Experiment Setting Details

Data We use the AMAZON-WILDS dataset [4] for sentiment analysis. The dataset consists of Amazon reviews of many different categories and their ratings. We select HOME-AND-KITCHEN category for training and 10 other categories for evaluation. In domain adaptation, the HOME-AND-KITCHEN category would be referred to as source domain, while other categories are target domains. Following previous work in domain adaptation [3], we convert ratings to binary labels (positive when ¿3, negative when ¡3) and sample a balanced dataset for each category (domain).

Model To obtain a series of models with different accuracies, we finetuned a pre-trained BERT model [2, 5] 100 times on different random seeds.

Capability	**Keywords**
negation	not, n't
negation *(variant)*	no, never, neither, nobody, none, nor, nothing
shifter	refuse, reject, deny, doubt, abandon, miss, question, abort, stop
modality	would have, could have, should have
comparative	better, worse, than
mixed	but, however, though, although, despite, even if, rather than, except that
reducer	kind of, all that, less, a little, somewhat, still
amplifier	really, very, super, so, incredibly, extremely, at all, whatsoever, much

Table 1: Capabilities and their instantitation keywords for sentiment analysis, selected based on existing work [1]. We slice the validation data on keywords to instantiate these capabilities.

Method Our goal is to observe the correlations between models' accuracies across source domain and target domains, as well as how extra information (e.g., accuracies on capability test suites) would affect the correlations.

We first use a linear model to compute the correlations of accuracies between source domain and target domains, without any extra variables. We then introduce three sets of variables to the linear model:

1. *Capability test suites' accuracies*: we first selected 8 capabilities for sentiment analysis based on an existing study [1]. We instantiated these capabilities by slicing the source domain dataset on their corresponding keywords (see Tab. 1). The final capabilities we used are *shifter*, *modality*, and *comparative*.

2. *Random subsets' accuracies*: we selected three random subsets from source dataset. These subsets are of the same size of three capability test suites. We computed models' accuracies on the random subsets.

3. *Noisy accuracies*: we added random Gaussian noise to models' validation accuracies.

For the last two settings, we repeat the process on 100 different random seeds and average their results.

We then fit the linear model with these extra variables for each setting. We looked at adjusted R^2 to see whether the model has a better fit (i.e., whether these extra variables help predict out-

of-distribution scores). [1] We also performed ANOVA testing to see whether the improvement is statistically significant.

To understand the relation between capabilities' predictiveness and distribution distance, we followed the method from previous work [3] to compute a proxy $\mathcal{A}$-distance between different domains.

References

[1] Barnes, J., Øvrelid, L., and Velldal, E. Sentiment analysis is not solved! assessing and probing sentiment classification. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (Florence, Italy, August 2019), Association for Computational Linguistics, pp. 12–23.

[2] Bhargava, P., Drozd, A., and Rogers, A. Generalization in nli: Ways (not) to go beyond simple heuristics, 2021.

[3] Blitzer, J., Dredze, M., and Pereira, F. Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics* (Prague, Czech Republic, June 2007), Association for Computational Linguistics, pp. 440–447.

[4] Koh, P. W., et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning* (18–24 Jul 2021), M. Meila and T. Zhang, Eds., vol. 139 of *Proceedings of Machine Learning Research*, PMLR, pp. 5637–5664.

[5] Turc, I., Chang, M., Lee, K., and Toutanova, K. Well-read students learn better: The impact of student initialization on knowledge distillation. *CoRR abs/1908.08962* (2019).

[1] R^2 represents the proportion of the variance that could be explained by input variables. Higher R^2 implies a better fit of the linear model and higher predictive power of input variables. We used adjusted R^2 to compensate the effect of extra degrees of freedom. It only increases if the new variable enhances the model above what would be obtained by chance.