# Unicorn on Rainbow: A Universal Commonsense Reasoning Model on a New Multitask Benchmark

**Nicholas Lourie,**[1] **Ronan Le Bras,**[1] **Chandra Bhagavatula,**[1] **Yejin Choi**[1,2]

[1]Allen Institute for AI, WA, USA,

[2]Paul G. Allen School of Computer Science & Engineering, University of Washington, WA, USA

{nicholasl, ronanlb, chandrab, yejinc}@allenai.org

## Abstract

Commonsense AI has long been seen as a near impossible goal—until recently. Now, research interest has sharply increased with an influx of new benchmarks and models.

We propose two new ways to evaluate commonsense models, emphasizing their generality on new tasks and building on diverse, recently introduced benchmarks. First, we propose a new multitask benchmark, Rainbow, to promote research on commonsense models that generalize well over multiple tasks and datasets. Second, we propose a novel evaluation, the **cost equivalent curve**, that sheds new insight on how the choice of source datasets, pretrained language models, and transfer learning methods impacts performance and *data efficiency*.

We perform extensive experiments—over 200 experiments encompassing 4800 models—and report multiple valuable and sometimes surprising findings, e.g., that transfer almost always leads to better or equivalent performance if following a particular recipe, that QA-based commonsense datasets transfer well with each other, while commonsense knowledge graphs do not, and that perhaps counter-intuitively, larger models benefit more from transfer than smaller ones.

Last but not least, we introduce a new universal commonsense reasoning model, Unicorn, that establishes new state-of-the-art performance across 8 popular commonsense benchmarks, $\alpha$NLI (→**87.3%**), CosmosQA (→**91.8%**), HellaSWAG (→**93.9%**), PIQA (→**90.1%**), SocialIQa (→**83.2%**), WinoGrande (→**86.6%**), CycIC (→**94.0%**) and CommonsenseQA (→**79.3%**).

## 1 Introduction

In AI's early years, researchers sought to build machines with common sense (McCarthy 1959); however, in the following decades, common sense came to be viewed as a near impossible goal. It is only recently that we see a sudden increase in research interest toward commonsense AI, with an influx of new benchmarks and models (Mostafazadeh et al. 2016; Talmor et al. 2019; Sakaguchi et al. 2020).

This renewed interest in common sense is ironically encouraged by both the great empirical strengths and limitations of large-scale pretrained neural language models. On one hand, pretrained models have led to remarkable progress across the board, often surpassing human performance on leaderboards (Radford et al. 2018; Devlin et al. 2019; Liu et al. 2019b; Raffel et al. 2019). On the other hand, pretrained language models continue to make surprisingly silly and *nonsensical* mistakes, even the recently introduced GPT-3.[1] This motivates new, relatively under-explored research avenues in commonsense knowledge and reasoning.

In pursuing commonsense AI, we can learn a great deal from mainstream NLP research. In particular, the introduction of multitask benchmarks such as GLUE (Wang et al. 2019b) and SuperGLUE (Wang et al. 2019a) has encouraged fundamental advances in the NLP community, accelerating research into models that robustly solve many tasks and datasets instead of overfitting to one in particular. In contrast, commonsense benchmarks and models are relatively nascent, thus there has been no organized effort, to

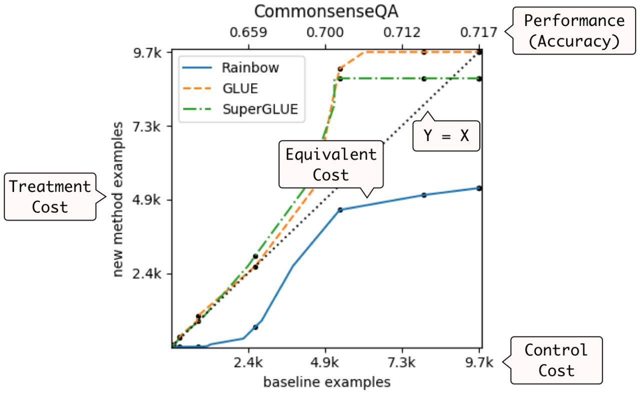

Figure 1: Cost equivalent curves comparing transfer learning from GLUE, SuperGLUE, and Rainbow onto CommonsenseQA. Each curve plots how much training data the single-task baseline (the $x$-axis) needs compared to the multitask method (the $y$-axis) to achieve the same performance (shown on the top axis in accuracy). Curves below the diagonal line ($y = x$) indicate that the multitask method needs less training data from the target dataset than the single-task baseline for the same performance. Thus, lower curves mean more successful transfer learning.

[1] https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/

date, at administering a collection of diverse commonsense benchmarks and investigating transfer learning across them.

We address exactly this need, proposing two new ways to evaluate commonsense models with a distinct emphasis on their generality across tasks and domains. First, we propose a new multi-task benchmark, RAINBOW, to facilitate research into commonsense models that generalize well over multiple different tasks and datasets. Second, we propose a novel evaluation, the **cost equivalent curve**, that sheds new insight on how different choices of source datasets, pretrained language models, and transfer learning methods affect performance and data efficiency in the target dataset.

The primary motivation for cost equivalent curves is **data efficiency**. The necessary condition for state-of-the-art neural models to maintain top performance on any dataset is a sufficiently large amount of training data for fine-tuning. Importantly, building a dataset for a new task or a domain is an expensive feat, easily costing tens of thousands of dollars (Zellers et al. 2018). Therefore, we want the models to *generalize systematically* across multiple datasets, instead of relying solely on the target dataset.

Shown in Figure 1, the cost equivalent curve aims to answer the following intuitive question: *how much data does a transfer learning approach save over the baseline that doesn't benefit from transfer learning?* We provide a more detailed walk-through of this chart in §2. As will be seen, cost equivalent curves have distinct advantages over simple evaluations at the full dataset size or classical learning curves drawn for each method and dataset separately, as they provide more accurate comparative insights into data efficiency in the context of multitasking and transfer learning.

We leverage these new tools to reevaluate common approaches for *intermediate-task transfer* (Pruksachatkun et al. 2020). Through extensive experiments, we identify multiple valuable and sometimes surprising findings, e.g., that intermediate-task transfer can always lead to better or equivalent performance if following a particular recipe, that QA-based commonsense datasets transfer well to each other, while commonsense knowledge graphs do not, and that perhaps counter-intuitively, larger models benefit much more from transfer learning compared to smaller ones.

In addition to the empirical insights, we also introduce a new universal commonsense reasoning model: UNICORN, establishing new state-of-the-art performances across 8 benchmarks: $\alpha$NLI (**87.3%**) (Bhagavatula et al. 2020), COSMOSQA (**91.8%**) (Huang et al. 2019), HELLASWAG (**93.9%**) (Zellers et al. 2019), PIQA (**90.1%**) (Bisk et al. 2020), SOCIALIQA (**83.2%**) (Sap et al. 2019b), WINOGRANDE (**86.6%**) (Sakaguchi et al. 2020), CYCIC (**94.0%**),[2] as well as the popular COMMONSENSEQA dataset (**79.3%**) (Talmor et al. 2019). Beyond setting records with the full training sets, our ablations show UNICORN also improves data efficiency for all training dataset sizes.

For reproducibility, we publicly release the UNICORN model and code, all the experimental results, and the RAINBOW leaderboard at https://github.com/allenai/rainbow.

## 2 Cost Equivalent Curves

Cost equivalent curves show *equivalent costs* between the single-task baseline and a new transfer-based approach. In this work, we define *cost* as *the number of training examples in the target dataset*. Intuitively, we want to measure how many examples the new approach needs to match the single-task baseline's performance as the amount of data varies.

Figure 1 illustrates cost equivalent curves with COMMONSENSEQA as the target dataset. The $x$-axis shows the number of examples used by the single-task baseline, while the $y$-axis shows the examples from the target dataset used by the new multitask method. The curve is where they achieve the same performance. The numbers on top of the figure show the performance corresponding to the number of baseline examples from the $x$-axis. For example, with 4.9k examples, the baseline achieves 70% accuracy. For any number of examples the baseline might use, we can see how many examples the new approach would require to match it. In Figure 1, to match the baseline's performance on $\sim$10k examples, multitasking with RAINBOW requires about 5k, while multitasking with GLUE requires more than 10k. Thus, *lower is better*, with curves below the diagonal ($y = x$) indicating that the new method improves over the baseline.

The construction of cost equivalent curves makes one technical assumption: the relationship between performance and cost is continuous and strictly monotonic (i.e., increasing or decreasing). This assumption holds empirically for parameters, compute, and data (Kaplan et al. 2020). Thus, we can safely estimate each learning curve with isotonic regression (Barlow et al. 1972), then construct the cost equivalent curve by mapping each dataset size to the baseline performance, finding the matching performance on the new method's curve, and seeing how many examples are required.

Cost equivalent curves visualize how a new approach impacts the cost-benefit trade-off, i.e. examples required for a given performance. This reframes the goal from pushing up performance on a fixed-size benchmark to most efficiently solving the problem. While we focus on data efficiency in this work, the idea of cost equivalent curves can be applied to other definitions of cost as well (e.g., GPU compute).

## 3 RAINBOW

We define RAINBOW, a suite of commonsense benchmarks, with the following datasets. To keep evaluation clean-cut, we only chose multiple-choice question-answering datasets.

**$\alpha$NLI** (Bhagavatula et al. 2020) tests abductive reasoning in narratives. It asks models to identify the best explanation among several connecting a beginning and ending.

**COSMOSQA** (Huang et al. 2019) asks commonsense reading comprehension questions about everyday narratives.

**HELLASWAG** (Zellers et al. 2019) requires models to choose the most plausible ending to a short context.

**PIQA** (Bisk et al. 2020) is a multiple-choice question answering benchmark for physical commonsense reasoning.

**SOCIALIQA** (Sap et al. 2019b) evaluates commonsense reasoning about social situations and interactions.

[2] The CYCIC dataset and leaderboard are available at https://leaderboard.allenai.org/cycic.

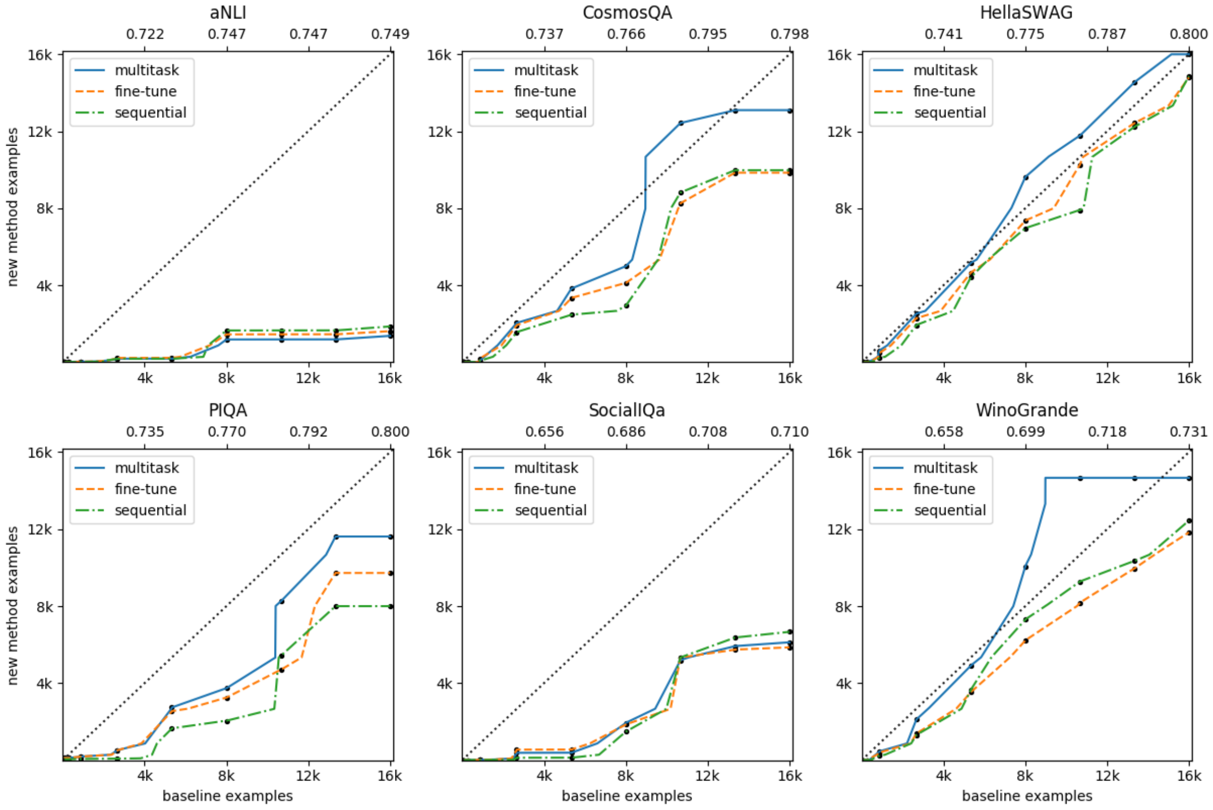

Figure 2: A comparison of transfer methods on RAINBOW tasks with T5-LARGE. Each plot varies the data available for one task while using all data from the other five to generate the cost equivalent curve. Performance is measured by dev set accuracy.

| TRANSFER | $\alpha$NLI | COSMOSQA | HELLASWAG | PIQA | SOCIALIQA | WINOGRANDE |
|---|---|---|---|---|---|---|
| multitask | 78.4 | 81.1 | 81.3 | 80.7 | 74.8 | 72.1 |
| fine-tune | 79.2 | 82.6 | **83.1** | **82.2** | 75.2 | 78.2 |
| sequential | **79.5** | **83.2** | 83.0 | **82.2** | **75.5** | **78.7** |
| none | 77.8 | 81.9 | 82.8 | 80.2 | 73.8 | 77.0 |

Table 1: A comparison of transfer methods' dev accuracy (%) on the RAINBOW tasks, using the T5-LARGE model.

- WINOGRANDE (Sakaguchi et al. 2020) is a large-scale collection of Winograd schema-inspired problems requiring reasoning about both social and physical interactions.

## 4 Empirical Insights

We present results from our large-scale empirical study, using pretrained T5-LARGE to transfer between datasets. We've grouped our findings and their relevant figures around the four following thematic questions.

### 4.1 What's the Best Approach for Transfer?

We compare three recipes for intermediate-task transfer:

**(1) multitask training** (Caruana 1995): training on multiple datasets (*including* the target dataset) all at once,

**(2) sequential training** (Pratt, Mostow, and Kamm 1991): first training on multiple datasets (*excluding* the target dataset) through multitask training, and then continuing to train on the target dataset alone,

**(3) multitask fine-tuning** (Liu et al. 2019a): first training on all datasets (*including* the target dataset) through multitask training, and then continuing to fine-tune on the target dataset alone.

Figure 2 compares these three methods on each of the six RAINBOW tasks, using the other five datasets for transfer.

**Finding 1: Sequential training almost always matches or beats other approaches.** Generally, sequential and multitask fine-tune training use fewer examples to achieve the same performance as multitask training or the single task baseline. For some tasks ($\alpha$NLI and SOCIALIQA), all three methods perform similarly; however, on the rest, sequential and multitask fine-tune training greatly improve data effi-

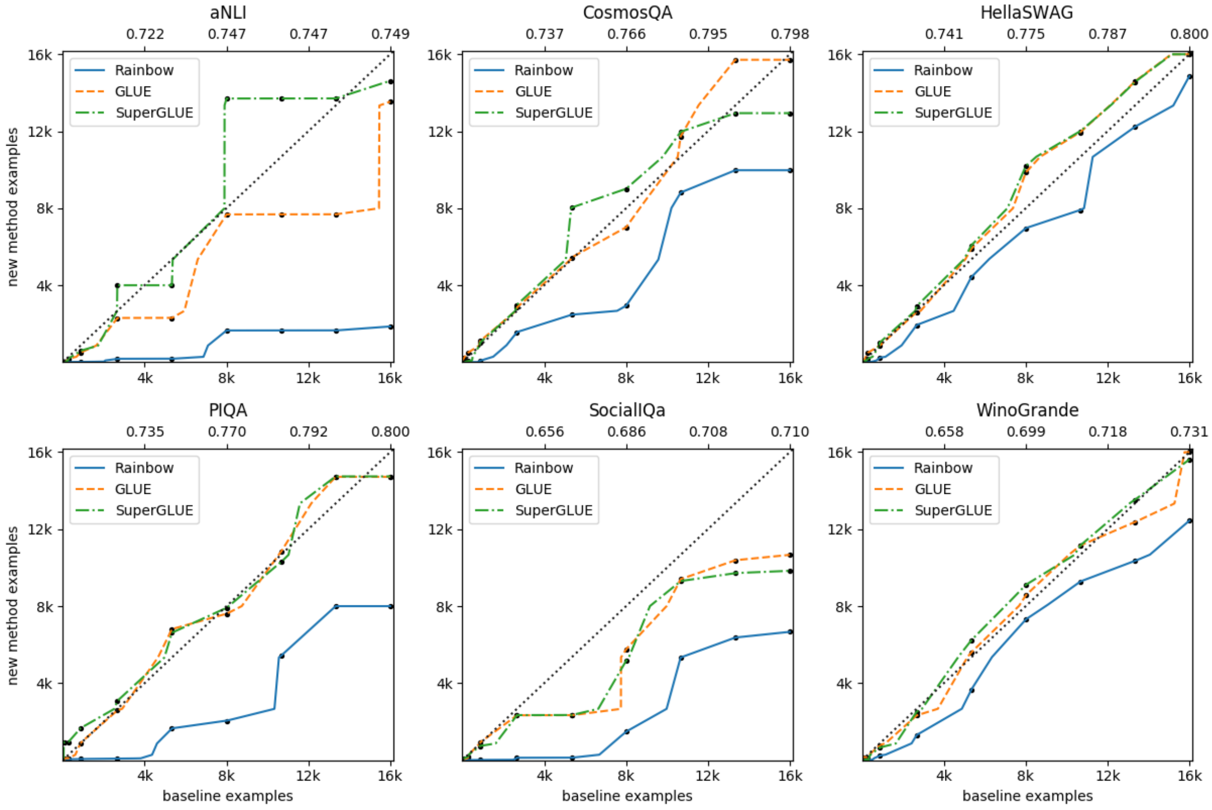

Figure 3: A comparison of multisets' transfer to RAINBOW tasks using sequential training with T5-LARGE. Performance is measured by dev set accuracy. For transfer from RAINBOW, we hold out the end task from the first round of fine-tuning.

| MULTISET | $\alpha$NLI | COSMOSQA | HELLASWAG | PIQA | SOCIALIQA | WINOGRANDE |
|---|---|---|---|---|---|---|
| GLUE | 78.5 | 81.4 | 82.3 | 80.8 | 74.3 | 77.7 |
| SUPERGLUE | 79.1 | 82.2 | 82.5 | 80.7 | 74.6 | 77.6 |
| RAINBOW | **79.5** | **83.2** | **83.0** | **82.2** | **75.5** | **78.7** |
| single task | 77.8 | 81.9 | 82.8 | 80.2 | 73.8 | 77.0 |

Table 2: A comparison of dev accuracy for multisets' transfer to RAINBOW via sequential training with T5-LARGE.

ciency. While sequential and multitask fine-tune training are often comparable, sequential training appears to be slightly more data efficient, both from comparing cost equivalent curves in Figure 2 and full dataset performance in Table 1.

**Finding 2: Sequential training rarely hurts performance.** While multitask training doesn't always beat the single task baseline, sequential and multitask fine-tune training uniformly outperform it—for all RAINBOW tasks and dataset sizes (including full datasets). This pattern mostly holds with other source and target tasks, especially for sequential training which rarely significantly harms performance.

**Finding 3: Multitask training helps most often in the low-data regime.** One mystery researchers currently face is the inconsistent effect of multitask learning: sometimes it helps, sometimes it hurts, sometimes it has no effect. Cost equivalent curves reveal one potential explanation: multitask learning tends to help when data is scarce, but may hurt performance if data is plentiful. In Figure 2, all cost equivalent curves initially require fewer examples than the single-task baseline (the $y = x$ line), while on some tasks (HELLASWAG and WINOGRANDE) multitasking eventually needs more data than the baseline. Table 1 reinforces this story, where multitask learning hurts performance on three of the six tasks (COSMOSQA, HELLASWAG, and WINOGRANDE), with WINOGRANDE dropping from 77.0% to 72.1% accuracy. The fact that such trends depend on things like data size shows the importance of examining a range of scenarios: changing the context can even reverse one's conclusions.

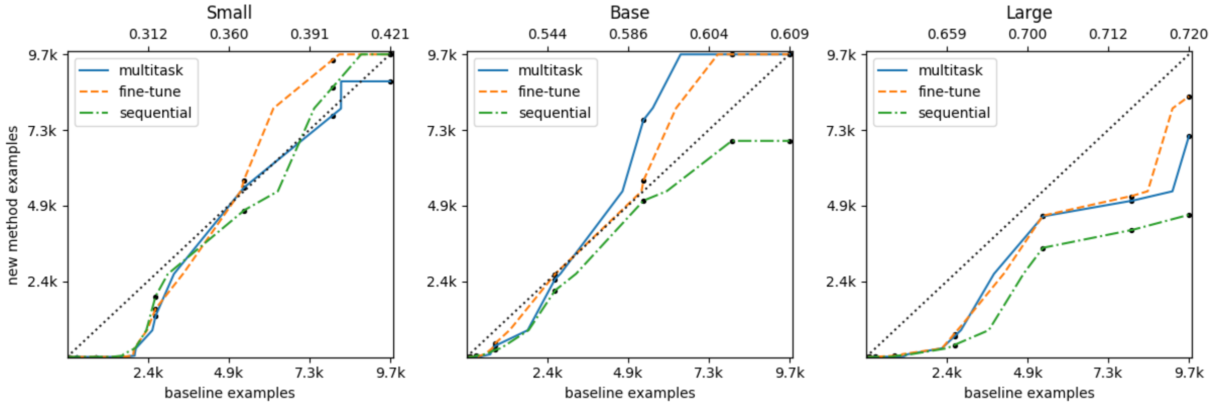

Figure 4: Cost equivalent curves comparing the effect of transfer across differently sized models on COMMONSENSEQA.

## 4.2 What Transfers Best for Common Sense?

Understanding when datasets transfer well is still an open and active area of research (Vu et al. 2020; Pruksachatkun et al. 2020). At present, modelers usually pick datasets that seem similar to the target, whether due to format, domain, or something else. To investigate common sense transfer, we compare how the RAINBOW tasks transfer to each other against two other popular dataset collections: GLUE and SUPERGLUE. Following the insights from Section 4.1, we use the strongest transfer method, sequential training, for the comparison. Figure 3 presents cost equivalent curves and Table 2 provides full dataset numbers.

**Finding 4: RAINBOW transfers best for common sense.** Across all six RAINBOW tasks and all training set sizes, the RAINBOW tasks transfer better to each other than GLUE and SUPERGLUE do to them. The same result also holds for the popular benchmark COMMONSENSEQA when multitask training (Figure 1); though, when multitasking with JOCI (Zhang et al. 2017), an ordinal commonsense variant of natural language inference, RAINBOW appears either not to help or to slightly hurt data efficiency—potentially more so than GLUE and SUPERGLUE.[3]

**Finding 5: Only RAINBOW uniformly beats the baseline.** With sequential training and T5-BASE or larger, RAINBOW improves data efficiency and performance for *every* task considered. Importantly, this pattern breaks down when multitask training, for which no multiset uniformly improved performance. Thus, sequential training can unlock useful transfer even in contexts where multitask training cannot. Likewise, smaller models demonstrated less transfer, as discussed further in Section 4.3. Consequently, T5-SMALL (the smallest model) did not always benefit. In contrast to RAINBOW, GLUE and SUPERGLUE often had little effect or slightly decreased data efficiency.

**Caveats about GLUE, SUPERGLUE, and T5.** There's an important caveat to note about T5, the model used in our experiments, and its relationship to GLUE and SUPERGLUE. The off-the-shelf T5's weights come from multitask pretraining, where many tasks are mixed with a language modeling objective to learn a powerful initialization for the weights. In fact, both GLUE and SUPERGLUE were mixed into the pretraining (Raffel et al. 2019). So, while RAINBOW clearly improves data efficiency and performance, our experiments do not determine whether some of the benefit comes from the novelty of RAINBOW's knowledge to T5, as opposed to containing more general information than GLUE and SUPERGLUE.

## 4.3 Does Model Size Affect Transfer?

Most of our exhaustive experiments use T5-LARGE (770M parameters), but in practice, we might prefer to use smaller models due to computational limitations. Thus, we investigate the impact of model size on intermediate-task transfer using the T5-BASE (220M parameters) and T5-SMALL (60M parameters) models. Figure 4 presents the results for transferring with different model sizes from RAINBOW to COMMONSENSEQA.

**Finding 6: Larger models benefit more from transfer.** Since larger pretrained models achieve substantially higher performance, it's difficult to compare transfer's effect across model size. The baselines start from very different places. Cost equivalent curves place everything in comparable units, *equivalent baseline cost* (e.g., number of training examples). Capitalizing on this fact, Figure 4 compares transfer from RAINBOW to COMMONSENSEQA across model size. The cost equivalent curves reveal a trend: larger models seem to benefit more from transfer, saving more examples over the relevant baselines. Since smaller models require more gradient updates to converge (Kaplan et al. 2020), it's important to note that we held the number of gradient updates fixed for comparison. Exploring whether this trend holds in different contexts, as well as theoretical explanations, are promising directions for future work.

[3] For these additional experiments, see the extended experimental results at https://github.com/allenai/rainbow.

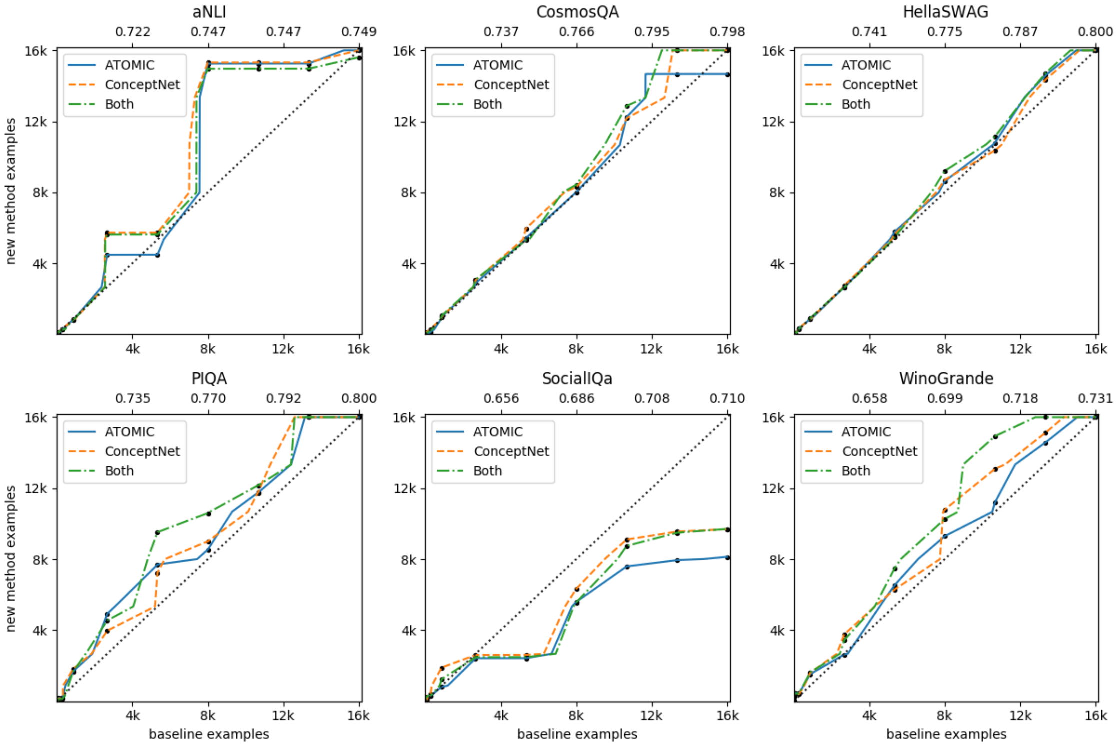

Figure 5: Cost equivalent curves comparing transfer from generative training on different common sense knowledge graphs using multitask training with T5-LARGE, across different RAINBOW tasks. Performance is measured by dev set accuracy.

| KNOWLEDGE GRAPH | $\alpha$NLI | COSMOSQA | HELLASWAG | PIQA | SOCIALIQA | WINOGRANDE |
|---|---|---|---|---|---|---|
| ATOMIC | **78.3** | 81.8 | **82.8** | 79.9 | **75.0** | **78.2** |
| CONCEPTNET | 78.0 | 81.8 | 82.5 | 80.5 | 74.3 | 76.3 |
| BOTH | 78.0 | 81.8 | 82.7 | **81.1** | 74.8 | 76.6 |
| single task | 77.8 | **81.9** | **82.8** | 80.2 | 73.8 | 77.0 |

Table 3: A comparison of dev accuracy when generatively training on knowledge graphs in a multitask setup using T5-LARGE.

**Finding 7: Sequential training wins across model sizes.** Figure 4 expands Finding 1, that sequential training generally matches or beats the other transfer approaches, by supporting it across model sizes. In all three plots, sequential training appears in line with or better than the other transfer methods.

## 4.4 Can Models Transfer from Knowledge Graphs to QA Datasets?

Due to reporting bias (Gordon and Van Durme 2013), common sense rarely appears explicitly in text, though it does appear implicitly. While language models learn much of the common sense implicit in natural language (Trinh and Le 2018), crowdsourced and expert curated knowledge might provide complementary information. To investigate, we explored multitask transfer from two popular common sense knowledge graphs, CONCEPTNET (Speer, Chin, and Havasi 2017) and ATOMIC (Sap et al. 2019a). Using the knowledge graphs, we created subject-relation-object triples and had the model generate either the subject or object from the other two components (Bosselut et al. 2019). Each triple's parts were wrapped in XML-like tags and concatenated before being fed into the model. The results are summarized in Figure 5 and Table 3.

**Finding 8: Knowledge graph multitasking shows little impact.** The results are generally negative. Only SOCIALIQA benefits, which might come from the use of ATOMIC during its construction. We offer two possible explanations: the serialized language from the knowledge graphs is not in a QA format, and the knowledge graph completion task is generative while all other tasks are discriminative. These discrepancies may present too large an obstacle for effective transfer. Our findings encourage future research to better close the gap between knowledge graphs and datasets. Given sequential training's strength, as exemplified in Findings 1, 2, and 7, it may lead to different results than the multitask transfer we explore here.

## 5 Unicorn

Finally, we present our universal commonsense reasoning model, Unicorn. Motivated by Finding 1, our primary goal with Unicorn is to provide a pretrained commonsense reasoning model ready to be fine-tuned on other downstream commonsense tasks. This is analogous to how off-the-shelf T5 models are multitasked on NLP benchmarks such as GLUE and SuperGLUE as part of their pretraining.

In order to see the limit of the best performance achievable, we start by multitasking T5-11B on Rainbow. We then trained Unicorn on each task individually, except for WinoGrande which required separate handling since it evaluates models via a learning curve. For WinoGrande, we multitasked the other five Rainbow datasets and then trained on WinoGrande.[4] In each case, we used the same hyper-parameters as Unicorn did during its initial multitask training, extending each of the 8 combinations tried at that stage. The best checkpoints were chosen using accuracy on dev.

**SOTA on Rainbow.** We establish new SOTA on all Rainbow datasets: $\alpha$NLI (**87.3%**), CosmosQA (**91.8%**), HellaSWAG (**93.9%**), PIQA (**90.1%**), SocialIQa (**83.2%**), and WinoGrande (**86.6%**).[5]

**SOTA on datasets beyond Rainbow.** While SOTA results on Rainbow are encouraging, we still need to check if Unicorn's strong performance is confined to Rainbow or generalizes beyond it. Thus, we evaluated on two additional commonsense benchmarks: CycIC (**94.0%**) and CommonsenseQA (**79.3%**). Again, Unicorn achieved SOTA on both.

## 6 Related Work

**Scaling Laws** In contemporary machine learning, simple methods that scale often outperform complex ones (Sutton 2019). Accordingly, recent years have seen a sharp rise in compute used by state-of-the-art methods (Amodei and Hernandez 2018). Performance gains from increasing data, parameters, and training are not only reliable, but empirically predictable (Hestness et al. 2017; Sun et al. 2017; Rosenfeld et al. 2020; Kaplan et al. 2020). For example, Sun et al. (2017) found that models need exponential data for improvements in accuracy.[6] These observations, that scaling is reliable, predictable, and critical to the current successes, motivate our focus on evaluation based on *cost-benefit trade-offs*, i.e. the cost equivalent curve.

**Commonsense Benchmarks** Rapid progress in modeling has led to a major challenge for NLP: the creation of suitable benchmarks. Neural models often cue off statistical biases and annotation artifacts to solve datasets without understanding tasks (Gururangan et al. 2018). To address this issue, recent commonsense benchmarks often use adversarial filtering (Zellers et al. 2018; Le Bras et al. 2020): a family of techniques that remove easily predicted examples from datasets. Besides CosmosQA, all Rainbow tasks use this technique. Many more common sense benchmarks exist beyond what we could explore here (Roemmele, Bejan, and Gordon 2011; Levesque, Davis, and Morgenstern 2011; Mostafazadeh et al. 2016).

**Transfer Learning** Semi-supervised and transfer learning have grown into cornerstones of NLP. Early work learned unsupervised representations of words (Brown et al. 1992; Mikolov et al. 2013), while more recent work employs contextualized representations from neural language models (Peters et al. 2018). Radford et al. (2018) demonstrated that language models could be fine-tuned directly to solve a wide-variety of tasks by providing the inputs encoded as text, while Devlin et al. (2019) and others improved upon the technique (Yang et al. 2019; Liu et al. 2019b; Lan et al. 2019). Most relevant to this work, Raffel et al. (2019) introduced T5 which built off previous work to reframe any NLP task as text-to-text, dispensing with the need for task-specific model adaptations.

**Data Efficiency & Evaluation** Other researchers have noted the importance of cost-benefit trade-offs in evaluation (Schwartz et al. 2019). Dodge et al. (2019) advocate reporting the compute-performance trade-off caused by hyper-parameter tuning for new models, and provide an estimator for expected validation performance as a function of hyper-parameter evaluations. In an older work, Clark and Matwin (1993) evaluated the use of qualitative knowledge in terms of saved training examples, similarly to our cost equivalent curves. In contrast to our work, they fitted a linear trend to the learning curve and counted examples saved rather than plotting the numbers of examples that achieve equivalent performance.

## 7 Conclusion

Motivated by the fact that increased scale reliably improves performance for neural networks, we reevaluated existing techniques based on their data efficiency. To enable such comparisons, we introduced a new evaluation, the cost equivalent curve, which improves over traditional learning curves by facilitating comparisons across otherwise hard-to-compare contexts. Our large-scale empirical study analyzed state-of-the-art techniques for transfer on pretrained language models, focusing on learning general, commonsense knowledge and evaluating on common sense tasks. In particular, we introduced a new collection of common sense datasets, Rainbow, and using the lessons from our empirical study trained a new model, Unicorn, improving state-of-the-art results across 8 benchmarks. We hope others find our empirical study, new evaluation, Rainbow, and Unicorn useful in their future work.

[4] While sequential training for the Rainbow tasks would likely yield the best results, it would have required much more compute.

[5] All tasks use accuracy for evaluation except WinoGrande which uses area under the dataset size–accuracy learning curve.

[6] Eventually, models saturate and need *super-exponential* data.

## Acknowledgements

We would like to thank the anonymous reviewers for their valuable feedback. This research was supported in part by NSF (IIS-1524371), the National Science Foundation Graduate Research Fellowship under Grant No. DGE 1256082, DARPA CwC through ARO (W911NF15-1- 0543), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI. Computations on beaker.org were supported in part by credits from Google Cloud. TPU machines for conducting experiments were generously provided by Google through the TensorFlow Research Cloud (TFRC) program.

## References

Amodei, D.; and Hernandez, D. 2018. AI and Compute. URL https://openai.com/blog/ai-and-compute/. (Last accessed 2021-03-24).

Barlow, R.; Bartholomew, D.; Bremner, J.; and Brunk, H. 1972. *Statistical Inference Under Order Restrictions: The Theory and Application of Isotonic Regression*. J. Wiley. ISBN 9780471049708.

Bhagavatula, C.; Le Bras, R.; Malaviya, C.; Sakaguchi, K.; Holtzman, A.; Rashkin, H.; Downey, D.; Yih, S. W.-t.; and Choi, Y. 2020. Abductive commonsense reasoning. *ICLR* .

Bisk, Y.; Zellers, R.; Le Bras, R.; Gao, J.; and Choi, Y. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Çelikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Brown, P. F.; Della Pietra, V. J.; deSouza, P. V.; Lai, J. C.; and Mercer, R. L. 1992. Class-Based *n*-gram Models of Natural Language. *Computational Linguistics* 18(4): 467–480. URL https://www.aclweb.org/anthology/J92-4003.

Caruana, R. 1995. Learning Many Related Tasks at the Same Time with Backpropagation. In Tesauro, G.; Touretzky, D. S.; and Leen, T. K., eds., *Advances in Neural Information Processing Systems 7*, 657–664. MIT Press. URL http://papers.nips.cc/paper/959-learning-many-related-tasks-at-the-same-time-with-backpropagation.pdf.

Clark, P.; and Matwin, S. 1993. Using qualitative models to guide inductive learning. In *Proceedings of the 1993 international conference on machine learning*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

Dodge, J.; Gururangan, S.; Card, D.; Schwartz, R.; and Smith, N. A. 2019. Show Your Work: Improved Reporting of Experimental Results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2185–2194. Hong Kong, China: Association for Computational Linguistics. doi:10.18653/v1/D19-1224. URL https://www.aclweb.org/anthology/D19-1224.

Gordon, J.; and Van Durme, B. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30. ACM.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL*. URL https://www.aclweb.org/anthology/N18-2017/.

Hestness, J.; Narang, S.; Ardalani, N.; Diamos, G. F.; Jun, H.; Kianinejad, H.; Patwary, M. M. A.; Yang, Y.; and Zhou, Y. 2017. Deep Learning Scaling is Predictable, Empirically. *ArXiv* abs/1712.00409.

Huang, L.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2019. Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. In *EMNLP/IJCNLP*.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* .

Lan, Z.; Chen, M.; Goodman, S.; Gimpel, K.; Sharma, P.; and Soricut, R. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* .

Le Bras, R.; Swayamdipta, S.; Bhagavatula, C.; Zellers, R.; Peters, M. E.; Sabharwal, A.; and Choi, Y. 2020. Adversarial Filters of Dataset Biases. *ArXiv* abs/2002.04108.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2011. The Winograd schema challenge. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, volume 46, 47.

Liu, X.; He, P.; Chen, W.; and Gao, J. 2019a. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4487–4496. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1441. URL https://www.aclweb.org/anthology/P19-1441.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

McCarthy, J. 1959. Programs with Common Sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, 75–91. London: Her Majesty's Stationary Office.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 26*, 3111–3119. Curran Associates, Inc. URL http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849. San Diego, California: Association for Computational Linguistics. doi:10.18653/v1/N16-1098. URL https://www.aclweb.org/anthology/N16-1098.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. doi:10.18653/v1/N18-1202. URL https://www.aclweb.org/anthology/N18-1202.

Pratt, L.; Mostow, J.; and Kamm, C. 1991. Direct Transfer of Learned Information Among Neural Networks. In *AAAI*.

Pruksachatkun, Y.; Phang, J.; Liu, H.; Htut, P. M.; Zhang, X.; Pang, R. Y.; Vania, C.; Kann, K.; and Bowman, S. R. 2020. Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work? *arXiv preprint arXiv:2005.00628* .

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. (Last accessed 2021-03-24).

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv e-prints* .

Roemmele, M.; Bejan, C. A.; and Gordon, A. S. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Rosenfeld, J. S.; Rosenfeld, A.; Belinkov, Y.; and Shavit, N. 2020. A Constructive Prediction of the Generalization Error Across Scales. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=ryenvpEKDr.

Sakaguchi, K.; Le Bras, R.; Bhagavatula, C.; and Choi, Y. 2020. WINOGRANDE: An Adversarial Winograd Schema Challenge at Scale. In *AAAI*.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3027–3035.

Sap, M.; Rashkin, H.; Chen, D.; Le Bras, R.; and Choi, Y. 2019b. Social IQA: Commonsense Reasoning about Social Interactions. In *EMNLP 2019*.

Schwartz, R.; Dodge, J.; Smith, N. A.; and Etzioni, O. 2019. Green ai. *arXiv preprint arXiv:1907.10597* .

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4444–4451. AAAI Press.

Sun, C.; Shrivastava, A.; Singh, S.; and Gupta, A. 2017. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, 843–852.

Sutton, R. S. 2019. The Bitter Lesson. URL http://incompleteideas.net/IncIdeas/BitterLesson.html. (Last accessed 2021-03-24).

Talmor, A.; Herzig, J.; Lourie, N.; and Berant, J. 2019. CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge. 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)*, 4149–4158. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1421. URL https://www.aclweb.org/anthology/N19-1421.

Tange, O. 2011. GNU Parallel - The Command-Line Power Tool. *;login: The USENIX Magazine* 36(1): 42–47. doi:10.5281/zenodo.16303. URL http://www.gnu.org/s/parallel.

Trinh, T. H.; and Le, Q. V. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847* .

Vu, T.; Wang, T.; Munkhdalai, T.; Sordoni, A.; Trischler, A.; Mattarella-Micke, A.; Maji, S.; and Iyyer, M. 2020. Exploring and predicting transferability across nlp tasks. *arXiv preprint arXiv:2005.00770* .

Wang, A.; Pruksachatkun, Y.; Nangia, N.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019a. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *arXiv preprint 1905.00537* .

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019b. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In the Proceedings of ICLR.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R. R.; and Le, Q. V. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, 5753–5763.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Zellers, R.; Holtzman, A.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. HellaSwag: Can a Machine Really Finish Your Sentence? In *ACL*.

Zhang, S.; Rudinger, R.; Duh, K.; and Van Durme, B. 2017. Ordinal Common-sense Inference. *Transactions of the Association for Computational Linguistics* 5: 379–395. doi:10.1162/tacl_a_00068. URL https://www.aclweb.org/anthology/Q17-1027.