and the weak-to-strong learning process can lead to trustworthiness generalization.

- **Trustworthy weak model:** The weak model is the same as in the Weak TFT strategy.
- **Trustworthy weak-to-strong learning:** Instead of the standard weak-to-strong learning, regularization is applied to the fine-tuning process on weak labels; we call this trustworthy weak-to-strong learning. We provide details on this training objective in Appendix F.1.

3. Experimental Evaluation

Datasets. We evaluate trustworthiness generalization using 20 datasets, previously explored by Wang et al. (2023), including the Enron Email dataset (Klimt & Yang, 2004), the AG News dataset, the Adult dataset (Ding et al., 2021), the PUMS ACS dataset (Ding et al., 2021), the OOD Style Transfer datasets (Wang et al., 2023), and the AdvGLUE++ datasets (Wang et al., 2023). For all datasets, we show average results from multiple runs and report ±1 standard deviation. While the main paper's plots focus on Enron, Adult, OOD Style Transfer, and AdvGlue++ datasets, supporting results on the other datasets can be found in the Appendix. Additional dataset details are in Appendix I.

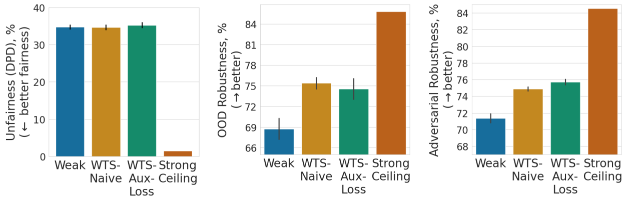
Large language models. We fine-tune models from the Pythia suite spanning five model sizes (14M, 70M, 410M, 1B, 6.9B parameters) (Biderman et al., 2023). The wide range of sizes allows us to systematically explore how model size impacts weak-to-strong trustworthiness.

3.1. Evaluating Weak-to-Strong Trustworthiness

We define weak-to-strong trustworthiness as a consistent trend – starting from the weak model at the lowest trustworthiness, increasing through the WTS-Naive and WTS-Aux-Loss models in the middle, and reaching its peak with the strong ceiling baseline. This monotonic trend indicates that trustworthiness is successfully generalized by the weak-to-strong model, despite only fine-tuning on the weak model' labels. The weak-to-strong model is able to recover some of the trustworthiness gap from the weak model to the strong ceiling model with ground truth data access. Sensitivity analyses are included in Appendix E and H.

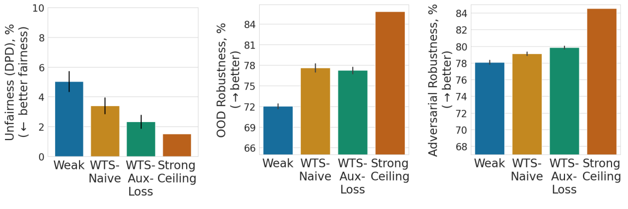
No TFT. The No TFT fine-tuning strategy does not achieve consistent weak-to-strong trustworthiness (Figure 2). For fairness experiments, the level of unfairness (demographic parity difference) remains constant at around 35% across all weak and weak-to-strong models. Similarly, we do not observe privacy generalization (Figure A1). We expected no consistent weak-to-strong trustworthiness for No TFT (standard weak-to-strong) as the strategy lacks regularization to explicitly enforce trustworthiness. Surprisingly, we observe a weak-to-strong trustworthiness trend for OOD and adversarial robustness. Despite the absence of regularization, the WTS-Naive and WTS-Aux-Loss models exhibited improved robustness compared to the weak models, suggesting that some trustworthiness properties may naturally generalize without explicit constraints.

Weak TFT. The Weak TFT fine-tuning strategy significantly improves the trustworthiness of weak models across all four



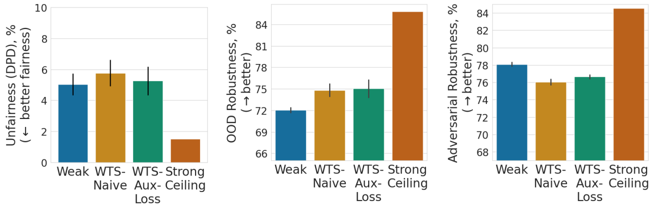
(a) Fairness (b) OOD Robustness (c) Adv. Robustness

Figure 2. **No TFT (standard weak-to-strong) is insufficient for trustworthiness generalization.** Weak-to-strong trustworthiness is inconsistent across properties, from no generalization of fairness to generalization of OOD and adversarial robustness.



(a) Fairness (b) OOD Robustness (c) Adv. Robustness

Figure 3. **Weak+WTS TFT achieves consistent WTS trustworthiness.** Weak+WTS TFT significantly improves trustworthiness generalization for fairness, OOD robustness, and adversarial robustness.



(a) Fairness (b) OOD Robustness (c) Adv. Robustness

Figure 4. **Weak TFT improves trustworthiness of weak models.** However, weak-to-strong trustworthiness is still inconsistent across properties, from no generalization of fairness and adversarial robustness to generalization of OOD robustness.

properties (Figures 4, A1). The effect of the additional regularization applied to weak models aligns with our expectations, as weak models are now explicitly regularized to enhance trustworthiness. Compared to No TFT, the weak models achieve lower unfairness (5% from 35%), increased OOD robustness (72% from 69%), increased adversarial robustness (78% from 71%), and lower privacy extraction (15% from 19%). Despite the trustworthy weak models, Weak TFT does not achieve consistent weak-to-strong trustworthiness. We only observe generalization for OOD robustness (Figure 4b). The weak-to-strong models are not more trustworthy than the weak models for fairness, adversarial robustness, and privacy (Figures 4a, 4c, A1).

Weak+WTS TFT. The Weak+WTS TFT fine-tuning strategy significantly improves the trustworthiness of weak-to-strong models across all four properties (Figures 3, A1). The effect of the additional regularization applied to weak and weak-to-strong models aligns with our expectations, as both models are now explicitly regularized to enhance trustworthiness. Compared to No TFT, the weak-to-strong models achieve lower unfairness (2% from 35%), increased OOD robustness (78% from 75%), increased adversarial robustness (80% from 75%), and lower privacy extraction (26% from 45%).

Unlike previous strategies, Weak+WTS TFT achieves consistent weak-to-strong trustworthiness for fairness, OOD robustness, and adversarial robustness (Figure 4a). The weak-to-strong models are significantly more trustworthy than the weak models, indicating successful trustworthy generalization through Weak+WTS TFT. For fairness and adversarial robustness, the WTS-Aux-Loss models generalize more effectively than the WTS-Naive models, suggesting that the auxiliary loss enables more weak-to-strong trustworthiness.

Through Weak+WTS TFT, weak-to-strong models are able to recover a significant portion of the trustworthiness gap between the weak model and strong ceiling baseline (strong models with access to ground truth labels). Despite their lack of ground truth labels, weak-to-strong models recover 88% of the fairness gap (2.8% out of 3.2%), 41% of the OOD robustness gap (5.5% out of 13.5%), and 31% of the adversarial robustness gap (2% out of 6.5%) (Figure 3).

However, weak-to-strong privacy trends are not observed. We discuss privacy in more detail in Appendix D.

4. Conclusion

This research is the first systematic study of weak-to-strong trustworthiness generalization (fairness, robustness, privacy) in language models. Standard weak-to-strong learning alone is insufficient. We introduce Weak TFT and Weak+WTS TFT fine-tuning, with Weak+WTS TFT recovering up to 88% of the trustworthiness gap while maintaining task performance. Fairness and robustness generalize well; privacy shows nuanced behavior. Our findings offer a path to build trustworthy AI with imperfect supervision, vital for safe deployment.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Bach, S. H., He, B., Ratner, A., and Ré, C. Learning the structure of generative models without labeled data. In *International Conference on Machine Learning*, pp. 273–282. PMLR, 2017.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, pp. 2397–2430. PMLR, 2023.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv:2303.12712*, 2023.

Burns, C., Izmailov, P., Kirchner, J. H., Baker, B., Gao, L., Aschenbrenner, L., Chen, Y., Ecoffet, A., Joglekar, M., Leike, J., et al. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *International Conference on Machine Learning (ICML)*. PMLR, 2024.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Chen, Z., Deng, Y., Yuan, H., Ji, K., and Gu, Q. Self-play fine-tuning converts weak language models to strong language models. ICML'24. JMLR.org, 2024.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pp. 265–284. Springer, 2006.

Garg, S. and Ramakrishnan, G. BAE: BERT-based adversarial examples for text classification. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6174–6181, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.498. URL https://aclanthology.org/2020.emnlp-main.498.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

Huang, P.-S., Zhang, H., Jiang, R., Stanforth, R., Welbl, J., Rae, J., Maini, V., Yogatama, D., and Kohli, P. Reducing sentiment bias in language models via counterfactual evaluation. In Cohn, T., He, Y., and Liu, Y. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 65–83, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.7. URL https://aclanthology.org/2020.findings-emnlp.7.

Jagielski, M., Nasr, M., Lee, K., Choquette-Choo, C. A., Carlini, N., and Tramer, F. Students parrot their teachers: Membership inference on model distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Jin, D., Jin, Z., Zhou, J. T., and Szolovits, P. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pp. 8018–8025. AAAI Press, 2020. doi: 10.1609/AAAI.V34I05.6311. URL https://doi.org/10.1609/aaai.v34i05.6311.

Klimt, B. and Yang, Y. The enron corpus: A new dataset for email classification research. In *European conference on machine learning*, pp. 217–226. Springer, 2004.

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., and Mullainathan, S. The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 275–284, 2017.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*, pp. 656–672. IEEE, 2019.

Leemann, T., Pawelczyk, M., and Kasneci, G. Gaussian membership inference privacy. *Advances in Neural Information Processing Systems*, 36, 2024.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019.

Li, L., Ma, R., Guo, Q., Xue, X., and Qiu, X. BERT-ATTACK: Adversarial attack against BERT using BERT. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6193–6202, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.500. URL https://aclanthology.org/2020.emnlp-main.500.

Lin, T. Focal loss for dense object detection. *arXiv preprint arXiv:1708.02002*, 2017.

Madry, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Mazzone, F., van den Heuvel, L., Huber, M., Verdecchia, C., Everts, M., Hahn, F., and Peter, A. Repeated knowledge distillation with confidence masking to mitigate membership inference attacks. In *Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security*, pp. 13–24, 2022.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., and Ré, C. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, volume 11, pp. 269. NIH Public Access, 2017.

Shejwalkar, V. and Houmansadr, A. Membership privacy for machine learning models through knowledge transfer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9549–9557, 2021.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6199.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Tang, X., Mahloujifar, S., Song, L., Shejwalkar, V., Nasr, M., Houmansadr, A., and Mittal, P. Mitigating membership inference attacks by {Self-Distillation} through a novel ensemble architecture. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1433–1450, 2022.

Wang, B., Pei, H., Pan, B., Chen, Q., Wang, S., and Li, B. T3: Tree-autoencoder constrained adversarial text generation for targeted attack. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6134–6150, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.495. URL https://aclanthology.org/2020.emnlp-main.495.

Wang, B., Xu, C., Wang, S., Gan, Z., Cheng, Y., Gao, J., Awadallah, A. H., and Li, B. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. In Vanschoren, J. and Yeung, S. (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*, 2021.

Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., et al. Decodingtrust: A comprehensive assessment of trustworthiness in gpt models. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2023.

Wei, J. and Zou, K. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*, 2019.

Yang, Y., Ma, Y., and Liu, P. Weak-to-strong reasoning. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 8350–8367, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.490. URL https://aclanthology.org/2024.findings-emnlp.490/.

Ye, M., Yin, Z., Zhang, T., Du, T., Chen, J., Wang, T., and Ma, F. Unit: a unified look at certified robust training against text adversarial perturbation. *Advances in Neural Information Processing Systems*, 36:22351–22368, 2023.

Yuan, L., Chen, Y., Cui, G., Gao, H., Zou, F., Cheng, X., Ji, H., Liu, Z., and Sun, M. Revisiting out-of-distribution robustness in nlp: Benchmarks, analysis, and llms evaluations. *Advances in Neural Information Processing Systems*, 36: 58478–58507, 2023.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics (AISTATS)*, pp. 962–970. PMLR, 2017.

Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., and Sun, M. Word-level textual adversarial attacking as combinatorial optimization. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 6066–6080. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.540. URL https://doi.org/10.18653/v1/2020.acl-main.540.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. Gender bias in coreference resolution: Evaluation and debiasing methods. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

Zheng, J., Cao, Y., and Wang, H. Resisting membership inference attacks through knowledge distillation. *Neurocomputing*, 452:114–126, 2021.

Zhu, C., Cheng, Y., Gan, Z., Sun, S., Goldstein, T., and Liu, J. Freelb: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*, 2019.

A. Related Work

This work is the first to study trustworthiness generalization from a weak supervisor to a weak-to-strong model. We discuss related works for the topics below.

Fairness. Unfair outcomes can arise in language models when they inadvertently encode biases present in the training data, leading to discriminatory practices against certain groups based on sensitive attributes like race, gender, or age (Bolukbasi et al., 2016). Recent efforts to improve fairness in LLMs include data pre-processing, post-processing, and adversarial training such as augmenting training data to balance gender representations (Zhao et al., 2018) and debiasing word embeddings (Huang et al., 2020). Our study is distinguished by its weak-to-strong setting and integration of fairness directly into the model's learning objective during fine-tuning.

Out-of-distribution robustness. OOD robustness describes a model's ability to perform well on inputs that differ from its training distribution. Various methods aim to enhance OOD robustness, including data augmentation techniques like adversarial perturbations (Madry, 2017; Lecuyer et al., 2019), EDA (Wei & Zou, 2019), as well as training modifications like label smoothing (Szegedy et al., 2016) and focal loss (Lin, 2017). However, recent research has shown that many of these methods do not reliably improve OOD robustness and may even degrade performance on in-distribution tasks; standard fine-tuning often remains a strong baseline (Yuan et al., 2023). In this work, we employ adversarial perturbation as a representative robustness technique, which has been explored in existing LLM robustness literature (Zhu et al., 2019; Ye et al., 2023). Unlike prior approaches, we focus on generalizing OOD robustness from weak models to larger strong models, both with and without the use of robustness-enhancing regularization.

Adversarial robustness. Machine learning model outputs can be changed by introducing minimal perturbations to a benign input, causing the model to malfunction (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018). Existing adversarial attack algorithms have been shown to degrade a large language model's performance on natural language processing tasks such as sentiment analysis, question answering, text classification, and entailment (Jin et al., 2020; Zang et al., 2020; Wang et al., 2020; Li et al., 2020; Garg & Ramakrishnan, 2020). Our work differs from these existing studies and is the first to examine if adversarial robustness can generalize from a weak model to a larger strong model fine-tuned on weak labels.

Privacy and model distillation. Prior research has explored knowledge distillation as a mechanism to mitigate privacy attacks. One example is the PATE framework (Papernot et al., 2016), where knowledge distillation is employed to reduce an ensemble of teacher models into a single model with provable privacy guarantees (Dwork et al., 2006). Other works have built on this idea, such as Zheng et al. (2021) and Tang et al. (2022), to similarly construct privacy-preserving model ensembles and consolidate them through distillation. Some research suggests that distillation alone can serve as an effective privacy defense (Shejwalkar & Houmansadr, 2021). Building on this, Mazzone et al. (2022) investigate the use of repeated distillation to protect against membership inference attacks. However, Jagielski et al. (2024) demonstrate through privacy attacks that distilled models without privacy guarantees can still leak sensitive information. In contrast to prior work, our research focuses on the privacy implications of weak-to-strong learning. This approach is the inverse of traditional model distillation. Nothing is known about the privacy risks when this process is reversed, making our work an important contribution to the field.

B. Preliminaries

First, we discuss how we adapt the weak-to-strong learning framework introduced by Burns et al. (2024). Following this, we examine regularization strategies to enhance trustworthiness properties such as fairness, robustness, and privacy.

Notation. We consider training datasets of the form $\{(x_i, y_i)\}_{i=1}^N$ where $y_i \in \mathcal{Y}$ is the ground-truth label. We denote a classifier $f_\theta : \mathcal{X} \to \mathcal{Y}$ parametrized by $\theta \in \mathbb{R}^d$, mapping inputs $x \in \mathcal{X}$, to labels $\mathcal{Y}$. We define the outputs of a fine-tuned smaller classifier $f_w(x)$ as *weak labels*, where $w \in \mathbb{R}^k$ denotes a lower-capacity parameterization than θ where $k \ll d$. Let $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ represent an appropriate loss function such as cross-entropy loss.

Weak-to-strong learning. In the weak-to-strong (WTS) framework, a pre-trained strong model accomplishes knowledge generalization by fine-tuning on a weak model's labels. This process incorporates an additional auxiliary loss, weighted by $\alpha \in [0, 1]$ to adjust the confidence in the strong model's predictions relative to the weak labels. This auxiliary loss encourages the strong model to make confident predictions, even when they diverge from the weak labels, potentially enhancing generalization. When $\alpha = 0$, we refer to the weak-to-strong learning as WTS-Naive, since we train on only the weak labels. When $\alpha > 0$, we refer to the weak-to-strong learning as WTS-Aux-Loss:

$$\ell_{\text{WTS}}^{\text{AUX}} = (1-\alpha)\ell\big(f_\theta(x), f_w(x;\lambda)\big) + \alpha\ell\big(f_\theta(x;\lambda), \hat{f}_{t,\theta}(x)\big). \tag{1}$$

We define our weak-to-strong loss function as a linear combination of the cross-entropy losses from the weak and strong models where $f_w(x;\lambda)$ denotes the weak model fine-tuned with trustworthiness regularization strength λ and $f_\theta(x)$ denotes the strong model. Further, $\hat{f}_{t,\theta}(x)$ represents the hardened strong model predictions according to threshold t set proportional to the dataset class weights. When $\lambda = 0$, we are in the standard weak-to-strong setting studied by Burns et al. (2024) (No TFT). In our proposed fine-tuning strategies, we apply trustworthiness regularization with $\lambda > 0$ to the weak model (Weak TFT, Weak+WTS TFT) and the weak-to-strong learning (Weak+WTS TFT).

Next, we describe our techniques for enhancing trustworthiness to obtain weak trustworthy models $f_w(\cdot;\lambda)$ through various regularization techniques.

Fairness. We enhance fairness through various definitions, one of which is Demographic Parity, which requires: $\mathbb{P}(f_w(x) = 1|a = 1) = \mathbb{P}(f_w(x) = 1|a = 0)$. We denote a as a protected attribute, like gender. To enforce this fairness constraint during fine-tuning, we use the following objective function from Zafar et al. (2017):

$$\min_w \frac{1}{N}\sum_{i=1}^{N} \ell(f_w(x_i), y_i) + \lambda_{\text{Fair}}(a_i - \bar{a})f_w(x_i), \tag{2}$$

where $\bar{a} = \frac{1}{N}\sum_{i=1}^N a_i$ is the base rate of the protected attribute. The first term incentivizes correct predictions while the second term acts as a fairness regularizer. Specifically, this term minimizes the covariance between the sensitive attribute a_i and the model outputs $f_w(x_i)$, incentivizing the model to satisfy demographic parity by becoming independent of the protected attribute a. Hyperparameter λ_{Fair} controls the trade-off between accuracy and fairness, where increasing λ_{Fair} emphasizes more fairness. We construct a similar objective function for Equalized Odds, another definition of fairness. Equalized Odds requires that true positive rates, $\mathbb{P}(f_w(x) = 1|y = 1)$, and false positive rates, $\mathbb{P}(f_w(x) = 1|y = 0)$, are equal across sensitive attributes.

Adversarial robustness. To enhance adversarial robustness, we introduce adversarially perturbed samples during the training process. As a result, the model learns to become invariant to small input perturbations and more robust to adversarial attacks. In this setting, the training dataset consists of triplets (x, x', y), where x is a clean input sample, x' is an adversarially manipulated version of x, and y is the ground truth label of x. The objective function combines the losses from both clean and adversarial samples:

$$\min_w \frac{1}{N}\sum_{i=1}^{N}(1-\lambda_{\text{Adv}})\ell(f_w(x_i), y_i) + \lambda_{\text{Adv}}\ell(f_w(x_i'), y_i), \tag{3}$$

where λ_{Adv} controls the trade-off between clean and adversarial losses. Increasing λ_{Adv} places greater emphasis on robustness to adversarial perturbations.

Out-of-distribution robustness. We use embedding perturbations to enhance out-of-distribution robustness, following approaches from Madry (2017); Lecuyer et al. (2019); Zhu et al. (2019). Specifically, we experiment with a setting that adds independent and identically distributed Gaussian noise to the word embeddings (Bowman et al., 2015; Li et al., 2019). Define $e(x) \in \mathbb{R}^d$ as the word embedding of input x, where d is the embedding dimension. We add Gaussian noise $z \sim \mathcal{N}(0, \lambda_{\text{OOD}} \cdot \mathrm{I}_d)$ drawn from a distribution with mean $\mathbf{0}$ and covariance matrix $\lambda_{\text{OOD}} \cdot \mathrm{I}_d$ to the word embedding, yielding a noisy embedding: $\tilde{e}(x; \lambda_{\text{OOD}}) = e(x) + z$. The noisy embedding is used to fine-tune the model. Denote $f_w(x; \lambda_{\text{OOD}}) = g_w(\tilde{e}(x; \lambda_{\text{OOD}}))$ as the output of the LLM parametrized by w. We use the following objective function:

$$\min_w \frac{1}{N} \sum_{i=1}^{N} \ell\big(y_i, f_w(x_i; \lambda_{\text{OOD}})\big)\big), \tag{4}$$

where λ_{OOD} controls the strength of the OOD regularizer. Decreasing $\lambda_{\text{OOD}} \to 0$ decreases OOD robustness.

Privacy. In (λ_P, δ)-differential privacy, the goal is to ensure that an algorithm's output $\mathcal{A}$ is nearly indistinguishable to whether any single data point is included in the dataset. Specifically, for any two datasets D_1 and D_2 that differ by only one element, the algorithm $\mathcal{A}$ satisfies (λ_P, δ)-differential privacy if:

$$\mathbb{P}(\mathcal{A}(D_1) \in S) \leq \exp(\lambda_P) \cdot \mathbb{P}(\mathcal{A}(D_2) \in S) + \delta, \tag{5}$$

for any possible output set S. Here, λ_P controls the privacy loss, with smaller values indicating stronger privacy guarantees, while δ allows for a small probability of the privacy guarantee being violated. To operationalize (λ_P, δ)-differential privacy, we use the most popular privacy algorithm called DP-SGD (Abadi et al., 2016), which is a variant of classical SGD with privacy guarantees. In summary, the algorithm consists of three fundamental steps: *gradient clipping* with clipping constant C (i.e. $\gamma = g(x_i, y_i) \cdot \max(1, C/\|g(x_i, y_i)\|)$ where $g(x_i, y_i) = \nabla_w \mathcal{L}(x_i, y_i)$ is the gradient of the loss function ℓ with respect to the model parameters), *aggregation* (i.e. $m = \frac{1}{n} \sum_{i=1}^{n} \gamma_i$), and *adding Gaussian noise* (i.e. $\tilde{m} = m + Y$ where $Y \sim \mathcal{N}(0, \tau^2 \mathrm{I})$ with variance parameter τ^2). By tuning the noise level τ^2, we ensure that the model satisfies the privacy guarantees specified by λ_P and δ.

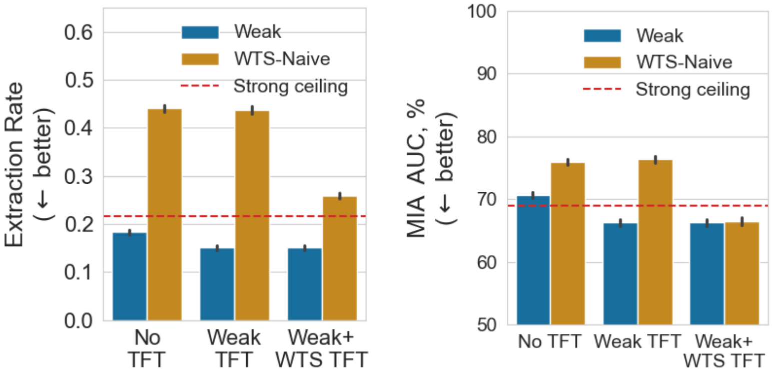
C. Metrics

Metrics. We evaluate a model's trustworthiness as follows:

- **Fairness**: We evaluate fairness using the demographic parity and equalized odds. For both definitions, lower values indicate better fairness, as they reflect minimal disparity in predictions between protected groups. We conduct comprehensive experiments using using Demographic Parity Difference (DPD), defined as DPD $= \mathbb{P}(f_\theta(x) = 1|a = 1) = \mathbb{P}(f_\theta(x) = 1|a = 0)$. Additional experiments on Equalized Odds Difference support the trends observed using Demographic Parity (Figure A14).
- **Robustness**: For robustness, we measure both OOD accuracy and adversarial accuracy, abbreviated as Robust Accuracy (RA), by evaluating the model's performance on OOD and adversarially perturbed test data. Specifically, we compute the RA $= \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \mathbb{I}[f_\theta(x'_i) = y_i]$, where x' represents either an OOD sample or an adversarially perturbed input, and $\mathbb{I}$ denotes the indicator function that equals 1 if the prediction is correct.
- **Privacy**: We evaluate privacy using targeted data extraction attacks and membership inference attacks (Shokri et al., 2017; Carlini et al., 2021). We conduct comprehensive experiments using extraction attacks, where given a prefix sequence and a generated response of k tokens, we compute the extraction rate by determining what fraction of the k-token continuation (suffix) matches the ground truth continuation of the sample. A higher extraction rate indicates a greater risk that the model memorizes and extracts private information. We also evaluate using standard membership inference attacks.

Strong ceiling baselines. For comparison, we establish baselines for both trustworthiness and task performance. We fine-tune strong models using ground truth labels with varying levels of trustworthiness regularization. We then select the model that achieves the best trade-off between task performance and trustworthiness. We provide an illustrative example of this selection in Figure A2. This value, referred to as the *strong ceiling*, represents the empirical upper bound of the strong model's capabilities for both task performance and trustworthiness.

D. Understanding weak-to-strong privacy



(a) Extraction Attack

(b) Membership Inference

Figure A1. **No weak-to-strong privacy.** While Weak+WTS TFT does not achieve privacy generalization, it still improves the privacy of weak-to-strong models compared to other strategies.

Privacy presents a unique situation, being the only property to not demonstrate consistent WTS trustworthiness under the Weak+WTS TFT strategy. However, note that the strong ceiling does not achieve better privacy than the weak model, which prevents any monotonic weak-to-strong privacy trend.

One reason for the distinction is that privacy is measured with respect to the underlying training dataset (Appendix I provides a more detailed discussion on how evaluating privacy differs from other properties). Larger models, all else being equal, tend to memorize more information, leading to a greater risk of private information leakage (Leemann et al., 2024). As a result, larger models are more susceptible to leak private data than smaller models. Therefore, we observe that privacy, measured by the extraction rate or membership inference attack success in Figure A1, degrades when learning a the weak model to a strong model. This is primarily due to weak-to-strong model privacy violations being measured for the larger model, which is more capable of memorizing information than the smaller one.

E. Sensitivity Analysis

Sensitivity to regularization strength (λ). We also investigated how varying the regularization strength in the trustworthiness objective functions affects weak-to-strong trustworthiness. For each property—fairness, OOD robustness, adversarial robustness, and privacy—we experimented with a range of λ values to observe their impact on trustworthiness generalization.

The Weak+WTS TFT strategy's ability to achieve consistent weak-to-strong trustworthiness, described in Section 3.1, maintained across a wide range of λ values. The plots of trustworthiness metrics against varying λ values demonstrate that the weak-to-strong models consistently generalized trustworthiness for fairness, OOD robustness, and adversarial robustness (Figure A4). The results suggest that the effectiveness of the Weak+WTS TFT strategy is robust to the choice of λ, provided it is within a reasonable range.

Compared to the Weak TFT strategy, the Weak+WTS TFT strategy demonstrates more significant trustworthiness generalization across various λ values (Figure A5). This behavior confirms our analysis in Section 3.1 that weak-to-strong trustworthiness is enhanced with increased regularization. Applying regularization to both the weak and weak-to-strong models enhances the trustworthiness generalization (from Figure A5 to Figure A4). Detailed sensitivity analyses are included in Appendix H

Sensitivity to model size. To assess the effect of model capacity on weak-to-strong, we experimented with multiple combinations of weak and strong model sizes. We analyzed experiments for five weak/strong configurations: Pythia 14M/410M, Pythia 14M/1B, Pythia 70M/410M, Pythia 70M/1B, and Pythia 14M/6.9B.

Our analysis reveals that the trustworthiness generalization trends observed in Section 3.1 hold consistently across a wide range of model sizes. No TFT remains unable to achieve consistent weak-to-strong trustworthiness, Weak TFT continues to improve weak model trustworthiness, and Weak+WTS TFT continues to consistently achieve weak-to-strong trustworthiness for fairness, OOD robustness, and adversarial robustness (Figures A9, A11, A10, A12, A13).

While increasing the strong model size led to some trustworthiness improvements, we saw significant improvement in weak-to-strong trustworthiness after increasing the weak model size (Figures A11, A10, A12 in Appendix). As weak models become more capable, their weak labels enable weak-to-strong models to generalize trustworthiness more effectively through Weak+WTF TFT.

F. Weak to Strong Learning Process

F.1. Training Objective for Weak+WTS TFT

In this section, we give a detailed description of the loss used for the third fine-tuning strategy presented in Section 2.1.

Fairness. To incorporate the fairness constraint into the fine-tuning process, we apply regularization twice yielding the following objective

$$\begin{aligned}\theta^* &\in \arg\min_\theta \mathcal{L}_{\text{Fair}}^{\text{WTS}}(\theta; \lambda_{\text{Fair}}^{\text{W}}, \lambda_{\text{Fair}}^{\text{WTS}}, \alpha, f_w) \\ &= \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell_{\text{WTS-AUX}}(x_i, f_\theta; \alpha, \lambda_{\text{Fair}}^{\text{W}}, f_w) + \lambda_{\text{Fair}}^{\text{WTS}} \cdot (a_i - \bar{a}) \cdot f_\theta(x_i),\end{aligned} \tag{6}$$

where $\alpha \in [0,1]$ is the auxiliary confidence loss weight and where $\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_i$ is the base rate of the protected attribute. The first term in equation 6 encourages the weak-to-strong model to make correct predictions while the second term acts as an additional fairness regularizer. The hyperparameter $\lambda_{\text{Fair}}^{\text{W}}$ corresponds to the regularization strength of the weak model while $\lambda_{\text{Fair}}^{\text{WTS}}$ controls the regularization strength for training in this stage.

Out-of-distribution robustness. The objective during fine-tuning is to minimize the following loss

$$\begin{aligned}\theta^* &\in \arg\min_\theta \mathcal{L}_{\text{OOD}}(\theta; \lambda_{\text{OOD}}^{\text{W}}, \lambda_{\text{OOD}}^{\text{WTS}}, \alpha, f_w) \\ &= \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell_{\text{WTS-AUX}}\big(x_i, f_\theta(x_i; \lambda_{\text{OOD}}^{\text{WTS}}); \alpha, \lambda_{\text{OOD}}^{\text{W}}, f_w\big),\end{aligned} \tag{7}$$

where $\alpha \in [0,1]$ is the auxiliary confidence loss weight. Further, $\lambda_{\text{OOD}}^{\text{W}}$ controls the regularization strength of the fixed weak classifier, while $\lambda_{\text{OOD}}^{\text{WTS}}$ controls the regularization strength of the transfer process. As $\lambda_{\text{OOD}}^{\text{WTS}} = 0$, we are back to our Weak TFT strategy, and as $\lambda_{\text{OOD}}^{\text{WTS}} = \lambda_{\text{OOD}}^{\text{W}} = 0$ the model is trained without any regularization, reverting to the No TFT strategy.

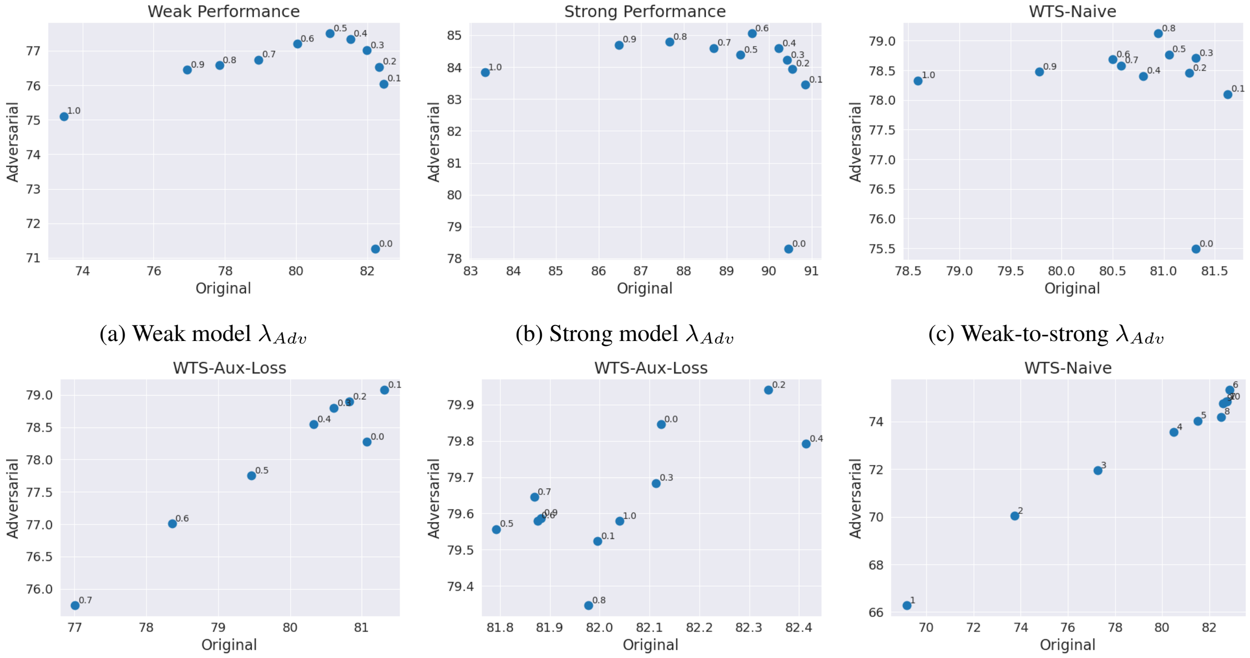
Adversarial Robustness. The training objective combines the losses from both clean and adversarial samples:

$$\begin{aligned}\theta^* &\in \arg\min_\theta \mathcal{L}_{\text{Adv}}(\theta; \lambda_{\text{Adv}}^{\text{W}}, \lambda_{\text{Adv}}^{\text{WTS}}, \alpha, f_w) \\ &= \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N}(1 - \lambda_{\text{Adv}}^{\text{WTS}})\, \ell_{\text{WTS-AUX}}(x_i, f_\theta; \alpha, \lambda_{\text{Adv}}^{\text{W}}, f_w) + \lambda_{\text{Adv}}^{\text{WTS}}\, \ell_{\text{WTS-AUX}}(x_i', f_\theta; \alpha, \lambda_{\text{Adv}}^{\text{W}}, f_w),\end{aligned} \tag{8}$$

where $\lambda_{\text{Adv}}^{\text{W}}$ controls the regularization strength of the fixed weak classifier, while $\lambda_{\text{Adv}}^{\text{WTS}}$ controls the regularization strength of the transfer process. As $\lambda_{\text{Adv}}^{\text{WTS}} = 0$, we are back to our Weak TFT strategy, and as $\lambda_{\text{Adv}}^{\text{WTS}} = \lambda_{\text{Adv}}^{\text{W}} = 0$ the model is trained without any regularization, reverting to the No TFT strategy.

F.2. Choosing the Hyperparameters Based on Trade-off Curves

In this section, we provide an illustrative example of how we selected the parameters for the strong baselines, using adversarial robustness as a case study. We plotted trade-off curves between the trustworthiness properties and task performance, selecting the parameter that corresponds to the optimal trade-off in the top right corner of the Figure A2. We set λ_{Adv} for the weak and strong model by independently fine-tuning them on training subset and evaluating on the test subset. We plot original task performance vs. adversarial performance for different values of λ_{Adv} and pick the value that offers the best trade-off between clean and adversarial accuracy. Figures A2a and A2b show that $\lambda_{Adv} = 0.3$ achieves the best combined accuracies on original and adversarial samples for both models. Fixing λ_{Adv} for the weak model to 0.3, we repeat the same analysis for the weak-to-strong model trained with the naive loss function. Figure A2c shows that $\lambda_{Adv} = 0.3$ offers the best trade-off for the weak-to-strong model as well. Fixing the λ_{Adv} parameter to 0.3 for the weak and weak-to-strong models, we vary the α parameter for the auxiliary loss function and plot in figure A2d. We observe that $\alpha = 0.1$ achieves the highest accuracy on both original and adversarial samples. We perform similar analyses for the warm-up period, α, and the number of fine-tuning epochs in Figures A2e and A2f. We select the values 0.2 and 6, respectively, for these training parameters.



(a) Weak model λ_{Adv}

(b) Strong model λ_{Adv}

(c) Weak-to-strong λ_{Adv}

(d) Weak-to-strong α

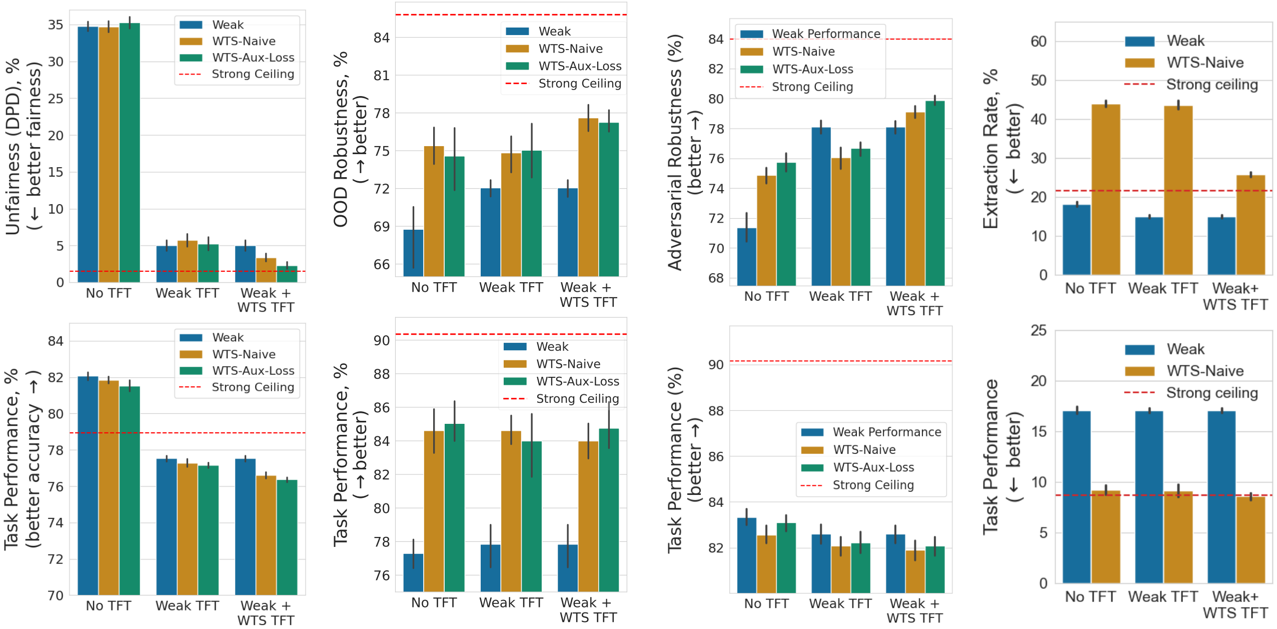
(e) Warm-up period.

(f) Number of epochs for WTS.

Figure A2. Trade-off between original and adversarial accuracy for different training parameters.

Similarly, for OOD robustness, we set the standard deviation of the Gaussian Noise to $2e - 3$ for both the weak model (Pythia 14M) and the strong model (Pythia 410M). This value was chosen as it allows both models to achieve a balanced trade-off between OOD robustness and task performance. With the noise standard deviation fixed, we conduct trade-off experiments by separately adjusting the maximum α value for auxiliary loss, the warm-up period, and the number of training epochs. For optimal balance between OOD robustness and task performance, these parameters are set to 0.25, 0.2, and 1, respectively.

G. Comprehensive Plots Across Strategies



(a) Fairness

(b) OOD Robustness

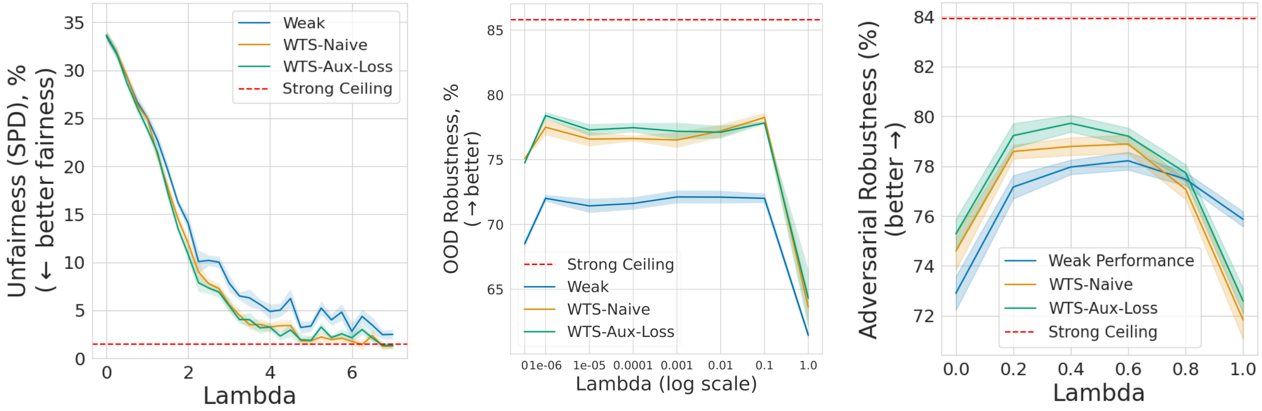
(c) Adv. Robustness

(d) Privacy

Figure A3. **Weak-to-strong trustworthiness for Pythia 14M/410M models**. Trustworthiness properties and task performance for our four properties: Fairness, OOD Robustness, Adversarial Robustness, and Privacy. Note that lower values are better for the top plot in Figure A3a as the y-axis is Unfairness (DPD). Similarly, lower values are better for the top plot in Figure A3d as the the y-axis is Extraction Rate. Results for WTS-Aux-Loss for privacy are omitted since it was the only task involving free data generation, making the auxiliary loss function inapplicable.

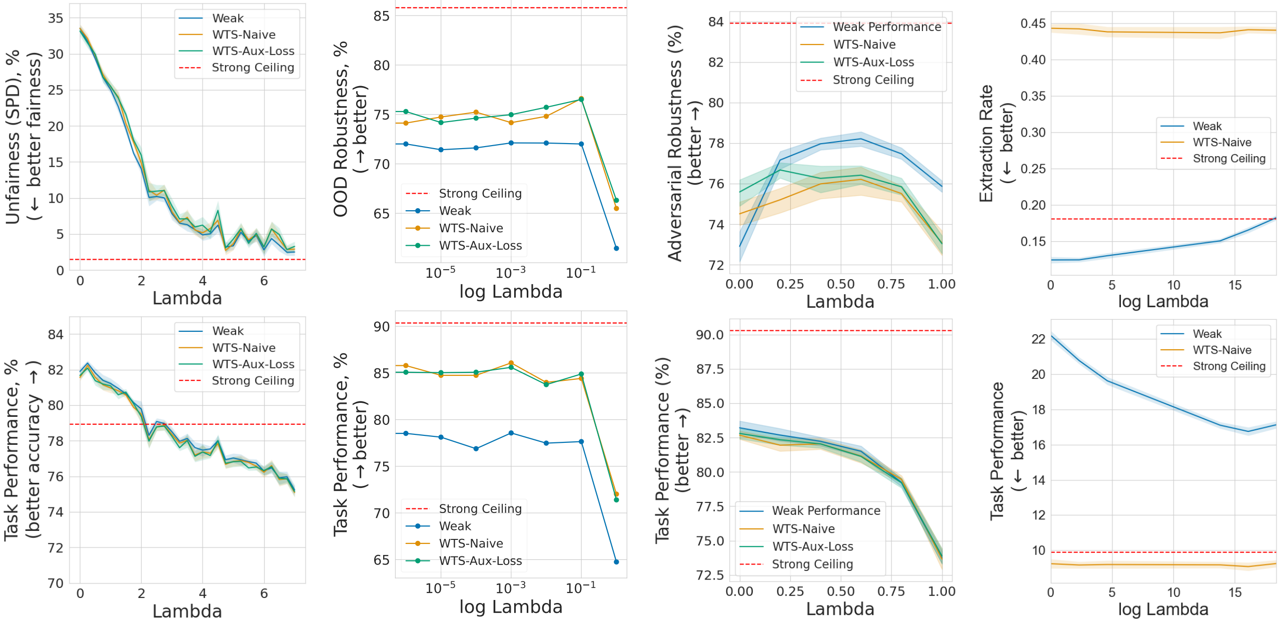
H. Detailed Sensitivity Analysis

In this section, we study the sensitivity of the weak-to-strong trustworthiness fine-tuning to key training parameters like λ and α.



(a) Fairness (b) OOD Robustness (c) Adv. Robustness

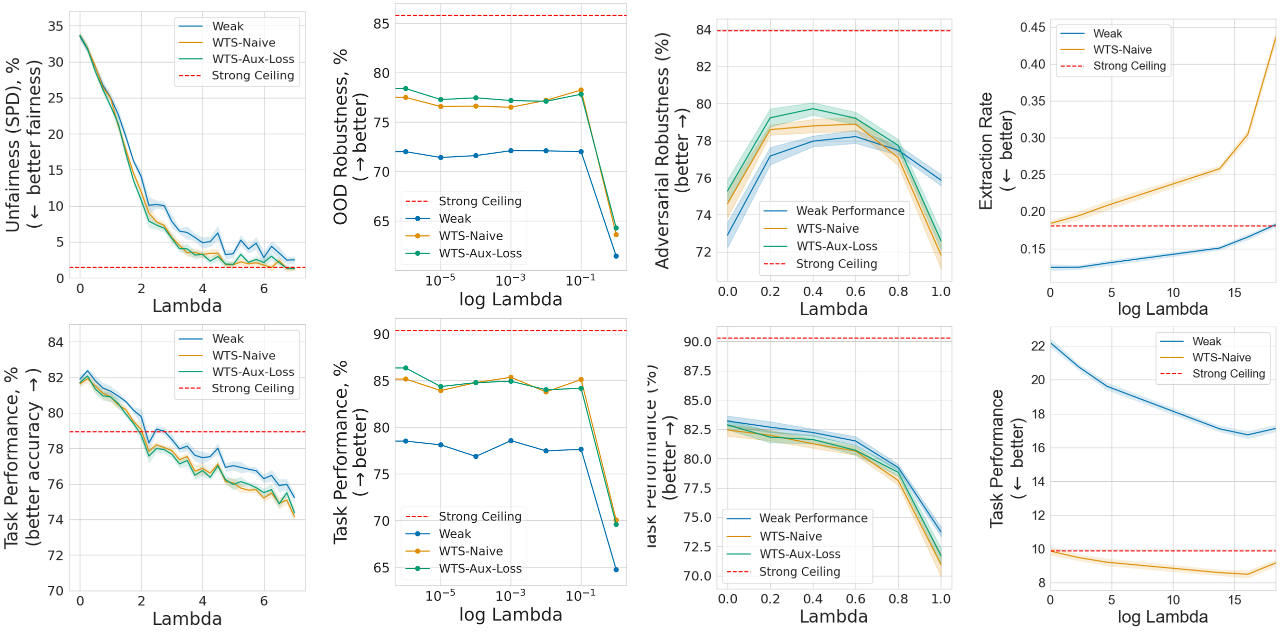
Figure A4. **Weak+WTS TFT improves trustworthiness generalization across regularization strengths** (λ)**.** Sensitivity analysis demonstrates the consistency of Weak+WTS TFT strategy to generalize fairness, OOD robustness, and adversarial robustness across a wide range of λ values for weak-to-strong trustworthiness regularization.



(a) Fairness (b) OOD Robustness (c) Adv. Robustness (d) Privacy

Figure A5. **Full Plot for Varying Lambda for Weak TFT**. Results for WTS-Aux-Loss for privacy are omitted since it was the only task involving free data generation, making the auxiliary loss function inapplicable.

Impact of Auxiliary Loss Weighting (α_{max}). The auxiliary loss weighting parameter α_{max} (maximum alpha) plays a crucial role in balancing the adherence to the weak model's outputs and the strong model's confidence in its predictions. Higher values of α_{max} place more emphasis on the strong model's own predictions rather than closely following the weak model's outputs. We examine the effect of varying α_{max} from 0 to 1 on the performance of the weak-to-strong models. Our experiments showed a degradation of performance with increasing α_{max}. As α_{max} increases from 0 to 1, the performance



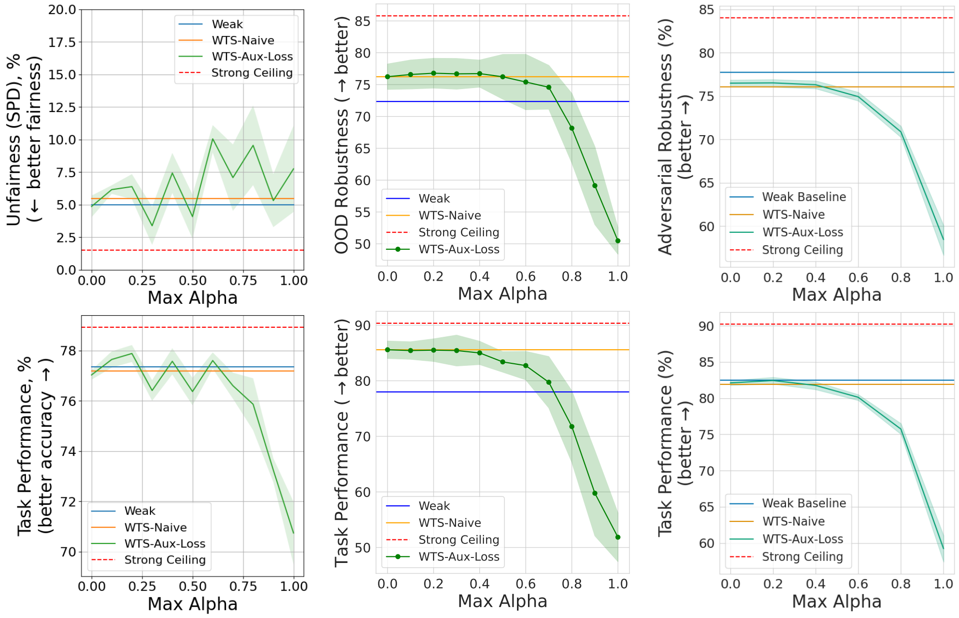
(a) Fairness (b) OOD Robustness (c) Adv. Robustness (d) Privacy

Figure A6. **Full Plot for Varying Lambda for Weak+WTS TFT**. Results for WTS-Aux-Loss for privacy are omitted since it was the only task involving free data generation, making the auxiliary loss function inapplicable.

of the weak-to-strong models trained with the auxiliary loss (WTS-Aux-Loss) tends to worsen. Therefore, selecting an appropriate value of α_{max} is essential to maintain a balance between leveraging the weak model's trustworthiness and allowing the strong model to develop its capabilities. Our results suggest that lower α_{max} values are preferable for effective weak-to-strong trustworthiness transfer. For our models, we chose α_{max} values from 0.1 to 0.4.

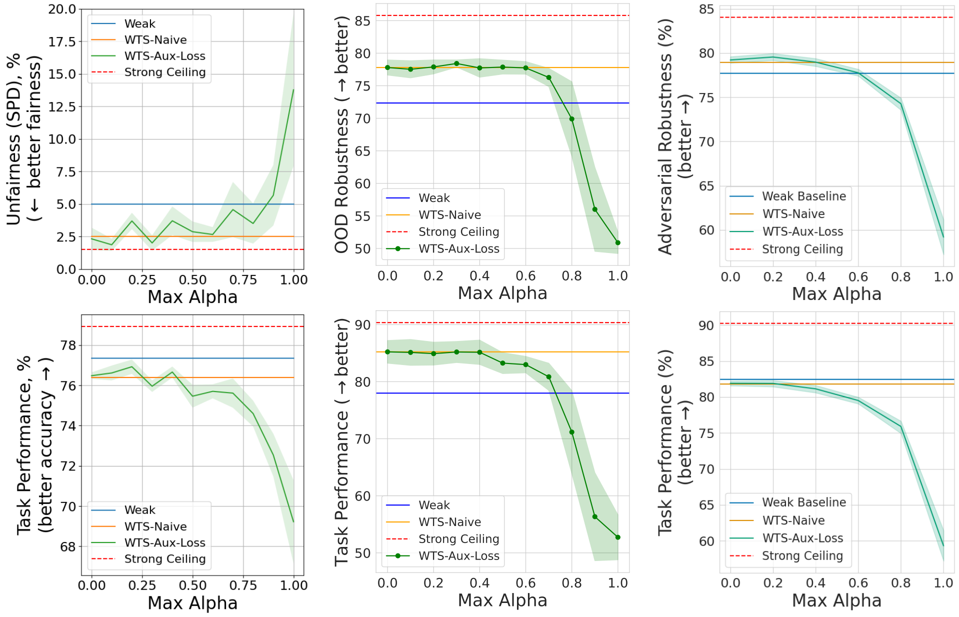
Impact of Larger Models (6.9B). We show that WTS trustworthiness trends are consistent when scaling up the strong model. As referenced in Section E, Figures A9 to A12, show four different weak/strong model size configurations (14M/410M, 70M/410M, 14M/1B, 70M/1B) with consistent property-specific weak-to-strong trustworthiness trends holding across model sizes. We also extended our model size sensitivity analysis to include Pythia 6.9B as the strong model for fairness, OOD robustness, and adversarial robustness. The 6.9B model required multiple GPUs to train, and DP-SGD currently does not support multi-GPU computations, so we did not provide 6.9B results for privacy. Figure A13 displays the results and demonstrates similar weak-to-strong trustworthiness trends as the previous model configurations. While weak-to-strong trustworthiness is inconsistent at the Weak TFT strategy, we see consistent weak-to-strong trustworthiness at the Weak+WTS TFT strategy.

Impact of Additional Metrics. We include multiple trustworthiness definitions to further support the weak-to-strong trustworthiness trends we observed. In Figure A14, we examine an additional fairness metric: equalized odds (true positive rate). The consistent weak-to-strong fairness trend is maintained across both demographic parity and equalized odds. In Figure A15, we examine an additional privacy metric: membership inference attack. We continue to see no weak-to-strong privacy across both extraction and membership inference attacks.



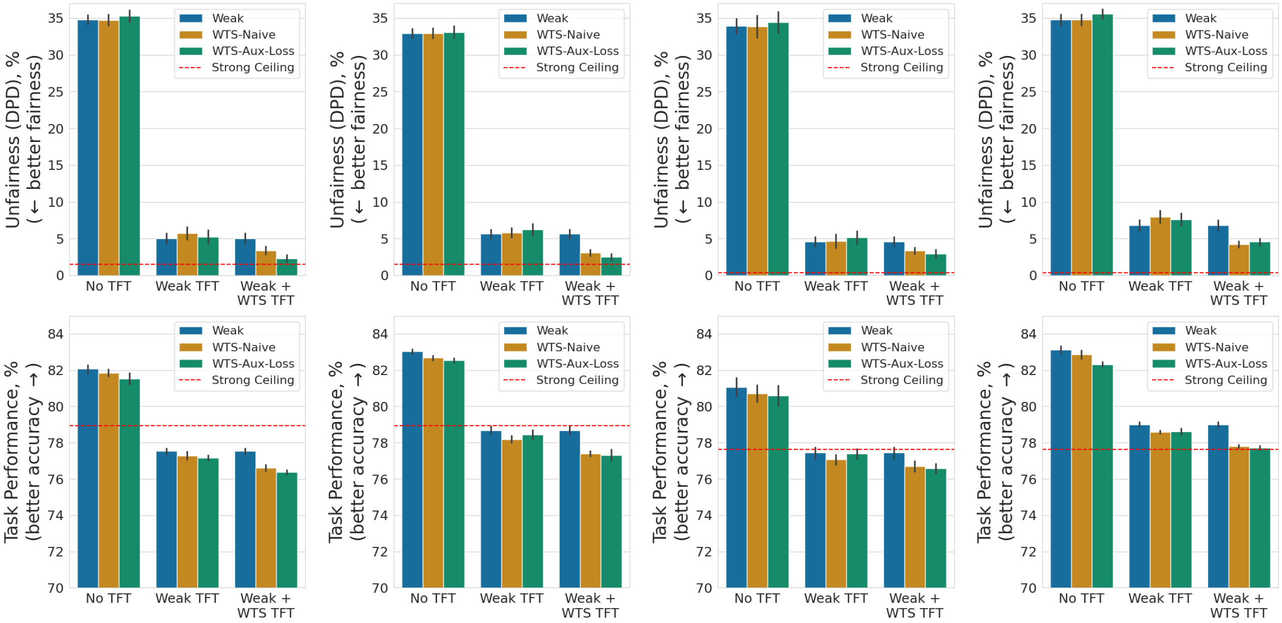
(a) Fairness (b) OOD Robustness (c) Adv. Robustness

Figure A7. **Varying Max Alpha for Weak TFT.** Results on privacy are omitted since it was the only task involving free data generation, making the auxiliary loss function inapplicable.



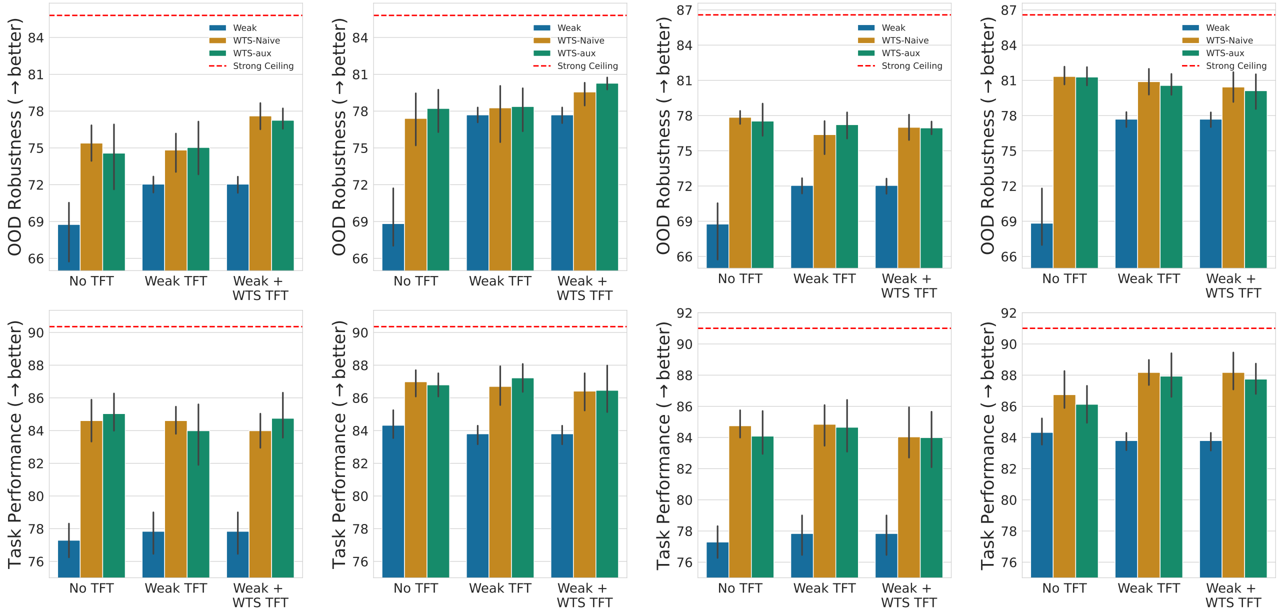
(a) Fairness (b) OOD Robustness (c) Adv. Robustness

Figure A8. **Varying Max Alpha for Weak+WTS TFT**. Results for WTS-Aux-Loss for privacy are omitted since it was the only task involving free data generation, making the auxiliary loss function inapplicable.



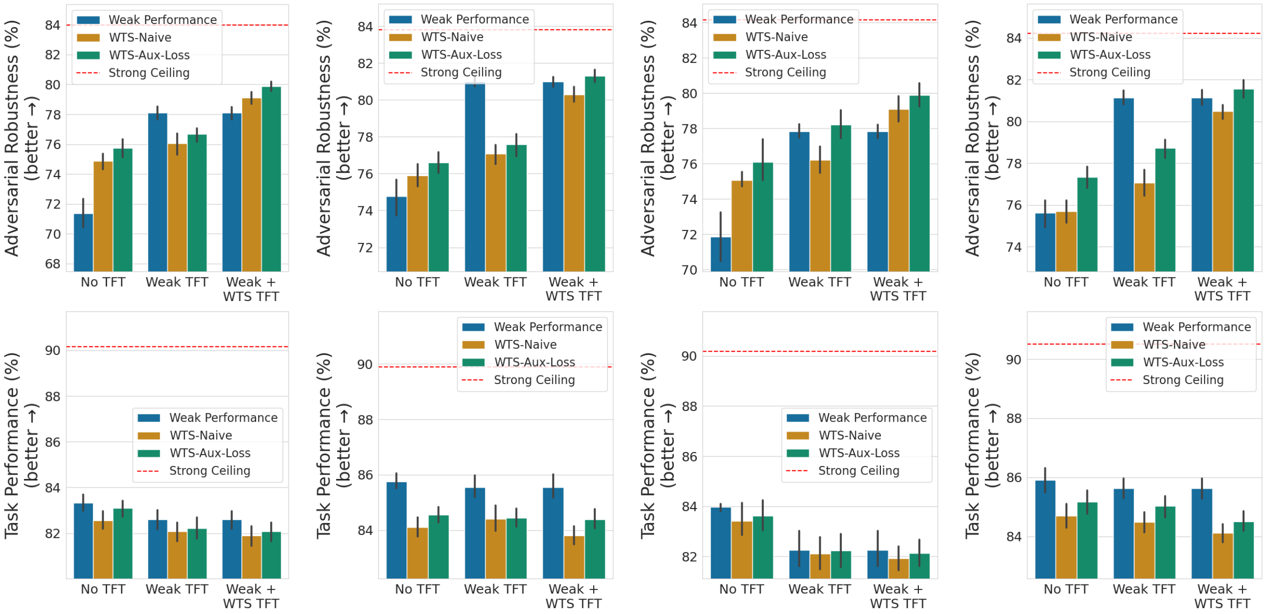
(a) 14M - 410M (b) 70M - 410M (c) 14M - 1B (d) 70M - 1B

Figure A9. **Varying model size for fairness.** Weak-to-strong trustworthiness trends hold for fairness cross multiple model size configurations.



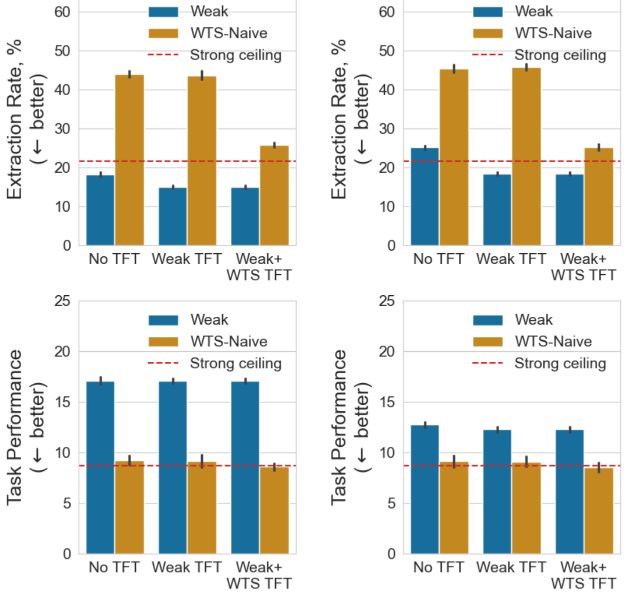
(a) 14M - 410M (b) 70M - 410M (c) 14M - 1B (d) 70M - 1B

Figure A10. **Varying model size for OOD Robustness.** Weak-to-strong trustworthiness trends hold for OOD robustness cross multiple model size configurations.



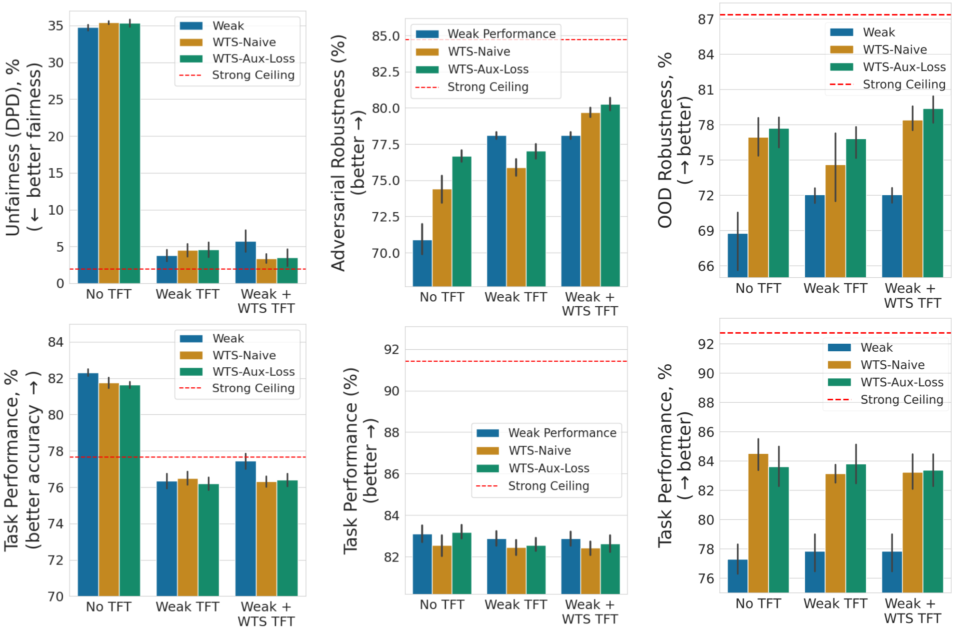
(a) 14M - 410M (b) 70M - 410M (c) 14M - 1B (d) 70M - 1B

Figure A11. **Varying model size for adversarial robustness.** Weak-to-strong trustworthiness trends hold for adversarial robustness cross multiple model size configurations.



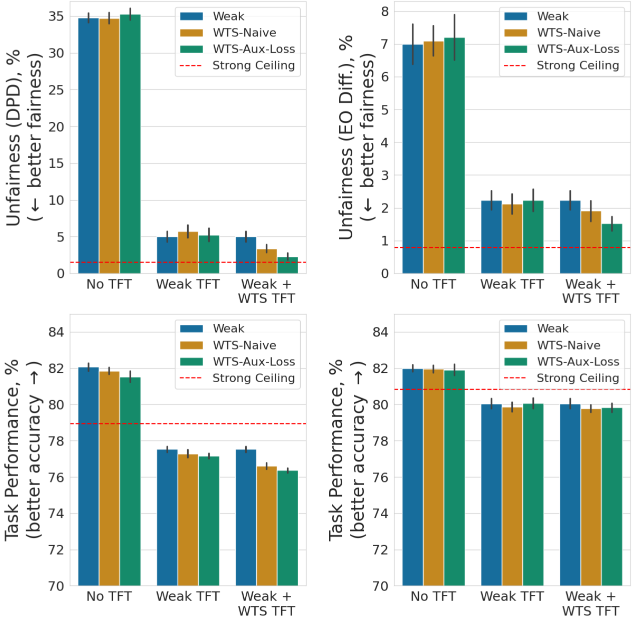
(a) 14M - 410M (b) 70M - 410M

Figure A12. **Varying model size for privacy.** No weak-to-strong trustworthiness trends hold for privacy cross multiple model size configurations. Due to memory limitations of training models with DP-SGD we did not train the 1B or 6.9B models.



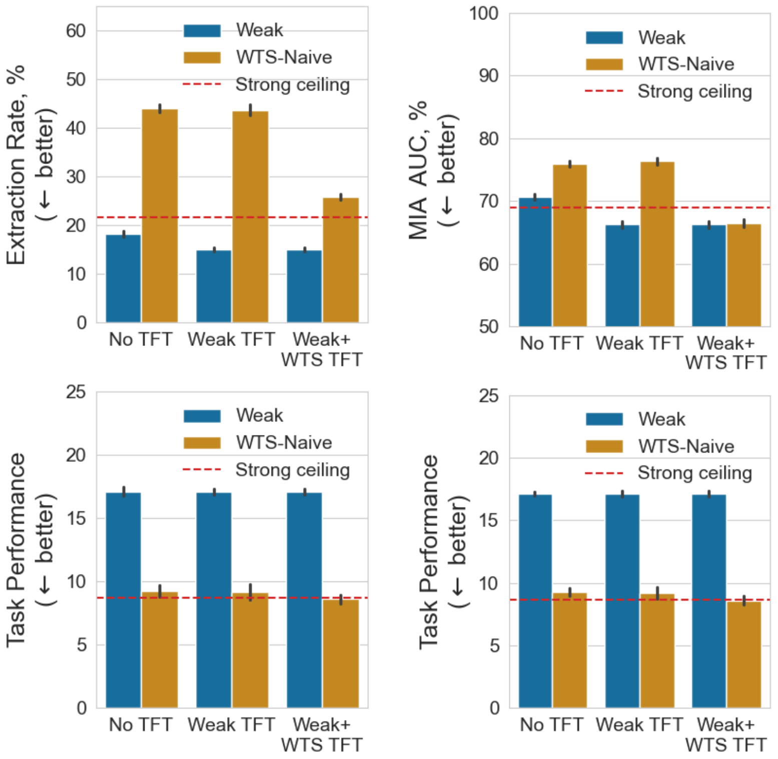
(a) Fairness (b) Adv. Robustness (c) OOD Robustness

Figure A13. **Model Size Analysis on Pythia 6.9B**. Results for model size sensitivity with Pythia 14M as the weak model and Pythia 6.9B as the strong model for fairness, adversarial robustness, and OOD robustness properties. We see that the WTS trends we identified earlier are maintained for the larger strong model.



(a) Demographic Parity (b) Equalized Odds

Figure A14. **Sensitivity to Fairness Metrics**. Side-by-side results for two fairness metrics: Demographic Parity and Equalized Odds (True Positive Rate). The weak-to-strong trustworthiness trends are maintained across both metrics.



(a) Extraction Attack

(b) Membership Inference Attack

Figure A15. **Sensitivity to Privacy Metrics**. Side-by-side results for two privacy metrics: Extraction Attack and Membership Inference Attack. While Weak+WTS TFT does not achieve weak-to-strong trustworthiness, it still leads to simultaneous improvement of privacy and performance for weak-to-strong models.

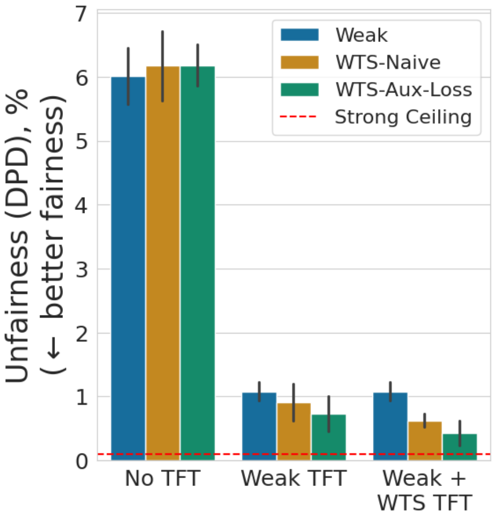


Figure A16. Additional Fairness Dataset: ACS PUMS Employment

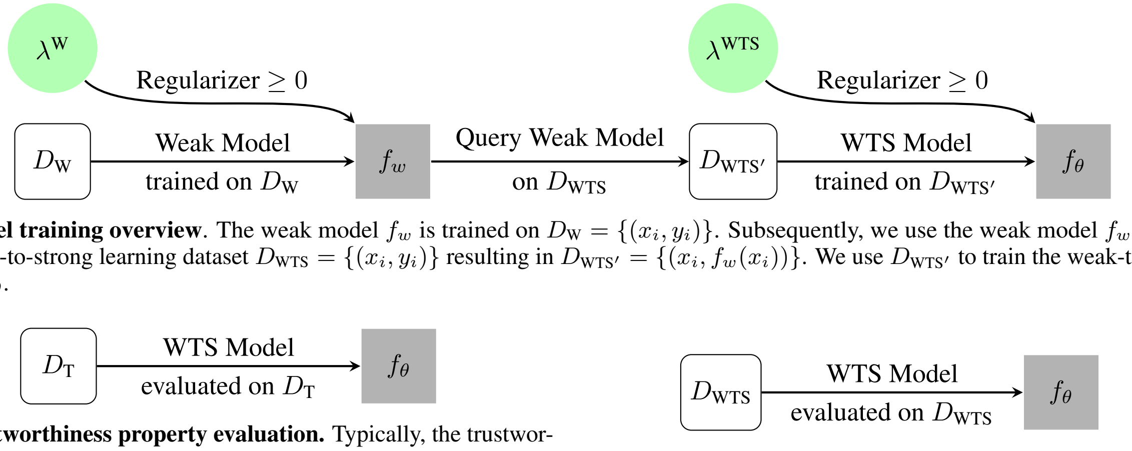
Table 1. Additional Privacy Dataset: AG News

Strategy	Model	Extraction Rate
No TFT	Weak	0.059
No TFT	WTS-Naive	0.081
Weak TFT	Weak	0.050
Weak TFT	WTS-Naive	0.102
Weak+WTS TFT	Weak	0.051
Weak+WTS TFT	WTS-Naive	0.092

I. Dataset and Evaluation Details

I.1. Dataset Details

- **Adult:** The Adult dataset is derived from the 1994 U.S. Census database and contains 48,842 instances with 14 attributes. The task is to classify whether an individual's income exceeds \$50K (USD) per year. We selected the "sex" feature as the sensitive attribute to evaluate fairness-related properties. Extraction was done by Barry Becker from the 1994 Census database. Adult dataset has a CC-BY-4.0 license, which we abide by.
- **ACS PUMS Employment:** The Census Bureau's American Community Survey (ACS) Public Use Microdata Sample (PUMS) includes information about U.S. residents' age, sex, race, education, employment, and other demographics. The task is to classify whether an individual is employed. ACS PUMS dataset has a CC-BY-4.0 license, which we abide by.
- **OOD Style Transfer**: The OOD Style Transfer dataset is based on the SST-2 sentiment classification dataset but incorporates a variety of text and style transformations. The transformations (e.g., shifts in language style, vocabulary, syntax, and tone) are applied at both the word and sentence level while preserving the original meaning (Wang et al., 2023). The task is to correctly classify the sentiment of inputs. OOD Style Transfer dataset has a CC-BY-SA-4.0 license, which we abide by.
- **AdvGLUE++**: AdvGLUE++ is a collection of six datasets contain clean and adversarial input samples for six NLP tasks: Sentiment analysis (SST-2), duplicate question detection (QQP), multi-genre natural language inference (MNLI, MNLI-mm), recognizing textual entailment (RTE), and question answering (QNLI) (Wang et al., 2023). It contains around 2K to 15K samples for each of the six tasks. We randomly sample up to 10K samples for each task and aggregate the performance by averaging over these six tasks. AdvGLUE++ datasets have a CC-BY-SA-4.0 license, which we abide by.
- **Enron Emails**: The Enron Emails dataset contains over 600K emails generated by employees of the Enron Corporation (Klimt & Yang, 2004). it includes sensitive personal information, such as email addresses, phone numbers, credit card numbers, and Social Security Numbers, which could be memorized and extracted by language models. For fine-tuning, we randomly subsampled 10K data points. Enron Emails dataset has a Apache License 2.0, which we abide by.
- **AG News**: The AG News dataset consists of 120,000 training samples and 7,600 test samples of news articles categorized into 4 classes: World, Sports, Business, and Science/Technology. Each sample contains a title and description extracted from AG's news corpus, with balanced distribution across classes. AG News data was made by Antonio Gulli (`http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html`) and permitted for non-commercial use, which we abide by.



(a) **Model training overview**. The weak model f_w is trained on $D_{\mathrm{W}} = \{(x_i, y_i)\}$. Subsequently, we use the weak model f_w to label the weak-to-strong learning dataset $D_{\mathrm{WTS}} = \{(x_i, y_i)\}$ resulting in $D_{\mathrm{WTS'}} = \{(x_i, f_w(x_i))\}$. We use $D_{\mathrm{WTS'}}$ to train the weak-to-strong model f_θ.

(b) **Trustworthiness property evaluation.** Typically, the trustworthiness properties for the WTS model are evaluated on a separate test set D_{T}.

(c) **Privacy Leakage Evaluation.** The privacy leakage for the WTS model is evaluated using the ground truth train set D_{WTS}.

Figure A17. **Data usage during training and evaluation.** In Figure A17a, we describe which data is used to train the weak and the weak-to-strong models, while Figures A17b and A17c describe which data is used for evaluation.

I.2. Data Usage During Training and Evaluation

Figure A17 describes which data is used for training the weak and the weak-to-strong models as well as for evaluating of the weak-to-strong model.

Data used to train the WTS model. The weak model f_w is trained on the labeled dataset $D_{\text{W}} = \{(x_i, y_i)\}$. Once trained, we use the weak model f_w to label the weak-to-strong learning dataset $D_{\text{WTS}} = \{(x_i, y_i)\}$ resulting in $D_{\text{WTS}'} = \{(x_i, f_w(x_i))\}$. We use $D_{\text{WTS}'}$ to train the weak-to-strong model f_θ. Notably, there is no overlap between D_{WTS} and D_{W}.

Trustworthiness Evaluation. We evaluate the trustworthiness properties adversarial robustness, OOD robustness as well as Demographic Parity and Equalized Odds for all models (weak model, weak-to-strong model, and strong ceiling) on the same held out test set for the respective problem. For privacy, we evaluate the trustworthiness properties of the weak and the weak-to-strong model on their training set D_{W} while the privacy leakage for the WTS model is evaluated on D_{WTS}. For privacy considerations, we evaluated the trustworthiness properties of models on their training set D_{W}, while the privacy leakage for the WTS model is assessed on D_{WTS}.

I.3. Additional Adversarial Robustness Dataset Details

We create training, holdout and test subsets of the AdvGLUE++ dataset using 40%, 40% and 20% of samples, respectively, from each task in the dataset. We use the training subset to fine-tune our models to be adversarially robust. We use the holdout subset to generate labels from the weak model to be used in the weak-to-strong learning process. To evaluate the clean and adversarial accuracy of our models, we evaluate them on a test subset of the AdvGLUE++ dataset and average the performance across the six NLP tasks in this dataset.

In particular, to evaluate weak-to-strong trends in adversarial robustness, we use the AdvGLUE++ dataset (Wang et al., 2023), an extension of the AdvGLUE dataset (Wang et al., 2021). AdvGLUE++ is a comprehensive benchmark designed to test adversarial robustness across multiple natural language processing (NLP) tasks and adversarial attack algorithms. This dataset includes adversarial examples for six widely used NLP tasks, each representing a distinct domain or linguistic challenge. The Stanford Sentiment Treebank (SST-2) task involves sentiment analysis, requiring the classification of sentences as having a positive or negative sentiment. The Quora Question Pairs (QQP) task identifies whether two questions convey the same meaning. The Multi-Genre Natural Language Inference (MNLI) task requires reasoning about entailment, contradiction, or neutrality between pairs of sentences. It includes a mismatched variant, MNLI-mm, where validation and test data originate from out-of-domain sources, increasing the challenge of generalization. The Question-answering NLI (QNLI) task is framed as an entailment problem between a question and an answer candidate. The Recognizing Textual Entailment (RTE) is a binary entailment task that aims to determine whether the meaning of one text can be inferred from another.

Adversarial examples in AdvGLUE++ are generated using a variety of attack algorithms, each representing a distinct perturbation strategy. TextBugger introduces typo-based perturbations that minimally alter characters while preserving the utility of benign text. TextFooler generates embedding similarity-based perturbations by substituting words with contextually plausible alternatives. BERT-ATTACK leverages BERT's language modeling capabilities to create context-aware adversarial samples. SememePSO relies on semantic representations and combinatorial optimization to generate knowledge-guided perturbations. SemAttack employs semantic optimization-based techniques by manipulating various semantic spaces to produce natural-looking adversarial texts.

The experimental results for adversarial robustness are presented as aggregated accuracy values across all six tasks and five attack algorithms. This approach enables us to evaluate the weak-to-strong trends in a comprehensive and robust manner. The results show that our findings are consistent across a wide range of NLP tasks and adversarial attacks, indicating that they are not influenced by the specific characteristics of any single setting.

I.4. Additional OOD Dataset Details

We use the same OOD data created by Wang et al. (2023). For ID data, we use the original SST-2 dataset but exclude the samples that are source samples for creating the OOD data. We split the ID data into training, validation, and heldout subsets. Specifically, 50% of the ID data is allocated for training and validation, where 95% of that portion is used for training and the remaining 5% is for validation. The other half represents the held-out data that is used for generating labels from the weak model for weak-to-strong fine-tuning. For evaluation, we use the in-distribution validation samples to measure ID performance and the OOD test samples to obtain OOD performance.

J. Overview Table

Table 2. **Overview table.** Trustworthiness properties, their corresponding metrics, datasets used, and tasks performed on each dataset.

Property	**Metrics**	**Datasets**	**Tasks**
Fairness	• Demographic Parity • Equalized Odds	• Adult • ACS PUMS	• Income classification with "sex" as the sensitive attribute
OOD Robustness	• Robust Accuracy (RA) on OOD test data	• OOD Style Transfer: a collection of 10 datasets with different text and style transformations (based on the SST-2 dataset)	• Sentiment classification on 10 different text and style transformations
Adversarial Robustness	• Robust Accuracy (RA) on adversarial test data	• AdvGLUE++: a collection of six datasets 1. SST-2 2. QQP 3. MNLI 4. MNLI-mm 5. RTE 6. QNLI	• Sentiment analysis • Duplicate question detection • Multi-genre natural language inference • Recognizing textual entailment • Question answering
Privacy	• Extraction attack • Membership inference attack	• Enron Emails • AG-News	• Sensitive data leakage detection

K. Experimental Details

Models: We use the Pythia models from EleutherAI (Biderman et al., 2023). They have a Apache License 2.0, which we abide by.

Statistical Significance: We report 1 standard deviations for our experiments over multiple trials (10 for fairness, 15 for OOD robustness, 15 for adversarial robustness, 3 for privacy).

Compute: Each experiment was run on 1 NVIDIA A100 80GB GPU on an internal cluster.

Table 3. Hyperparameters

Hyperparameter	**Fairness**	**OOD Robustness**	**Adversarial Robustness**	**Privacy**
Epochs	5	1	6	1
Learning rate	5e-5	1e-5	1e-5	5e-5
Optimizer	AdamW	AdamW	AdamW	Adam
Lambda	4.25	0.002	0.3	1e6
Alpha	0.3	0.2	0.1	N/A

L. Limitations

While this study investigated models up to 6.9 billion parameters, further exploration with even larger models was constrained by available computational resources. Future work with access to greater computational capacity could extend these findings to assess the weak-to-strong trustworthiness to the frontier of model sizes.