Capability Transfer from Large to Small Models with Synthetically-Generated Data

Kerem Dayı [1 *] **Lillian Sun** [1 *] **Emma Yang** [1 *]

Abstract

We investigate the transfer of capabilities from large language models to smaller models using synthetic, LLM-generated data. Instead of using human-annotated data, we explore whether a large model can effectively "teach" a smaller model natural language capabilities like summarization and question-answering through generated synthetic data. The large model acts as a teacher in generating both the training data and evaluation metrics, while a smaller student model learns exclusively from this synthetic data. We empirically investigate two key tasks, summarization and question-answering. Through this work, we aim to demonstrate the feasibility of a fully synthetic data driven pipeline for capability transfer. Our experiments demonstrate promising results for both tasks, displaying up to 56% performance improvement in summarization and at least on-par performance in question-answering on the synthetic capability metric. Our study highlights the potential of synthetic data as a scalable and cost-effective alternative to human annotation, paving the way for more efficient training of smaller models without sacrificing performance.

1. Introduction

As large language models continue to grow in size, so does the demand for model distillation methods. Research into scaling models down is critical to reduce their memory and computational footprint while preserving their performance. Larger models are often trained on very large datasets containing long documents or texts that may not fit in the context length of a smaller model. Also, smaller models may have decreased ability to reason about larger texts.

[*]Equal contribution [1]Harvard University, Cambridge, MA, USA. Correspondence to: Lillian Sun <lillian-sun@college.harvard.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

In addition to the overhead of training and deploying large models, obtaining high-quality, human-labeled or annotated data can be extremely costly. This is especially the case for data in specialized applications, such as understanding and summarizing texts that require highly-specialized domain knowledge to label and annotate. However, with the improvement of large language models' native comprehension abilities, we hypothesize that large models may be able to generate datasets for natural language tasks with minimal quality loss compared to human-generated data.

Thus, we study whether large models can generate training data to fine-tune smaller models and "teach" the small model capabilities, such summarizing or answering yes/no questions about a given text. The ability to train a smaller model to replicate the behavior of a larger model on a given task has many promises. First, it reduces the computation overhead introduced by the larger model due to prompt engineering, since the smaller model no longer needs to process a long prompt description of the task and the output format. Second, not relying on humans to generate data for a task creates more room for customizations of tasks such as generating output with particular linguistic features, particular lengths, and formats without extensive prompt engineering.

In order to investigate capability transfer from large to small models using data and performance metrics synthetically generated by the large model, we define the following experimental setups:

1. **Problem 1: Summarization.** To generate the synthetic dataset, a large model is given a text to summarize and asked to generate 5 key words that a good summary should have. The small model is then fine-tuned on the text-summary pairs and evaluated on a holdout dataset of texts. The summaries are evaluated on how many of the large model-generated keywords each includes.

2. **Problem 2: Q&A.** The synthetic dataset consists of summaries like in Problem 1, accompanied by a set of five yes/no questions generated by the large model that a reader should be able to answer upon seeing the summary. The small model is fine-tuned on the generated summaries and the question-answer pairs. For evalu-

ation, the small model is given unseen synthetically-generated summaries and evaluated on whether it can correctly answer yes/no questions about the summary.

The intuition behind these problem formulations is two-fold:

1. Both of the tasks outlined above help us investigate whether a large model can generate training data with enough signal such that a smaller model, with less capacity and a shorter effective context length, can learn to perform a capability from the larger model. The failure mode would be that the large model's generated data does not include sufficient information for the small model to learn from, and thus the small model does improve at the large model-generated performance metric.

2. By testing small models of various sizes, we can empirically determine how their ability to learn from synthetic data scales with model size relative to the size of the larger model and the generated training dataset.

2. Methodology

2.1. Synthetic Data Generation

To generate synthetic data for summarization and Q&A, we initially chose a base dataset which includes human written articles to be summarized. For both tasks, we base the human written articles on the CNN/Dailymail dataset (Hermann et al., 2015; Nallapati et al., 2016) and filter for CNN articles, resulting in approximately 80,000 unique news articles written in English. Each article is also accompanied by summarizing "highlights" written by the original author of the article. For the synthetic data used to fine-tune small models, we only use the CNN articles themselves and not the human-written summaries. With a budget of only $10, we were able to generate a summarization dataset of approximately 35,000 text-summary pairs, and a Q&A dataset of approximately 50,000 datapoints of summaries and question-answer pairs.

To generate our synthetic data, we primarily worked with GPT-4o-mini due to having a low cost and high speed while being a high-quality model. We used a mixture of the OpenAI Chat Completions API and the Batch API; the latter allowed us to process large amounts of requests in a single batch at a reduced cost. Nevertheless, our methodology is easily generalizable to other models such as GPT-4o or Gemini. Furthermore, our methodology is generalizable to other prompting strategies such as generating summaries with certain stylistic features or other natural language tasks. Prompts are provided in Section A.6.

2.1.1. Summarization Dataset

For the summarization task, we aim to jointly generate summaries and an evaluation metric for CNN articles. For this reason, we prompted GPT-4o-mini to propose five keywords that are important to be included in a summary, and a summary that includes those keywords. Note that our work generalizes to other ways of generating summaries based on other metrics, but we focus on important keywords for concreteness. Ultimately, we generated around 35,000 samples for the summarization dataset, which was later split into training and validation. An example of the prompts and the resulting generations, including an excerpt from the article being summarized, can be found in Figure 6 in Appendix Section A.3.

2.1.2. Q&A Dataset

To address shorter context lengths for smaller models, we generated a Q&A dataset where a summary for the article is sufficient to answering questions about that article. Hence, we prompted GPT-4o-mini to generate a summary and five questions that can be answered using the summary. Furthermore, we constrain the questions to be "Yes"/"No" questions so that we can evaluate the performance of different models through text classification. As a result, we generate approximately 50,000 data points that include an article, GPT-generated summary, and a question answer pair for that summary. An example of Q&A can be found in Figure 7 in Appendix Section A.3.

Experimental details are included in Appendix A.2.

3. Experimental Results

We fine-tuned the `flan-t5-small`, `flan-t5-base`, and `flan-t5-large` pre-trained models, available on HuggingFace, on the synthetic data generated with the methodology described in Section 3. All models were fine-tuned on a NVIDIA H100 GPU. Evaluation metrics on the validation dataset were then computed by performing inference on a NVIDIA A100 GPU using checkpoints from training. We performed experiments on each model size for both summarization and Q&A tasks to empirically observe capability transfer from large models to the small T5 models in both tasks and how capability transfer scales with small model size.

3.1. Summarization

Figure 1 presents the loss curve and capability trajectories for the three T5 models fine-tuned on the summarization dataset, along with their human-generated data baselines.

Synthetic Data vs Human-Generated Data The loss curves for all three models show that the model is able to

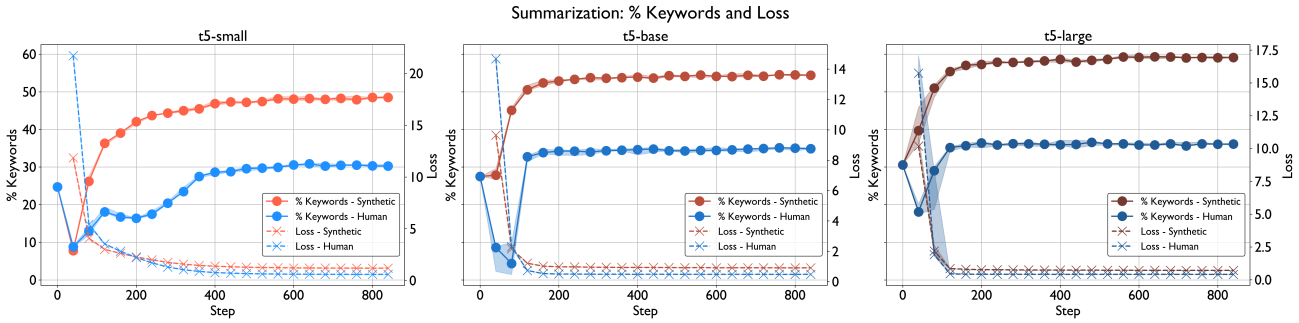


Figure 1. **Small models fine-tuned on synthetic summarization data outperform those fine-tuned on human-generated data**. For each model size, the loss curves and performance on the GPT-generated keywords metric are shown on the baseline model and 21 checkpoints throughout 1 epoch of fine-tuning for both synthetic summary data and human-generated highlights data. Each point marks the median loss/evaluation score and the shaded regions show the range from min to max loss/evaluation score.

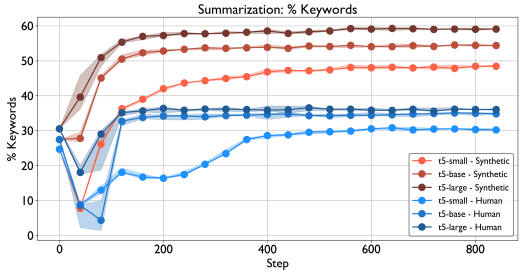
fit to both human-written and synthetically-generated summary data. Each model's improvement on the capability metric, the percentage of the five keywords included in the generated summary, throughout fine-tuning demonstrates that the synthetic data is more effective at transferring the summarization capability to the small model with respect to the GPT-generated metric. Figure 2(b) shows the best performance of each small model compared to the pre-finetuning model when trained on each dataset; each model fine-tuned on synthetic data outperforms the model trained on human-generated data by 52-57% for all three model sizes.

> ***Takeaway 1****: Small models can effectively learn from synthetic data and perform better than when fine-tuned on human data with respect to a synthetic data.*

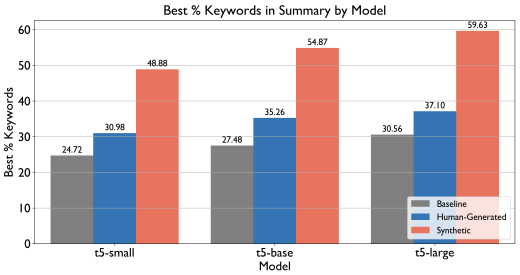
Impact of model size In all three models, the model fine-tuned on synthetic data outperforms the model fine-tuned on human-generated data when evaluated on the GPT-generated capability metric. In fact, as shown in Figure 2(b) and 2(a), not only does peak capability performance improve for both datasets as model size increases, but the smallest synthetic data model outperforms the largest human-generated data model. For the summarization task, increasing the model size leads to a noticeable 10% improvement in the capability metric when fine-tuned on the same synthetically-generated dataset. On the other hand, the improvement when training on the human-generated data is much more modest.

> ***Takeaway 2****: For learning to summarize text, increasing model size leads to better performance when learning from synthetic data, whereas the payoff when learning from human data is less obvious.*

Qualitative effects of fine-tuning In addition to a quantitative improvement in the evaluation metric as the model is fine-tuned, we also observe that the summaries generated by



(a) Keyword inclusion capability trajectory across fine-tuning steps for three model sizes.



(b) Best synthetic data performance exceeds that of human data on the keyword metric for all model sizes.

Figure 2. Performance comparison on the summarization task using the GPT-generated keyword inclusion metric. Synthetic data enables more effective capability transfer than human-generated data, particularly for smaller models.

the small models improve with training, with more marked improvement in the coherence and quality of the summaries when fine-tuning on synthetic data compared to the human-generated summaries. Table 1 in Appendix Section A.5 shows the summaries generated for the same prompt in the validation set by each small model. We show the generated summaries before fine-tuning, 240 steps into training (25% of 1 epoch), and at the end of 1 epoch of fine-tuning on synthetic and human data.

Figure 3 shows a sample of generated summaries of an article before fine-tuning and after fine-tuning on synthetic data and human-generated data. The pre-trained, un-finetuned

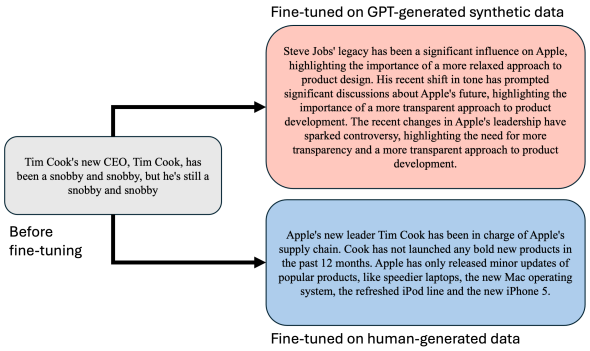


Figure 3. Generated summaries by the `flan-t5-small` model before and after fine-tuning on synthetic and human-generated data.

model exhibits significant repetition, making the summary incoherent, and fails to convey any information about the article itself. After finetuning on the synthetic data, some repetition (e.g. of the idea of transparency) is still present, but the summary becomes much more effective at including information expressed by the article. The sentences are longer than those generated by the un-finetuned model, showing that the model improves at generating longer, grammatically-correct, and meaningful sentences. When fine-tuned on the human-generated data, the model also improves in generating syntactically-correct and meaningful sentences. However, it fails to convey some of the important ideas expressed in the original article (e.g. changes in the leadership style between CEOs, which the GPT-generated summary highlights). While it is worth noting that the human-written highlights in the dataset may be in a different style than the paragraph-form highlights, using shorter sentences, the non-negligeable difference in the amount of information conveyed in the summary generated by the model fine-tuned on human data demonstrates the effect of each type of data on capability transfer.

> ***Takeaway 3****: Fine-tuning on synthetically-generated summaries generates more coherent and expressive summaries than learning from the human-generated summaries from the CNN/DailyMail dataset.*

3.2. Q&A

Figure 4 shows the loss curves and capability metric trajectories for fine-tuning each small T5 model on the synthetically-generated and human-generated Q&A data. As with the summarization data, all three models are able to fit to both datasets, but observations about the ability of each model to learn from the synthetic and human-generated data can be made from the trajectories of the synthetic capability metric.

Synthetic vs. Human-Generated Data Figure 5(a) show the trajectories and relative accuracy of the model across three sizes and fine-tuned on synthetic and human-generated datasets on the GPT-generated question-answering benchmarks.

While all three models of various sizes fine-tuned on synthetic data outperformed the models fine-tuned on human-generated data in the summarization task, the synthetic models only outperform their human-generated counterparts of the same size in the Q&A task. Concretely, fine-tuning on synthetic summaries and synthetic question-answer pairs improves the best performance of `flan-t5-small` over the human-generated data by 4.5%, of `flan-t5-base` by 0.68%, and of `flan-t5-large` by 0.08%. The relative best performances are also shown in Figure 5(b). It is important to note that 75% of the answers in the dataset (and the validation set, by random downsampling) are "yes" answers; the baseline `flan-t5-small` model effectively always answers "yes," whereas the two other baseline models always answer "no."

Impact of Model Size As already noted, whereas all models fine-tuned with synthetic data outperform models fine-tuned with human data in the summarization task, the performance improvement only manifests *pairwise* for the question-answering task. With increasing model size, the performance gap markedly shrinks, until the discrepancy is essentially negligible for the large T5 model.

> ***Takeaway 4****: Fine-tuning on synthetic data for the Q&A task, where answers are yes/no, reaches competitive performance compared to models of the same size fine-tuned on human data. The advantage gained by using synthetic data increases as model size decreases.*

4. Discussion

Our experiments highlight several core insights into how models learn from synthetic versus human-annotated data across the two tasks of summarization and question-answering (Q&A). We analyze our experimental results across training steps and model sizes to explore how capability transfer improves and scales.

4.1. Summarization Task

Before fine-tuning, the summaries generated by the T5 models have low performance on the keyword inclusion metric (35-43% as shown in Figure 1) and are largely incoherent or lack significant information about the article (see Table 1). However, the improvement observed in performance metrics and the decrease in loss with fine-tuning for both human and synthetic generate data indicates that there is sufficient signal in the data for the model to learn to generate meaningful summaries. Furthermore, the persistent performance gap be-

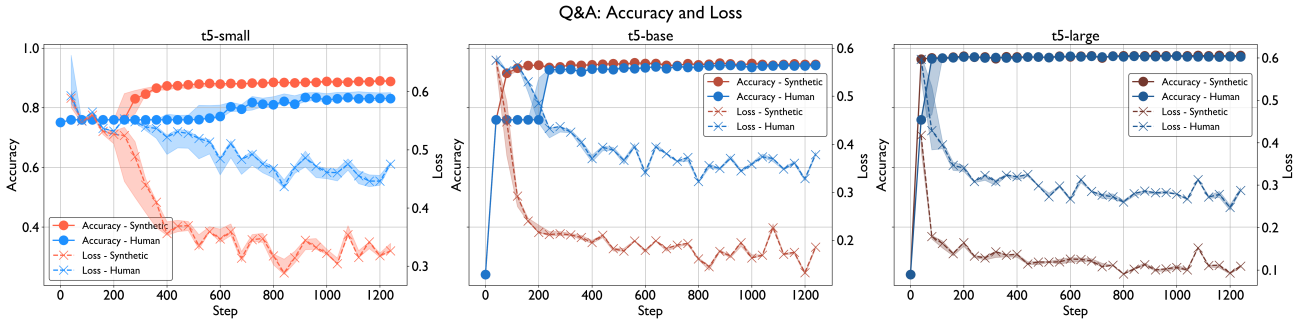
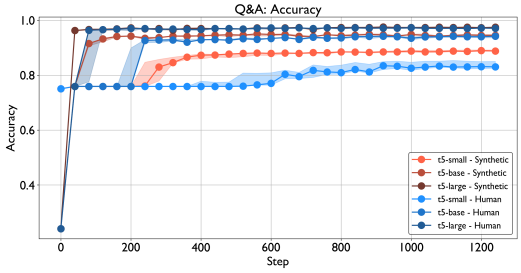
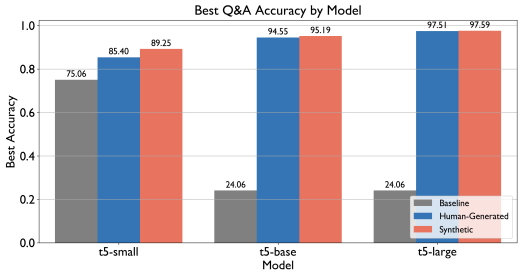


Figure 4. **Small models fine-tuned on synthetic Q&A data match the performance of models fine-tuned on human data.** For each model size, the loss curves and performance on the GPT-generated question-answering metric are shown on the baseline model and 32 checkpoints throughout 1 epoch of fine-tuning for both synthetic summary data and human-generated highlights data with the GPT-generated questions. Each point marks the median loss/evaluation score and the shaded regions show the range from min to max loss/evaluation score.



(a) Accuracy trajectories during fine-tuning on synthetic and human-generated data for three model sizes. Shaded regions indicate the min-max accuracy range.



(b) Best synthetic data performance exceeds or matches that of human data, especially in smaller models.

Figure 5. Comparison of model performance on the question-answering task when fine-tuned on synthetic versus human-generated data. Synthetic data yields comparable or superior results across model sizes, with the greatest gains observed in smaller models.

tween the synthetic and human data conditions suggest that fine-tuning the small models on synthetic data achieve better outcomes as measured by the keyword inclusion metric than those trained on human data.

We also find that increasing model size strictly increases evaluation performance on the keyword metric. Figure 2(a) shows that across all model sizes and configurations, models trained on synthetic data consistently outperform those trained on human-generated summaries. Furthermore, our results with scaling up the small model size indicate that our generated synthetic data for summarization task contains sufficient information content such that models can keep learning and improving performance as their size and capabilities increase.

It is important to note that our evaluation metric may be biased, given that the keywords were generated based on the synthetic summaries. Thus, the keywords may be more naturally aligned with the synthetic data's structure and style, as well as what the large model deemed to be key information expressed in the article. We can observe this bias qualitatively by examining the style of the summaries generated by the model in table 1. If we assume the metric is valid and meaningfully captures summary quality, then these results imply that synthetic data can achieve better and more efficient performance than costlier human-labeled data. Since synthetic data is typically **cheaper** and can be **produced at much larger scale**, these findings suggest a practical path for capability transfer: large models can "teach" smaller ones to summarize effectively without relying on expensive human efforts. In addition, the synthetic metric and evaluation criteria streamline the training and evaluation pipeline, offering an end-to-end synthetic solution, as the small model never has to see human data at all.

4.2. Question-Answering Task

As with the summarization task, over the course of training, the performance metric improves for both synthetic and human data settings, together with the decreasing training loss values.

While, in summarization, the models trained on synthetic data more clearly outperform those trained on human data throughout fine-tuning, the landscape for the Q&A task is more nuanced. In particular, the capability metric trajectories for the models trained in human-generated data appear to lag behind those for the synthetic models (see Figure 4, rather than plateauing at lower performance in the summarization setting as shown in Figure 1. Thus, if we assume

that the synthetic metric is sound, then there appears to be a difference in the nature of the two tasks: while summarization is much easier to learn from synthetic data than from human-generated data, it appears that human-generated data may not be strictly worst than synthetic data, albeit slower to learn from.

When looking at the gap between models' best capability performance, Figure 5(a) shows that as model size increases, the discrepancy in performance between synthetic data and human-generated data decreases until essentially negligible for the large model. This scaling effect suggests that while synthetic data might initially be easier to learn from for the smaller models, larger models have enough capacity to glean signal from the human summaries and questions, cutting through the noise of the data. Unlike for the summarization task, it is no longer true that any model trained on synthetic data perform better than all models trained on human data (Figure 5(a)). However, we still see pairwise improvement in metric performance for synthetic data models compared to human data models for equal model size. **When dealing with smaller student models that struggle to learn from sparse signals, synthetic data can still serve as a strong jumpstart to capability acquisition**, given the improved efficiency exemplified by the trajectories. Also, given that the synthetic data always performs on par with human data, synthetic data can offer a much cheaper alternative to human annotation.

4.3. Key Insights and Limitations

For both tasks, we examined the benefits for training models on the synthetic data compared to human-generated data.

One caveat in our experimental results is that we cannot directly compare the rate of improvement in summarization to that in Q&A due to their fundamentally different output modalities and evaluation criteria. The output variance is much higher for the summarization task than the Q&A task: a small improvement in the model may manifest itself much more clearly in the Q&A task compared to the summarization task. Thus, while the keyword metric may not capture all the improvements made by Seq2Seq model throughout training, it is also important to note this fundamental paradigm difference in the two tasks.

In conclusion, our results emphasize that synthetic data can serve as a powerful and efficient alternative to human-annotated data to transfer capabilities to smaller models, especially in tasks where the evaluation metric is tightly aligned with the synthetic data. While our synthetic metric may create some bias in our experimental evaluation towards models trained on synthetic data, the results point to promising opportunities where synthetic data can provide a valuable basis for fine-tuning, with impressive performance on small models and continued scaling of capability transfer as we increase model size. This suggests that an end-to-end synthetic approach, both for data generation and evaluation, can indeed facilitate capability transfer from large to small models, though the magnitude and ease of this transfer may depend on the complexity of the task and the representational capacity of the student model.

4.4. Future Work

While small model scale was varied in this study, the large model can also be scaled to assess how the effectiveness of the large model at producing learnable and useful synthetic data changes with model capacity. These experiments could help us empirically determine a "scaling law" and design space for capability transfer. Furthermore, when sampling from both the small model and the large model, changing the temperature and the decoding strategy can be explored to assess their impact on performance on the synthetic metric.

While the two tasks used in our experiments provide a proof-of-concept for capability transfer for natural language tasks, many more tasks can be explored in similar synthetic settings. These may include mathematical and analytical settings such as proofs and algebraic manipulations, translation tasks, and logical reasoning.

5. Conclusion

We conducted an empirical investigation of whether large models can transfer capabilities to small models with synthetically-generated data through two tasks, summarization and question-answering. Using the GPT-4o-mini model as the large model to fine-tune three small T5 models, we found that models fine-tuned on synthetic data outperformed those trained on human-generated data by up to 56% on the GPT-generated evaluation metric. On the Q&A task, models trained on synthetic data reached on-par performance with those trained on human-annotated data. Thus, our proof-of-concept tasks demonstrate that synthetic data can serve as a promising, cost-efficient alternative to human annotation for transferring capabilities from large models to small models, reducing the dependence on humans in the loop to produce high-quality data. By leveraging large models as data generators and evaluators, we enable more cost-efficient and customizable training pipelines, paving the way for smaller, high-performing models to be deployed in resource-constrained settings.

6. Impact Statement

This research into transferring capabilities from large language models (LLMs) to smaller models using synthetically generated data presents a significant step towards more accessible, cost-effective, and efficient AI. Positive societal impacts include the democratization of AI by enabling smaller

models on less powerful hardware, reduced development costs by replacing human annotation with synthetic data, a smaller environmental footprint due to increased model efficiency, and enhanced customization for specific tasks. However, potential negative impacts should also be considered, such as the propagation and amplification of biases from the "teacher" LLM to the "student" model. By thoughtfully developing evaluation methodologies, this approach holds great promise for responsibly unlocking substantial benefits and fostering a future where AI is both powerful and equitably deployed.

References

Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. URL http://arxiv.org/abs/1506.03340.

Kaddour, J. and Liu, Q. Text data augmentation in low-resource settings via fine-tuning of large language models. *arXiv preprint arXiv:2310.01119*, 2023.

Longpre, S., Hou, L., Vu, T., Webson, A., Chung, H. W., Tay, Y., Zhou, D., Le, Q. V., Zoph, B., Wei, J., et al. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning*, pp. 22631–22648. PMLR, 2023.

Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford alpaca: An instruction-following llama model, 2023.

Wang, Y., Kordi, Y., Mishra, S., Liu, A., Smith, N. A., Khashabi, D., and Hajishirzi, H. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

Xu, Y., Xu, R., Iter, D., Liu, Y., Wang, S., Zhu, C., and Zeng, M. Inheritsumm: A general, versatile and compact summarizer by distilling from gpt. *arXiv preprint arXiv:2305.13083*, 2023.

A. Appendix

A.1. Related Work

Knowledge distillation from large language models to smaller models has emerged as a promising approach for making advanced AI capabilities more accessible and efficient. Previous studies have demonstrated various techniques for transferring capabilities from teacher models to student models through synthetic data generation and instruction tuning. For example, Taori et al. (2023) showed that instruction-following capabilities could be effectively distilled from GPT models to smaller open-source models through supervised fine-tuning on generated data. Wang et al. (2022) pioneered self-instruct methods where teacher models generate instruction-following examples to train student models. Our work builds on these approaches but focuses specifically on investigating capability transfer for summarization and question-answering tasks using purely synthetic for both data generation and evaluation.

Recent work has also explored specialized distillation approaches for key NLP tasks like summarization and question answering. For summarization, Xu et al. (2023) demonstrated that knowledge from large models like GPT-3.5 can be effectively distilled into smaller models while maintaining strong abstractive summarization capabilities across zero-shot, few-shot and supervised settings (trained > 7 epochs, dataset size > 6 million). Their INHERITSUMM model (trained on ZCode++) achieved comparable ROUGE scores to GPT-3.5 while being significantly more compact and efficient. For distillation in low-resource settings, Kaddour & Liu (2023) showed that synthetic training data generated by finetuned teacher LLMs can dramatically improve downstream summarization and question answering performance of smaller student models (trained up to 320 epochs), even when only minimal labeled examples are available for teacher finetuning. Our study differs in several key aspects: synthetic evaluation, model sizes, and training scale. We explore synthetic evaluation metrics generated by the teacher model itself to assess capability transfer. We also investigate distillation across a broader range of small model sizes that includes even smaller models (80M - 780M parameters), providing insights into how distillation effectiveness scales with model capacity. The other two studies trained on datasets up to 2 magnitudes larger in dataset size and training epochs. Our study analyzes improvements in performance after training for only 1 epoch and on smaller datasets (28000 for summarization and 40000 for Q&A).

A.2. Experimental Details

We selected FLAN-T5 as the small trained on the synthetic data (Longpre et al., 2023). The pre-trained models are available in various sizes on HuggingFace. We used three model sizes: `small` (80M parameters), `base` (250M parameters), and `large` (780M parameters). For both the summarization and Q&A synthetic datasets, we trained the models on 80% of the data (approximately 28,000 for summarization and approximately 40,000 for Q&A).

For all experiments, we trained the model for 1 epoch, e.g. fine-tuning on each training example once. We set the batch size as 8 and saved a checkpoint every 40 steps, resulting in 21 checkpoints for summarization and 31 checkpoints for Q&A.

For generating the evaluation metrics, inference was performed on each model with the full-text articles for summarization and the GPT-generated summaries and questions for Q&A. The models were prompted in the same way as the data generation process. For the summarization task, the generated summary was scored on whether the summary included the five keywords. For the question-answering task, the generated answer was scored on accuracy.

A.3. Complimentary Figures for Synthetic Data Generation

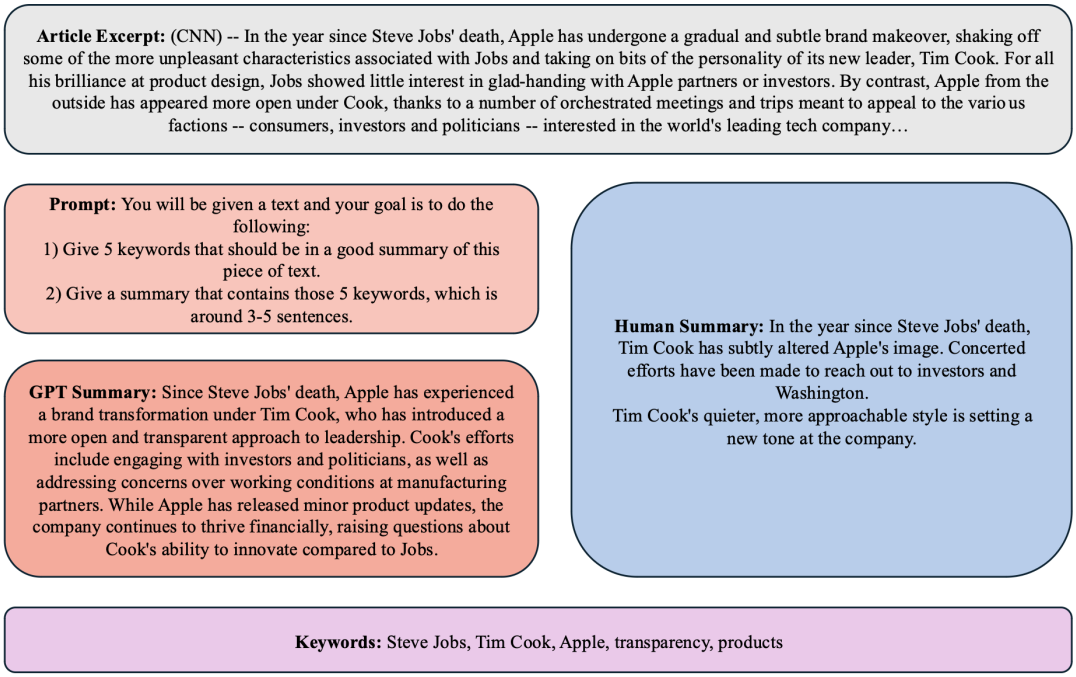


Figure 6. Overview of synthetic data generation for summarization task. The article excerpt is taken from the CNN articles dataset. When given the shown prompt, the large model generates the GPT summary and the keywords.

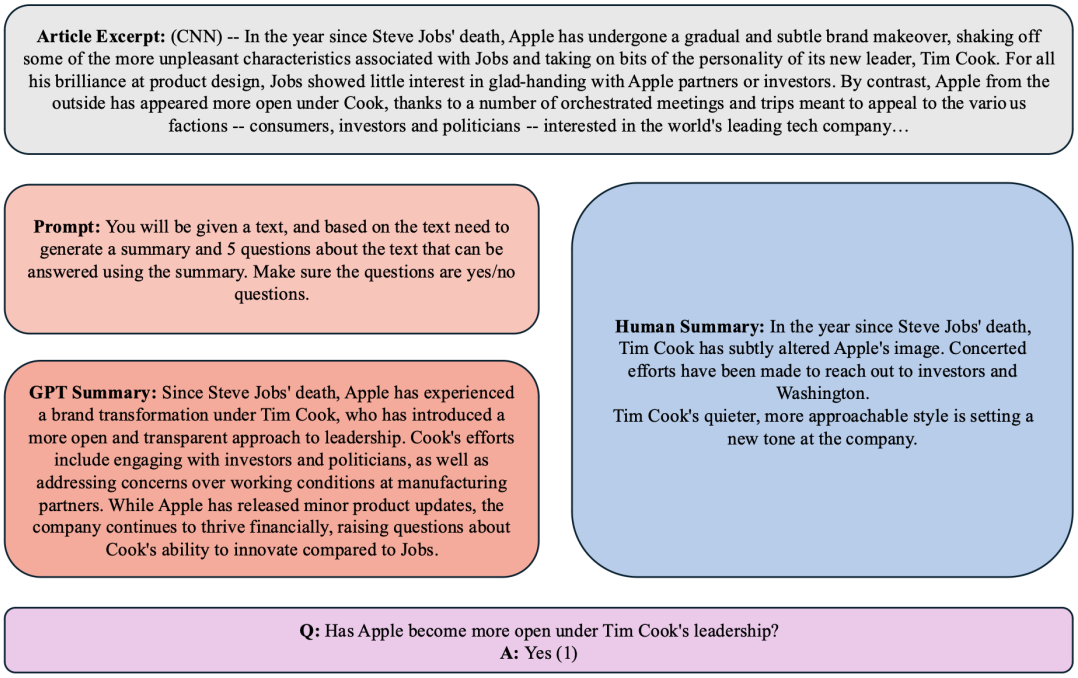


Figure 7. Overview of Q&A data generation. Given the CNN article, the large model is prompted to generate the summary shown and five question/answer pairs, such as the question shown above.

A.4. Algorithm for Capability Transfer Fine-tuning and Evaluation

Algorithm 1 Capability Transfer via Synthetic Data

Input: Large Model L, Small Pre-trained Model S
Output: Fine-tuned small model S'

```
// Step 1: Generate synthetic dataset using the large model
D ← ∅ // Initialize dataset
for T in Texts do
    // Generate task-specific outputs using the large model
    if Task == "Summarization" then
        Summary, KeyWords ← GenerateSummary(L, T) D.append(T, Summary, KeyWords)
    else if Task == "Q&A" then
        // Generate summary and QAs
        Summary, QAs, ← L.GenerateQA(T) // Split QAs for training and evaluation
        TrainQA, EvalQA ← split(QAs, train, eval) D.append(Summary, TrainQA, EvalQA)
// Step 2: Fine-tune small model S on the synthetic dataset
for (x, y, meta) in D do
    if Task == "Summarization" then
        // Compute loss based on summary alignment and key words
        ℓ ← compute_loss(S(x), y, meta.KeyWords)
    else if Task == "Q&A" then
        // Train on the summary and questions
        ℓ ← compute_loss(S(meta.Summary, meta.TrainQA), meta.EvalQA)
    // Backpropagate and update S using ℓ
    S' ← train(S, ℓ)
// Step 3: Evaluate the fine-tuned small model
for (x, y, meta) in D_test do
    if Task == "Summarization" then
        // Evaluate performance based on key word coverage
        performance ← evaluate((S'(x), meta.KeyWords))
    else if Task == "Q&A" then
        // Evaluate on unseen yes/no questions
        performance ← evaluate(S'(meta.Summary, meta.EvalQA))
return S'
```

A.5. Small Model-Generated Summaries

Table 1. T5 Model Outputs Across Fine-tuning Steps with Synthetic and Human-Generated Data

GPT Summary		Since Steve Jobs' death, Apple has experienced a brand transformation under Tim Cook, who has introduced a more open and transparent approach to leadership. Cook's efforts include engaging with investors and politicians, as well as addressing concerns over working conditions at manufacturing partners. While Apple has released minor product updates, the company continues to thrive financially, raising questions about Cook's ability to innovate compared to Jobs.	
Keywords		Steve Jobs, Tim Cook, Apple, transparency, products	
Model	**Step**	**Synthetic Data**	**Human-Generated Data**
t5-small	0	Tim Cook's new CEO, Tim Cook, has been a snobby and snobby, but he's still a snobby and snobby.	
	240	Tim Cook has been a key figure in the company's evolution, focusing on the personality of its new leader, Tim Cook. He has been a key figure in the company's evolution, focusing on the importance of a more relaxed and approachable personality. The company has a significant impact on Apple and iPhone sales, and the impact of Jobs' absence will not be known for two years or more.	mente: Apple has a huge amount of goodwill, but it's not the best company in history. Apple has been the most valuable company in history, but it has been the most valuable company in history.
	840	Steve Jobs' legacy has been a significant influence on Apple, highlighting the importance of a more relaxed approach to product design. His recent shift in tone has prompted significant discussions about Apple's future, highlighting the importance of a more transparent approach to product development. The recent changes in Apple's leadership have sparked controversy, highlighting the need for more transparency and a more transparent approach to product development.	Apple's new leader Tim Cook has been in charge of Apple's supply chain. Cook has not launched any bold new products in the past 12 months. Apple has only released minor updates of popular products, like speedier laptops, the new Mac operating system, the refreshed iPod line and the new iPhone 5.
t5-base	0	Apple's new chief executive Tim Cook has been compared to Steve Jobs after every presentation, success and failure.	
	240	Apple has undergone a gradual brand makeover since Steve Jobs' death, focusing on its new leader, Tim Cook. Cook has embraced a more open personality, focusing on a variety of meetings and trips to attract various factions. Cook's recent actions have largely reflected the company's positive attitude towards the company, highlighting the company's ability to innovate and innovate. However, Cook's lack of innovation has raised concerns about his ability to create innovative products, particularly in the face of significant challenges.	Apple has undergone a gradual and subtle brand makeover. Tim Cook has been incessantly compared to his predecessor after every presentation, successes and failures. Apple has about 12,000 employees in the United States, not counting retail and support staff.
	840	Apple has undergone a subtle brand makeover since Steve Jobs' death, focusing on its new leader, Tim Cook. Cook has embraced a more open personality, focusing on addressing issues like the bungled iOS 6 Maps app and addressing issues like the bungled iOS 6 Maps app. While Cook has not launched any significant new products, he has emphasized the importance of creativity and innovation in Apple's product development process	Apple has undergone a gradual and subtle brand makeover. Tim Cook has taken on bits of the personality of its new leader. Cook has been incessantly compared to his predecessor after every presentation, successes and failures. Apple has about 12,000 employees in the United States, not counting retail and support staff.
t5-large	0	In the year since Steve Jobs' death, Apple has undergone a gradual and subtle brand makeover, shaking off some of the more unpleasant characteristics associated with Jobs and taking on bits of the personality of its new leader, Tim Cook.	
	240	Apple has undergone a gradual brand makeover since Steve Jobs' death, embracing the personality of Tim Cook and embracing transparency. Despite the company's recent success, Cook has faced criticism for his lack of innovation and the recent Maps app controversy, raising concerns about his ability to manage the legacy of Steve Jobs. While Cook has not launched any bold new products, he has shown a more relaxed approach to product announcements, highlighting the company's commitment to transparency and goodwill.	Apple has undergone a gradual and subtle brand makeover since Steve Jobs' death. Tim Cook has taken on bits of the personality of Apple's new leader. Cook has not launched any bold new products in the past 12 months.
	840	In the year since Steve Jobs' death, Apple has undergone a gradual brand makeover, embracing the personality of Tim Cook and embracing a more open approach to communication. Despite the recent Maps app controversy, Cook has shown a more relaxed approach, addressing issues like working conditions and improving conditions at manufacturing partners. However, the company's recent success has not been as impressive as previously expected, with only minor updates and a few new products, raising questions about the future of Apple's leadership.	Apple has undergone a gradual and subtle brand makeover since Steve Jobs' death. Tim Cook has taken on bits of the personality of Apple's new leader. Cook has not launched any bold new products in the past 12 months.

A.6. Synthetic Data Generation Prompts

A.6.1. SUMMARIZATION

You will be given a text and your goal is to the following:

1) Give 5 keywords that should be in a good summary of this piece of text.

2) Give a summary that contains those 5 keywords, which is around 3 sentences.

Make your summary around 3 sentences. You should format your answer as follows where you replace the words in quotes with your answers.

[keywords] `keyword1`, `keyword2`, `keyword3`, `keyword4`, `keyword5`

[summary] `YOUR SUMMARY HERE`

A.6.2. Q&A

You will be given a text, and based on the text need to generate a summary and 5 questions about the text that can be answered using the summary. It is very important that the questions can be answered accurately only using the summary. Make sure the questions are yes/no questions.

Give your response in the following format.

[summary] `YOUR SUMMARY HERE`

[q1] `Question 1`

[a1] `Answer 1`

[q2] `Question 2`

[a2] `Answer 2`

[q3] `Question 3`

[a3] `Answer 3`

[q4] `Question 4`

[a4] `Answer 4`

[q5] `Question 5`

[a5] `Answer 5`