Impossible Distillation for Paraphrasing and Summarization: How to Make High-quality Lemonade out of Small, Low-quality Models

Jaehun Jung[†] **Peter West**[†] **Liwei Jiang**[†] **Faeze Brahman**[†‡]
Ximing Lu[†] **Jillian Fisher**[†] **Taylor Sorensen**[†] **Yejin Choi**[†‡]
[†]Paul G. Allen School of Computer Science & Engineering, University of Washington
[‡]Allen Institute for Artificial Intelligence
hoony123@cs.washington.edu

Abstract

We present IMPOSSIBLE DISTILLATION, a novel framework for paraphrasing and sentence summarization, that distills a high-quality dataset and model from a low-quality teacher that itself cannot perform these tasks. Unlike prior works that rely on an extreme-scale teacher model (*e.g.,* GPT3) or task-specific architecture, we hypothesize and verify the *paraphrastic proximity* intrinsic to pre-trained LMs (*e.g.,* GPT2), where paraphrases occupy a proximal subspace in the LM distribution. By identifying and distilling generations from these subspaces, IMPOSSIBLE DISTILLATION produces a high-quality dataset and model even from GPT2-scale LMs. We evaluate our method on multiple benchmarks spanning unconstrained / syntax-controlled paraphrase generation and sentence summarization. Our model with 770M parameters consistently outperforms strong baselines, including models distilled from ChatGPT, and sometimes, even ChatGPT itself. Also, we find that our distilled dataset from 1.5B LMs exhibits higher diversity and fidelity than up to 13 times larger datasets.

1 Introduction

Training a compact, yet performant model is a non-trivial challenge in modern NLP, even for classical tasks such as paraphrase generation and sentence summarization. While large-scale, high-quality data is central to this goal, human supervision is hard to scale; as such, research efforts have focused on training models with an unsupervised, automatically generated dataset. Common approaches include back-translation (Wieting and Gimpel, 2018) and auto-encoding (Févry and Phang, 2018), but are often limited in terms of corpus diversity and noisiness (Hu et al., 2019a,b).

Alternatively, recent works propose to train a compact task model by distilling knowledge from gigantic language models (LLMs) (West et al., 2022). As LLMs such as GPT3 – often multi-billion scale and instruction-tuned – are already competent in paraphrasing and summarizing sentences (Cegin et al., 2023), a specialized model can be trained by simply imitating LLM generations (Cegin et al., 2023; Xu et al., 2023). Despite with limitations (*e.g.,* significant budget requirement for data collection), LLM distillation outperforms previous methods without human supervision, giving out an impression that powerful teacher LM is all we need to train a better student.

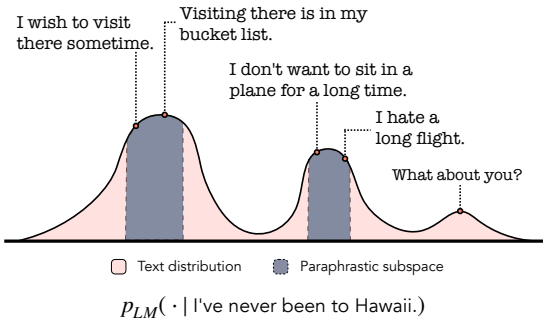

Figure 1: IMPOSSIBLE DISTILLATION develops upon *paraphrastic proximity*: LM's tendency to encode paraphrases on a proximal subspace in its distribution.

In this work, we envision a seemingly impossible alternative to LLM distillation: instead of an extreme-scale, frontier LLM (*e.g.,* GPT3), can we start off with a small, off-the-shelf LM that itself cannot perform paraphrase generation or sentence summarization? We present IMPOSSIBLE DISTILLATION, a novel framework to distill task-specialized dataset and model from GPT2-scale LMs. Our framework requires neither a strong LLM nor human-authored references, yet can distill high-quality paraphrases and summaries comparable to that of prompting the strongest LLMs.

The key observation behind our framework is that a sentence and its paraphrases tend to lie on a proximal subspace in the pretrained LM distribution – a property we call *paraphrastic proximity* (Fig. 1). In other words, by effectively reduc-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 4439–4454

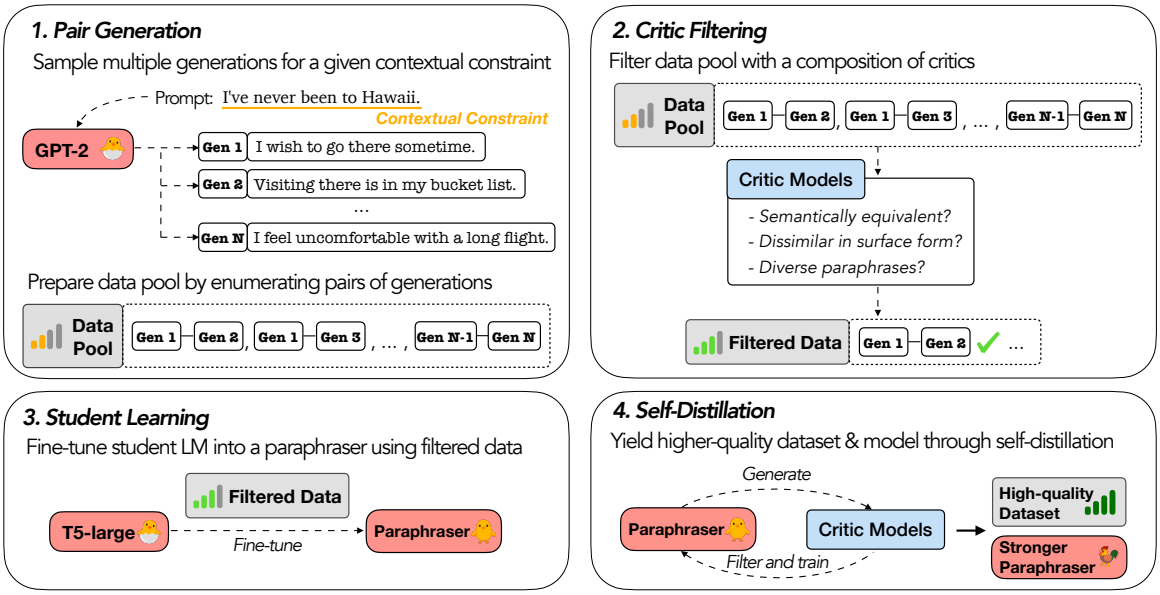

Figure 2: Overview of IMPOSSIBLE DISTILLATION. Starting from low-quality LM (GPT2), we generate a data pool of input-output pairs leveraging paraphrastic proximity, filter it with off-the-shelf critics, and distill a student model on this data pool. By self-distilling the student model, we obtain a high-quality dataset and model for target task.

ing down the LM search space (*e.g.,* by constraining the model with an informative context) toward the paraphrastic subspaces, we can encourage the model to generate multiple sequences that paraphrase each other. As shown in Fig. 2, we leverage this property by first constructing a data pool of *(source, paraphrase)* pairs by enumerating a batch of generations sampled given the context. Next, we filter the data pool with off-the-shelf critic models to keep only the pairs with high quality paraphrases, which we subsequently use to fine-tune a student LM. Finally, the student LM is further refined through self-distillation, where the model is trained on its own high-quality paraphrases; as a result, we obtain both a high-quality corpus and a compact, yet powerful model for paraphrasing. Moreover, as IMPOSSIBLE DISTILLATION is grounded on the explicit evaluation of generated pairs, the framework generalizes to sentence summarization by simply re-defining the filters.

Experimental results show that IMPOSSIBLE DISTILLATION is surprisingly effective, both in terms of the distilled data quality and model performance. We first evaluate the quality of our dataset by measuring the semantic fidelity, lexical diversity, and syntactic diversity against three state-of-the-art paraphrase corpora. We find that our dataset, as a purely synthetic corpus generated from 1.5B LMs, shows better metrics in all measures than state-of-the-art datasets: ParaBank (Hu et al., 2019a) that is 13 times larger than ours and ChatGPT-Para (Vorobev et al., 2023) generated by orders of magnitude larger ChatGPT. Furthermore, in benchmarks across three distinct tasks – unconstrained / syntax-controlled paraphrasing and sentence summarization, our model distilled from 1.5B LM outperforms competitive baselines, including both the task-specific methods and the models distilled from ChatGPT (OpenAI, 2022). In human evaluation, our model with 770M parameters is consistently preferred to the ChatGPT-distilled model, and sometimes, even ChatGPT itself.

2 Paraphrastic Proximity

We develop IMPOSSIBLE DISTILLATION based on the observation of *paraphrastic proximity – i.e.* when the LM decoding space is constrained with sufficiently informative context, the model can produce multiple generations that paraphrase each other. Notably, Meng et al. (2021) indirectly leverages paraphrastic proximity in *context LMs* – a set of encoder-decoder transformers pre-trained from scratch using specialized training objectives. We, on the other hand, show that paraphrastic proximity holds for off-the-shelf LMs such as GPT2, which we can make use of to distill a high-quality task model and dataset. In this section, we first verify this with GPT2-XL (Radford et al., 2019).

While the exact distribution of all possible generations is intractable, we can obtain an approximation by sampling a large number of generations given a contextual constraint. Concretely, we

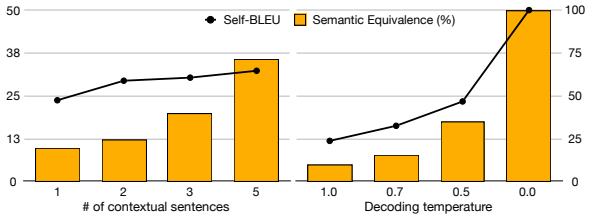

Figure 3: How paraphrastic are GPT2-XL generations? We compute the ratio of semantically equivalent pairs and their average Self-BLEU.

first sample 1000 context (each with 1-5 sentences) from news articles in XSUM dataset (Narayan et al., 2018). Then we prompt GPT2 to generate 100 next sentences per each contextual constraint. To evaluate whether these sentences are indeed paraphrastic to each other, we measure their (1) pair-wise semantic equivalence, and (2) surface-form dissimilarity. For semantic equivalence, we employ an off-the-shelf NLI model (Liu et al., 2022a), and determine a pair (x, y) to be semantically equivalent if entailment holds in both directions. For surface-form dissimilarity, we compute the Self-BLEU (Zhu et al., 2018) between sentences, for only the semantically equivalent pairs.

The results are presented in Fig. 3. In the left figure, as the context becomes longer (*i.e.* more informative), the generated sentences are more likely to be semantically equivalent, verifying our assumption. Notably, the high semantic equivalence does not come from merely generating sentences with similar surface form – even with longer context, the average pair-wise Self-BLEU is around 32[1]. The results indicate that GPT2 can generate a large number of paraphrases simply by over-sampling multiple completions to the given context.

On the right side of Fig 3, we also gauge the paraphrastic proximity under various decoding temperatures. Here, the ratio of semantically equivalent pairs dramatically increases as the temperature decreases. Low temperature adjusts the sampling distribution to be more skewed towards regions with high probability mass, hence allowing paraphrases to be more easily sampled from these subspaces. However, it is important that the temperature balance the trade-off between sample efficiency and diversity; when the temperature is too low, the generated sentences are almost identical and hence does not qualify as desirable paraphrase.

[1]The average Self-BLEU of human-authored paraphrases in MRPC (Dolan and Brockett, 2005) dataset is 39 (Herbold, 2023).

3 Impossible Distillation

IMPOSSIBLE DISTILLATION starts from an off-the-shelf teacher LM $\mathcal{M}_T$, and distills its knowledge into a student LM $\mathcal{M}_S$, yielding a specialized model $\mathcal{M}_{task}$ for paraphrasing and sentence summarization. As a byproduct of this process, we also obtain a high-quality dataset $\mathcal{D}_{task}$. Below, we detail the process focusing on paraphrase generation as the task of interest, then discuss how this generalizes to sentence summarization.

3.1 Pair Generation

We first generate a large pool of candidate *(source-paraphrase)* pairs $\mathcal{C}_T = \{(x_1, y_1), \cdots, (x_{|\mathcal{C}_T|}, y_{|\mathcal{C}_T|})\}$ from an off-the-shelf teacher $\mathcal{M}_T$. Our first step is to prepare contextual constraints c_i, by sampling 1-5 sentences from $\mathcal{M}_T$:

$$c_i \sim p_{\mathcal{M}_T}(\cdot)$$

The contextual constraints could be generated either unconditionally or conditioned on a simple prompt (Appendix A). Alternatively, one could sample contextual constraints from human-written corpus (as done in §2). While manually collecting contextual constraints allows fine-grained control over the generation style and domain, we show that LM-generated context suffices to yield a highly diverse and domain-specific data pool without resorting to an external source of data (§4.1).

Next, we generate a batch of next sentences conditioned on each c_i, then enumerate candidate pairs as the combinations of these sentences:

$$\{s_{i1}, \cdots, s_{ik}\} \sim p_{\mathcal{M}_T}(\cdot|c_i; \tau_{temp})$$
$$\mathcal{C}_i = \{(s_{im}, s_{in})|m, n \in [1, k], m \neq n\}$$

Concretely, we set $k = 100$, generating 100 samples per c_i using Nucleus-Sampling (Holtzman et al., 2020). Based on our preliminary experiments in §2, we set the decoding temperature $\tau_{temp} = 0.7$ to balance the diversity and sample efficiency of the generated pairs. Collecting the pairs across all c_is, we obtain the data pool $\mathcal{C}_T = \bigcup_i \mathcal{C}_i$.

3.2 Filtering with Critics

Despite producing a large population of valid paraphrases, our pair generation process is noisy in nature, as it enumerates all possible pairs of generated sentences. For example, Generation 1 (*I wish to visit there sometime.*) and Generation N (*I hate a long flight.*) in Fig. 2 will constitute a pair in the data pool, although the two sentences have no

logical relevance. A crucial step, therefore, is to filter out suboptimal pairs from the data pool and ensure the quality of the distilled dataset.

Semantic Equivalence Filter A faithful paraphrase should preserve the semantics of the source statement without hallucinating unsupported content. NLI models are well-suited to quantify this relationship, as they are trained to infer the logical entailment between an arbitrary pair of statements (Chen et al., 2021). Hence, we define a binary filter using a small NLI model (Liu et al., 2022a) as a critic, and discard the pairs that do not achieve the entailment score over the threshold $\tau_{semantic}$:

$$f_{semantic}(x, y) = \mathbb{1}\Big\{p_{NLI}(x \Rightarrow y) \geq \tau_{semantic} \wedge p_{NLI}(y \Rightarrow x)) \geq \tau_{semantic}\Big\}$$

Dissimilarity Filter A good paraphrase should significantly alter the surface form of the input while preserving its meaning. In IMPOSSIBLE DISTILLATION, the surface-form dissimilarity is achieved by filtering pairs based on (1) the token overlap between sentences and (2) their syntactic difference. For token overlap, we filter the pairs with higher ROUGE-L (Lin, 2004) than a threshold τ_{rouge}. To gauge the syntactic difference, we follow prior works (Kumar et al., 2020) by first parsing the constituency tree of the source and paraphrase, then filtering based on their tree edit distance (TED):

$$f_{dissim}(x, y) = \mathbb{1}\Big\{\text{ROUGE}(x, y) \leq \tau_{rouge} \wedge \text{TED}(x, y) \geq \tau_{TED}\Big\}$$

Intuitively, the two dimensions of dissimilarity complements each other – while ROUGE filter promotes lexical divergence in each pair, TED filter preempts "hacking" the token-overlap metric by simply switching a few words in the source sentence with corresponding synonyms.

Diversity Filter Constructing a high-quality corpus is not just about creating valid input-output pairs; ideally, the corpus should cover a diverse range of style and topic within its samples, as the data diversity directly correlates with the robustness of the trained model (Rebuffi et al., 2021). Our data pool might be limited in this regard, as it includes a large number of pairs from the same context c, often resulting in multiple pairs having similar x or y. To remove the duplicate pairs and promote diversity, we employ an additional critic $f_{diversity}$. Concretely, we define two pairs (x_1, y_1) and (x_2, y_2) to be duplicate when one pair entails another, either on the input side ($x_1 \Rightarrow x_2$) or the output side ($y_1 \Rightarrow y_2$). The diversity filter operates by first grouping all entailing pairs, then discarding all but one with the largest entailment score. In practice, this filter can be efficiently implemented using graph traversal; we describe the formal algorithm in Appendix A.

Incorporating all critics, we filter the candidate pool $\mathcal{C}_T$ into a refined dataset $\mathcal{D}_T$ as following:

$$\mathcal{D}_T = \{(x, y) | (x, y) \in \mathcal{C}_T, f_{semantic} \wedge f_{dissim} \wedge f_{diversity}(x, y) = 1\}$$

3.3 Distilling Student Model

Now that we extracted the paraphrastic knowledge of the teacher $\mathcal{M}_T$ into a dataset $\mathcal{D}_T$, we use the data to fine-tune the student model into a paraphrase generation model. The student model $\mathcal{M}_S$ is fine-tuned by maximizing $\mathbb{E}_{(x,y)\sim\mathcal{D}_T}[\log p_{\mathcal{M}_S}(y|x)]$, i.e. the conditional log-likelihood of y given x.

Next, the paraphrasing capability of the student is further amplified through self-distillation, by fine-tuning on its own generated high-quality paraphrases. We first sample the input sentence x from the teacher LM $\mathcal{M}_T$, then generate paraphrase y by feeding x into $\mathcal{M}_S$:

$$\mathcal{C}_S = \{(x_1, y_1), \cdots | x_i \sim p_{\mathcal{M}_T}(\cdot|c_i);\ y_i \sim p_{\mathcal{M}_S}(\cdot|x_i)\}$$

Using the same critics as in the previous stage, we filter $\mathcal{C}_S$ to obtain a high-quality dataset $\mathcal{D}_{para}$. Finally, we fine-tune $\mathcal{M}_S$ on $\mathcal{D}_{para}$, yielding the end-stage model $\mathcal{M}_{para}$. Consistent with prior findings on self-distillation (Pham et al., 2022; Allen-Zhu and Li, 2020), this simple process significantly improves the performance of our task model, as confirmed by our ablation study (§4.5). In addition, our self-distillation outputs a large-scale, standalone dataset $\mathcal{D}_{para}$ that can be evaluated and reused, *e.g.,* to directly train a paraphrasing model without re-iterating the distillation procedure.

3.4 Endowing Controllability

Recent works emphasize the importance of syntactic control in paraphrase generation, allowing the model to generate an output paraphrase tailored to users' need (Chen et al., 2019). In IMPOSSIBLE DISTILLATION, endowing controllability to the student model is straightforward. We first prepare the dataset $\mathcal{D}_{control}$ by parsing the constituency tree t of each paraphrase y in $\mathcal{D}_{para}$:

$$\mathcal{D}_{control} = \{(x, y, t) | (x, y) \in \mathcal{D}_{para}, t = \text{parse}(y)\}$$

Dataset (# Instances)	Semantic Similarity	Lexical Diversity				Syntactic Diversity	
	Cosine Sim. ↑	H_2 ↑	H_3 ↑	MSTTR ↑	Jaccard Sim. ↓	TED-3 ↑	TED-F ↑
ParaBank1 (57.0M)	81.77	17.07	21.66	45.52	48.41	3.59	14.53
ParaBank2 (19.7M)	82.50	17.48	21.44	46.16	**43.44**	4.04	17.41
ChatGPT-Para (2.1M)	85.44	17.67	21.41	35.83	44.56	4.26	20.15
DIMPLE (4.2M)	**87.68**	**17.75**	**22.46**	**53.08**	43.62	**5.02**	**29.84**

Table 1: Quality comparison between paraphrase datasets. DIMPLE, as a purely synthetic corpus generated from 1.5B LMs, exhibits better diversity compared to others, including the dataset constructed by prompting ChatGPT.

Then a controllable model $\mathcal{M}_{control}$ can be trained, by fine-tuning $\mathcal{M}_S$ to generate y given the source x and the tree t.

3.5 DIMPLE and Impossible-T5

To test IMPOSSIBLE DISTILLATION in both general and domain-specific paraphrasing, we use two teacher LMs – GPT2-XL and BioGPT (Luo et al., 2022; for biomedical domain), all with 1.5B parameters. We use T5-large (Raffel et al., 2020) with 770M parameters as our student LM, and train it with 400k filtered pairs distilled from the teacher models. After self-distilling the student model, we yield a specialized paraphrase generation model we call **Impossible-T5**, along with a large-scale corpus with 4M high-quality pairs (2M for general domain and 2M for biomedical domain). We name this dataset **DIMPLE**[2]. Additional implementation details such as generation parameters and filter thresholds are provided in Appendix A.

3.6 Sentence Summarization

In addition, we generalize IMPOSSIBLE DISTILLATION for abstractive sentence summarization, a task akin to paraphrasing but with different goals (Zhou and Rush, 2019). Whereas paraphrase generation searches for an alternative form of the source sentence while preserving all its information, summarization aims for a succinct representation of the given sentence, at the expense of losing tangential details. In our method, the distillation process is grounded on the explicit evaluation of generated pairs, hence the framework can generalize to summarization by simply redefining the filters. For example, one could add a length filter to the critics, guaranteeing that the output is strictly shorter than the input in the filtered pairs. In Appendix C, we describe the details of the generalized pipeline for sentence summarization.

[2]**D**ataset of **Imp**ossib**l**e Paraphrases.

4 Experiments

4.1 Dataset Evaluation

Evaluation Setup First, we directly compare the quality of DIMPLE against three large-scale paraphrase corpora: ParaBank1 (Hu et al., 2019a), ParaBank2 (Hu et al., 2019b), and ChatGPT-Para (Vorobev et al., 2023). Both ParaBank1 and ParaBank2 are based on back-translation; ParaBank1 imposes lexical constraints to promote the diversity of paraphrases, and ParaBank2 additionally clusters and resamples generations to further improve the syntactic diversity. ChatGPT-Para is a dataset distilled from ChatGPT, by instructing the LLM to paraphrase sentences from Quora (Sharma et al., 2019), SQUAD 2.0 (Rajpurkar et al., 2018) and CNN/DM (Nallapati et al., 2016).

Following Huang et al. (2023), we measure the semantic similarity, lexical diversity and syntactic diversity of each dataset. For semantic similarity, we compute the average cosine-similarity between source and paraphrase measured by SimCSE (Gao et al., 2021). To estimate lexical diversity, we use 2/3-gram entropy, mean-segmented token type ratio (MSTTR; Torruella and Capsada, 2013), and the token-level Jaccard similarity between source and paraphrase. For syntactic diversity, we compute the average pairwise tree edit distance, either for the top 3 layers (TED-3) or the full tree (TED-F).

Results The results are shown in Table 1. In all 3 dimensions, DIMPLE consistently outperforms all baseline datasets. Notably, this includes ParaBank1 that is more than 13 times larger than DIMPLE, demonstrating the sample efficiency of our framework in extracting diverse paraphrastic knowledge. In addition, the superior results of our dataset compared to ChatGPT-Para implies that the scale of the LM is not the only factor that determines the quality of generated data. By effectively constraining the LM search space and filtering pairs with a composition of critics, IMPOSSIBLE DISTILLA-

Dataset	MRPC				ParaNMT-small				ParaSCI-arXiv			
Model	*iBLEU*	*B-iB*	*BLEU*	*R-L*	*iBLEU*	*B-iB*	*BLEU*	*R-L*	*iBLEU*	*B-iB*	*BLEU*	*R-L*
Copy-Input	1.1	0.0	**44.4**	**65.7**	-13.7	0.0	**23.3**	**38.2**	0.0	0.0	**42.8**	**60.2**
GPT-3.5	5.7	66.5	18.1	35.7	6.2	79.6	14.1	32.2	**4.0**	63.2	16.2	37.5
ChatGPT	**5.9**	**67.7**	18.0	36.5	**6.7**	**81.3**	13.4	32.8	3.6	**63.9**	16.4	36.1
$T5_{ParaBank1}$	5.8	60.0	27.3	50.3	4.4	70.3	19.2	36.5	3.1	53.3	26.2	48.6
$T5_{ParaBank2}$	6.1	60.5	27.7	51.4	5.3	71.5	19.7	37.6	3.7	55.1	24.9	48.2
$T5_{ChatGPT-Para}$	5.4	62.9	21.2	40.7	5.2	74.4	15.2	33.5	3.8	59.6	23.8	41.0
Impossible-T5	**7.3**	**67.1**	26.0	46.8	**5.9**	**79.7**	16.3	31.8	**4.3**	**64.5**	25.6	44.9

Table 2: Experimental results of Impossible-T5 and baselines on unconstrained paraphrase generation. Impossible-T5 outperforms the same size model trained on much larger datasets, and is competitive to 175B LLM in both general and domain-specific benchmarks.

Model	Fleunt	Faithful	Dissimilar
ChatGPT	**2.8**	2.37	2.38
$T5_{ParaBank1}$	2.45	2.33	2.09
$T5_{ParaBank2}$	2.47	2.50	2.21
$T5_{ChatGPT-Para}$	2.64	2.30	2.33
Impossible-T5	2.74	**2.55**	**2.40**

Table 3: Human evaluation results (Krippendorff's alpha = 0.62; substantial inter-annotator agreement). To minimize subjectivity, we use strict 3-level scale, where 3 indicates perfect satisfaction, and 1 indicates complete dissatisfaction of the desired property.

TION makes it possible to generate a high-quality dataset from a small, low-quality LM.

4.2 Unconstrained Paraphrase Generation

Evaluation Setup In this section, we evaluate Impossible-T5 in multiple benchmarks for paraphrase generation without syntactic control. We use three human-curated benchmarks spanning general and domain-specific paraphrase generation: MRPC (Dolan and Brockett, 2005), ParaNMT-small (Kumar et al., 2020), and ParaSCI-arXiv (Dong et al., 2021). For baselines, we fine-tune T5-large with ParaBank1, ParaBank2 and ChatGPT-Para, the three large-scale paraphrase corpora analyzed in §4.1. We also consider LLM-based baselines, by zero-shot prompting GPT3.5 (*text-davinci-003*; Ouyang et al., 2022) and ChatGPT.

Results In Table 2, we report BLEU and ROUGE-L (R-L) along with iBLEU (Sun and Zhou, 2012) and BERT-iBLEU (B-iB), a metric known to better correlate with human judgements of paraphrase quality (Niu et al., 2021). Consistent to the prior findings on the brittleness of token-overlap metrics (Zhang et al., 2020), BLEU and ROUGE-L fail to accurately assess the paraphrase quality. In fact, a simple baseline that merely copies the input as an output (Copy-Input) marks state-of-the-art on these metrics, across all datasets.

A clearer tendency is shown with iBLEU and BERT-iBLEU: Impossible-T5 consistently outperforms the same size model trained on order of magnitude larger ParaBank, showing up to 10% relative improvement across all benchmarks. Moreover, Impossible-T5 is the only 770M model comparable to 175B GPT-3.5 across all benchmarks. Notably in an expert domain (ParaSCI), it even outperforms ChatGPT. Additional results against state-of-the-art unsupervised paraphrase generation methods are presented in Appendix B.

Human Evaluation We additionally conduct human evaluation to compare the quality of the generated paraphrases. We generate 200 paraphrases with each model using MRPC corpus, and ask six Mechanical Turk workers to evaluate whether each paraphrase is (1) fluent, (2) faithful to the source, and (3) dissimilar to the source (Appendix D). Table 3 shows the results. Consistent to the quantitative metrics, human annotators prefer paraphrases from Impossible-T5 than the competitive baselines. We find that our model is generally considered to be more faithful to the original statement than ChatGPT while sufficiently altering the surface form. Notably, the high faithfulness and dissimilarity does not come from sacrificing the soundness of generation, marking better fluency score than both $T5_{ParaBank}$ and $T5_{ChatGPT-Para}$.

4.3 Syntactically Controlled Paraphrase Generation

Evaluation Setup Next, we assess Impossible-T5 in syntactically controlled paraphrase generation. We use ParaNMT-small, where each sam-

Model	iBLEU ↑	B-iB ↑	R-L ↑	TED-F ↓
$\text{ChatGPT}_{\text{0-shot}}$	9.1	85.8	41.6	11.6
$\text{ChatGPT}_{\text{5-shot}}$	9.0	85.9	42.2	10.3
$\text{T5}_{\text{ParaBank1}}$	10.7	82.3	55.6	**8.4**
$\text{T5}_{\text{ParaBank2}}$	10.9	84.7	**57.5**	8.8
$\text{T5}_{\text{ChatGPT-Para}}$	10.5	79.4	47.6	10.4
Impossible-T5	**11.2**	**86.6**	51.8	8.5

Table 4: Results on syntactically controlled paraphrase generation. Impossible-T5 outperforms baselines in both paraphrase quality and controllability.

Model	Automatic		Human		
	B-F1	R-L	Fluent	Faithful	Concise
ChatGPT	84.8	**33.6**	**2.55**	2.44	2.32
Referee	78.2	29.2	2.45	2.33	2.41
$\text{T5}_{\text{ParaBank}}$	77.5	29.6	2.21	2.17	1.96
Impossible-T5	**85.1**	30.3	2.46	**2.53**	**2.49**

Table 5: Results on sentence summarization. We report BERTScore-F1 (Zhang et al., 2020) and ROUGE-L for automatic evaluation. In addition, six crowd-source workers qualitatively assessed the 100 summaries per each model with 3-level likert scale.

ple consists of a source x, a syntactic exemplar z, and a paraphrase y of x that follows the syntax of z. Since the controllable version of our model is trained with the constituency tree as input, we first parse z and feed the tree into our model (along with x) during inference. For baselines, we consider T5 trained with existing corpora, additionally annotated with the tree of target paraphrases. We also prompt ChatGPT to generate paraphrase using the same syntax with z.

Results The results are shown in Table 4. Impossible-T5 outperforms baselines across all metrics except ROUGE-L. Notably, the syntax conformity of ChatGPT is substantially poor, even with 5-shot examples of syntax-controlled paraphrases. The results imply that distilling a fine-grained controllable model could be a reasonable alternative to prompting LLM with a textual description of the desired output.

4.4 Sentence Summarization

Evaluation Setup We use Gigaword (Rush et al., 2015), a representative benchmark for sentence summarization. For baselines, we use ChatGPT and Referee (Sclar et al., 2022), an unsupervised summarizer distilled from GPT3. We also train T5-large on $\text{ParaBank}_{\text{summ}}$, a variant of ParaBank2 filtered using the same set of summarization critics as for IMPOSSIBLE DISTILLATION.

Model	BERT-iBLEU
Student model w/o Self-distillation	64.0
$\text{T5}_{\text{ChatGPT-Para}}$	62.9
$\text{T5}_{\text{ChatGPT-Para}}$ + Self-distillation	63.3
$\text{T5}_{\text{ChatGPT-Para}}$ + Critic Filtering	64.1
Impossible-T5	**67.1**

Table 6: Ablation study on MRPC. The best configuration is IMPOSSIBLE DISTILLATION incorporating both critic filtering and self-distillation.

Results The results are as seen in Table 5. We observe that re-purposing a paraphrase corpus for summarization ($\text{T5}_{\text{ParaBank}}$) leads to sub-optimal performance, as the back-translation does not reflect the concise nature of summaries. In contrast, the critic models in IMPOSSIBLE DISTILLATION explicitly participate in the data-generating process, by promoting the model to generate outputs that satisfy the desired properties of critic models. As a result, IMPOSSIBLE DISTILLATION successfully generalizes to summarization, only by plugging in the redefined composition of filters.

4.5 Ablation Study

In Table 6, we conduct an ablation study to analyze the contribution of different components in IMPOSSIBLE DISTILLATION.

Does self-distillation matter? We analyze the contribution of self-distillation in two ways. First, we omit the self-distillation stage in our framework and directly test the student model $\mathcal{M}_S$ distilled from $\mathcal{M}_T$. BERT-iBLEU in this case degrades by 3.1 from Impossible-T5, indicating the importance of self-distillation in amplifying model capability. Next, in order to verify whether self-distillation is a dominant contributor to the performance, we iterate self-distillation on $\text{T5}_{\text{ChatGPT-Para}}$. While the performance of $\text{T5}_{\text{ChatGPT-Para}}$ gets better with iterative distillation, the improvement is relatively small, leading to worse performance than our model without self-distillation. The result confirms that while self-distillation helps in improving the end-stage performance, the diversity of data distilled from teacher model is crucial to fully elicit the student model's capability.

Is it all about critics? At the core of IMPOSSIBLE DISTILLATION are the critic models, filtering the noisy data pool generated from the teacher and aligning it for the target task. Therefore, it would be reasonable to ask whether the performance of Impossible-T5 solely comes from the composition

Paraphrase Generation (General domain)	
Constraint c	As part of the process for the upcoming release of the Android M, Google is also adding a new camera API to the latest Android OS.
Sentence x	This API allows third-party apps to use the camera of Android devices.
Paraphrase y	The new API will allow developers to use Android's camera features to create custom apps.

Paraphrase Generation (Biomedical domain)	
Constraint c	The impact of obesity on health-related quality of life (HRQOL) in adolescents and young adults with spinal deformity is not well described.
Sentence x	The purpose of this study was to compare HRQOL measures in adolescent idiopathic scoliosis (AIS) patients with and without obesity.
Paraphrase y	This study aimed to investigate the relationship between HRQOL and obesity in adolescents with idiopathic scoliosis (AIS).

Summarization (General domain)	
Constraint c	There had been fears the flare could ignite the escaping gas at the Elgin platform, about 150 miles (240 km) east of the Scottish city of Aberdeen, potentially causing a huge explosion. Total said it had received the first indication that the flare might be out at lunchtime on Friday. The firm is "mobilizing all means to allow these options to be implemented," it said. The company, which is still investigating the cause of the leak, estimates that 200,000 cubic meters of gas a day are escaping.
Sentence x	"The gas cloud is fairly small in size and prevailing winds are blowing it away from the platform and dispersing it," Total said.
Summary y	The gas cloud is small and blowing away, Total said.

Summarization (Biomedical domain)	
Constraint c	A banana primarily consists of carbo hydrate chains (sugar), but also contains some minor amount of minerals and vitamins. Let's see what happens with this stuff - Sugar: Will be broken down to either be stored as fat (another form of carbo hydrate chains) or broken up and used to provide cell energy; the resulting "waste" hydrogen and carbon is disposed of in form of CO2 or H2O. Minerals: Are mainly used to regenerate organs/tissue and other organ functions; these could probably be still in your body, but even if they are, they are probably very rare. Vitamins: The atoms are very often disposed after use, so they too leave your body.
Sentence x	They do leave in rather short time frames, because the body can't store them well and needs it daily (that is why your diet should include them).
Summary y	They do leave in a short time, as the body does not store them long.

Table 7: Qualitative examples of pair generation. Along with each x and y, we present contextual constraint c used for pair generation.

of critics in our framework. To methodically verify this, we filter ChatGPT-Para using the same set of critics as in our framework, and train T5-large on the filtered dataset.

In this configuration, BERT-iBLEU on MRPC marks 64.1, improving over the original ChatGPT-Para but still falling behind Impossible-T5. We attribute this to the relatively small size of the filtered dataset ($n \approx 340k$), primarily due to a large portion of pairs not passing either the Dissimilarity or Diversity Filter. While ChatGPT can generate sensible paraphrases, it is not aligned with the specific evaluation criteria defined in the filtering stage, leading to the poor sample efficiency. Although the issue maybe mitigated via fine-tuning the teacher or over-sampling more generations, such solution would require substantially more compute than GPT2-scale LMs. In this sense, our framework provides an attractive alternative to LLM distillation, incorporating a small, cost-efficient data generator and a composition of filters, in replace of a gigantic data generator.

5 Related Works

Unsupervised Paraphrasing and Summarization Conventional approaches for unsupervised paraphrasing and summarization have focused on task-specific surrogates – *e.g.,* back-translation (Wi-

eting and Gimpel, 2018; Hu et al., 2019b) and autoencoding (Huang and Chang, 2021; Baziotis et al., 2019) – that guide the model toward desired output. These surrogate tasks inherently provide weak supervision signal compared to the complexity of the target task, often mandating carefully engineered perturbations (Huang et al., 2023; Niu et al., 2021), auxiliary constraints (Liu et al., 2022b; Chen et al., 2022b) or a complete re-training of teacher model (Meng et al., 2021). Apart from the task-specific methods, a growing line of research seeks to harness LLMs to paraphrase and summarize without supervision (Tang et al., 2023; Goyal et al., 2023). In fact, recent findings suggest that zero-shot generations prompted from LLMs exhibit human-level quality in various use-cases (Wahle et al., 2022).

Task-solving with Language Model More broadly, task-solving capabilities of LMs have been tested and analyzed across domains (Hendrycks et al., 2021). While large-scale pre-training allows models to acquire sufficient knowledge to solve complex tasks (Bommarito et al., 2023; Brahman et al., 2023; Jung et al., 2022), recent works suggest that their full capability is elicited from aligning the model knowledge with additional fine-tuning – e.g. using instruction data (Chung et al., 2022; Wang et al., 2022) and human feedback (Ouyang et al., 2022; Ziegler et al., 2020) – which often requires a curated set of annotated data. Our work suggests an alternative to this paradigm, by identifying and leveraging the paraphrastic knowledge intrinsic to the LM, rather than human annotation.

Data Generation with Language Model Another line of related works propose to directly distill models with LM-generated data, improving model reasoning (Zelikman et al., 2022; Hsieh et al., 2023), robustness (Chen et al., 2022a), controllability (Sclar et al., 2022), and language understanding (Ye et al., 2022). These works essentially follow the conceptual framework of Symbolic Knowledge Distillation (West et al., 2022), where a teacher model's knowledge is transferred to a student model via a symbolic, textual dataset. Other works explore to extract a standalone corpus from LMs, whether it be knowledge base (Alivanistos et al., 2022), dialogue (Kim et al., 2022), or evaluation suite (Perez et al., 2022). However, these works typically impose a strong assumption on the teacher LM (Wang et al., 2021), and require manually constructed set of prompts (Bhagavatula et al., 2022). Overcoming these limitations, IMPOSSIBLE DISTILLATION generalizes data generation into an off-the-shelf setup, removing the dependence to the teacher model's capability for the target task.

6 Conclusion

In this work, we propose IMPOSSIBLE DISTILLATION, a novel framework to distill high-quality paraphrase dataset and model from small, low-quality LMs. We show that by leveraging paraphrastic proximity and critic-guided distillation, IMPOSSIBLE DISTILLATION can empower small LMs to outperform competitive counterparts – in both performance and controllability, across domains and tasks, without training on human-authored paraphrases. Also, we find that DIMPLE, the natural byproduct of our method, presents higher fidelity and diversity than order of magnitude larger paraphrase datasets. IMPOSSIBLE DISTILLATION shows a promising direction to rediscover the under-explored capabilities of off-the-shelf language models, by accurately identifying their characteristics and amplifying them.

Acknowledgements

This work was funded in part by the DARPA MCS program through NIWC Pacific (N66001-19-2-4031) and IARPA HIATUS via 2022-22072200003.

References

Dimitrios Alivanistos, Selene Báez Santamaría, Michael Cochez, Jan-Christoph Kalo, Emile van Krieken, and Thiviyan Thanapalasingam. 2022. Prompting as probing: Using language models for knowledge base construction. *arXiv preprint arXiv:2208.11057*.

Zeyuan Allen-Zhu and Yuanzhi Li. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *CoRR*, abs/2012.09816.

Christos Baziotis, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos. 2019. Seq3: Differentiable sequence-to-sequence-to-sequence autoencoder for unsupervised abstractive sentence compression.

Chandra Bhagavatula, Jena D. Hwang, Doug Downey, Ronan Le Bras, Ximing Lu, Keisuke Sakaguchi, Swabha Swayamdipta, Peter West, and Yejin Choi. 2022. I2d2: Inductive knowledge distillation with neurologic and self-imitation.

BigScience. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Jillian Bommarito, Michael Bommarito, Daniel Martin Katz, and Jessica Katz. 2023. Gpt as knowledge worker: A zero-shot evaluation of (ai)cpa capabilities.

Faeze Brahman, Chandra Bhagavatula, Valentina Pyatkin, Jena D. Hwang, Xiang Lorraine Li, Hirona J. Arai, Soumya Sanyal, Keisuke Sakaguchi, Xiang Ren, and Yejin Choi. 2023. Plasma: Making small language models better procedural knowledge models for (counterfactual) planning.

Jan Cegin, Jakub Simko, and Peter Brusilovsky. 2023. Chatgpt to replace crowdsourcing of paraphrases for intent classification: Higher diversity and comparable model robustness.

Howard Chen, Jacqueline He, Karthik Narasimhan, and Danqi Chen. 2022a. Can rationalization improve robustness? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3792–3805, Seattle, United States. Association for Computational Linguistics.

Jifan Chen, Eunsol Choi, and Greg Durrett. 2021. Can NLI models verify QA systems' predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mingda Chen, Qingming Tang, Sam Wiseman, and Kevin Gimpel. 2019. Controllable paraphrase generation with a syntactic exemplar. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5972–5984, Florence, Italy. Association for Computational Linguistics.

Yi Chen, Haiyun Jiang, Lemao Liu, Rui Wang, Shuming Shi, and Ruifeng Xu. 2022b. MCPG: A flexible multi-level controllable framework for unsupervised paraphrase generation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5948–5958, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

William B. Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. ParaSCI: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434, Online. Association for Computational Linguistics.

Thibault Févry and Jason Phang. 2018. Unsupervised sentence compression using denoising auto-encoders. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 413–422, Brussels, Belgium. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2023. News summarization and evaluation in the era of gpt-3.

Junxian He, Daniel M. Spokoyny, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. Lagging inference networks and posterior collapse in variational autoencoders. *ArXiv*, abs/1901.05534.

Chaitra Hegde and Shrikumar Patil. 2020. Unsupervised paraphrase generation using pre-trained language models.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Steffen Herbold. 2023. Semantic similarity prediction is better than other semantic similarity measures.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes.

J. Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019a. Parabank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *AAAI Conference on Artificial Intelligence*.

J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019b. Large-scale, diverse, paraphrastic bitexts via sampling and clustering. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.

Kuan-Hao Huang and Kai-Wei Chang. 2021. Generating syntactically controlled paraphrases without using annotated parallel pairs. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1022–1033, Online. Association for Computational Linguistics.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023. Paraamr: A large-scale syntactically diverse paraphrase dataset by amr back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. *arXiv preprint arXiv:2205.11822*.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization.

Klaus Krippendorff. 2007. Computing krippendorff's alpha-reliability. annenberg school for communication departmental paper 43.

Ashutosh Kumar, Kabir Ahuja, Raghuram Vadapalli, and Partha Talukdar. 2020. Syntax-guided controlled generation of paraphrases. *Transactions of the Association for Computational Linguistics*, 8:329–345.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022a. WANLI: Worker and AI collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinxin Liu, Jiaxin Shi, Ji Qi, Lei Hou, Juanzi Li, and Qi Tian. 2022b. ParaMac: A general unsupervised paraphrase generation framework leveraging semantic constraints and diversifying mechanisms. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6193–6206, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xianggen Liu, Lili Mou, Fandong Meng, Hao Zhou, Jie Zhou, and Sen Song. 2020. Unsupervised paraphrasing by simulated annealing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 302–312, Online. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Yuxian Meng, Xiang Ao, Qing He, Xiaofei Sun, Qinghong Han, Fei Wu, Chun Fan, and Jiwei Li. 2021. ConRPG: Paraphrase generation using contexts as regularizer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2551–2562, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ning Miao, Hao Zhou, Lili Mou, Rui Yan, and Lei Li. 2019. Cgmh: Constrained sentence generation by metropolis-hastings sampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6834–6842.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Tong Niu, Semih Yavuz, Yingbo Zhou, Nitish Shirish Keskar, Huan Wang, and Caiming Xiong. 2021. Unsupervised paraphrasing with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*,

pages 5136–5150, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Ethan Perez, Sam Ringer, Kamilė Lukošiūtė, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, Andy Jones, Anna Chen, Ben Mann, Brian Israel, Bryan Seethor, Cameron McKinnon, Christopher Olah, Da Yan, Daniela Amodei, Dario Amodei, Dawn Drain, Dustin Li, Eli Tran-Johnson, Guro Khundadze, Jackson Kernion, James Landis, Jamie Kerr, Jared Mueller, Jeeyoon Hyun, Joshua Landau, Kamal Ndousse, Landon Goldberg, Liane Lovitt, Martin Lucas, Michael Sellitto, Miranda Zhang, Neerav Kingsland, Nelson Elhage, Nicholas Joseph, Noemí Mercado, Nova DasSarma, Oliver Rausch, Robin Larson, Sam McCandlish, Scott Johnston, Shauna Kravec, Sheer El Showk, Tamera Lanham, Timothy Telleen-Lawton, Tom Brown, Tom Henighan, Tristan Hume, Yuntao Bai, Zac Hatfield-Dodds, Jack Clark, Samuel R. Bowman, Amanda Askell, Roger Grosse, Danny Hernandez, Deep Ganguli, Evan Hubinger, Nicholas Schiefer, and Jared Kaplan. 2022. Discovering language model behaviors with model-written evaluations.

Minh Pham, Minsu Cho, Ameya Joshi, and Chinmay Hegde. 2022. Revisiting self-distillation.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. 2021. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*.

Melanie Sclar, Peter West, Sachin Kumar, Yulia Tsvetkov, and Yejin Choi. 2022. Referee: Reference-free sentence summarization with sharper controllability through symbolic knowledge distillation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9649–9668, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Lakshay Sharma, Laura Graesser, Nikita Nangia, and Utku Evci. 2019. Natural language understanding with the quora question pairs dataset.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea. Association for Computational Linguistics.

Tianyi Tang, Hongyuan Lu, Yuchen Eleanor Jiang, Haoyang Huang, Dongdong Zhang, Wayne Xin Zhao, and Furu Wei. 2023. Not all metrics are guilty: Improving nlg evaluation with llm paraphrasing.

Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160.

Joan Torruella and Ramon Capsada. 2013. Lexical statistics and tipological structures: A measure of lexical richness. *Procedia - Social and Behavioral Sciences*, 95:447–454.

Vladimir Vorobev, Maxim, and Kuznetsov. 2023. Chatgpt paraphrases dataset.

Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. How large language models are transforming machine-paraphrase plagiarism. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 952–963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Gary Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Maitreya Patel, Kuntal Kumar Pal, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza,

Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Shailaja Keyur Sampat, Savan Doshi, Siddhartha Mishra, Sujan Reddy, Sumanta Patro, Tanay Dixit, Xudong Shen, Chitta Baral, Yejin Choi, Noah A. Smith, Hannaneh Hajishirzi, and Daniel Khashabi. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Yichong Xu, Ruochen Xu, Dan Iter, Yang Liu, Shuohang Wang, Chenguang Zhu, and Michael Zeng. 2023. Inheritsumm: A general, versatile and compact summarizer by distilling from gpt.

Jiacheng Ye, Jiahui Gao, Qintong Li, Hang Xu, Jiangtao Feng, Zhiyong Wu, Tao Yu, and Lingpeng Kong. 2022. ZeroGen: Efficient zero-shot learning via dataset generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11653–11669, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Jiawei Zhou and Alexander Rush. 2019. Simple unsupervised summarization by contextual matching. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5101–5106, Florence, Italy. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 1097–1100, New York, NY, USA. Association for Computing Machinery.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences.

Algorithm 1 Diversity Filter

Require: A set of pairs $\mathcal{P}_{\text{in}} = \{(x_1, y_1), \cdots, (x_{|P|}, y_{|P|})\}$ generated using the same prefix c
Ensure: Filtered set of pairs $\mathcal{P}_{\text{out}}$
  $E \leftarrow \emptyset$
  for $i, j \in [1, |P|], i \neq j$ **do** ▷ search for duplicate pairs
    if $P_{NLI}(x_i \Rightarrow x_j) > \tau_{entail}$ **then**
      $E \leftarrow E \cup \{(x_i, y_i), (x_j, y_j)\}$
    else if $P_{NLI}(y_i \Rightarrow y_j) > \tau_{entail}$ **then**
      $E \leftarrow E \cup \{(x_i, y_i), (x_j, y_j)\}$
    end if
  end for
  $G \leftarrow (\mathcal{P}_{\text{in}}, E)$ ▷ define a graph where nodes are pairs and edges connect duplicate pairs
  $S \leftarrow$ Connected-Components(G)
  $\mathcal{P}_{\text{out}} \leftarrow \emptyset$
  for $\mathcal{C} \in S$ **do** ▷ find the max-entailing pair in each connected component
    $p_{\text{out}} = \text{argmax}_{(x,y) \in C} P_{NLI}(x \Leftrightarrow y)$
    $\mathcal{P}_{\text{out}} \leftarrow \mathcal{P}_{\text{out}} \cup \{p_{\text{out}}\}$
  end for

A Implementation Details

A.1 Pair Generation

As noted in Section 3.5, we start off using GPT2-XL and BioGPT-large as teacher LM – all with 1.5B parameters – generating paraphrases in general and biomedical domain respectively. To sample contextual constraints from these LMs, we use Nucleus Sampling with *top_p* $= 0.9$ and *temp* $= 1.0$. Additionally, we find that for general domain, generating new-style sentences by prompting GPT2 with a simple prefix (*e.g.,* `New York (CNN) -`) leads to less noisy and more diverse context. For BioGPT, we free-form generate without any prefix given. Throughout the distillation process, we used 4 Quadro RTX 8000 GPUs.

A.2 Critic Models for Paraphrase Generation

For semantic equivalence filter, we use Roberta-large-WANLI (Liu et al., 2022a), readily available at HuggingFace transformers (Wolf et al., 2020). To leave only the highly semantically equivalent pairs of paraphrases, we use $\tau_{\text{semantic}} = 0.75$, discarding all pairs with the bidirectional entailment score below this threshold. For dissimilarity filter, we use $\tau_{rouge} = 0.75$ and $\tau_{TED} = 12$. We use Stanford CoreNLP library to parse the constituency tree.

Finally, we present the formal algorithm of the diversity filter in Algorithm 1. We first create an undirected graph G where pairs are nodes and edges exist between duplicate pairs, then find the set S of all connected components in G. By discarding all but one with the maximal entailment score $p_{NLI}(x \Rightarrow y) + p_{NLI}(y \Rightarrow x)$ in each component, we remove the duplicate pairs in the candidate pool. As the duplicate pair search with NLI model is parallelizable, the time complexity follows that of the connected component search, *i.e.* $O(|P| + |E|)$ when using DFS-based algorithm (Tarjan, 1972).

Dataset	Quora		MSCOCO	
Model	*iBLEU*	*BLEU*	*iBLEU*	*BLEU*
Lag VAE	8.73	15.52	7.69	11.63
CGMH	9.94	15.73	7.84	11.45
UPSA	12.03	18.21	9.26	14.16
BT	11.64	11.59	9.72	14.36
Corruption	12.32	17.97	10.32	15.60
ConRPG	12.68	18.31	11.17	16.98
MCPG	13.58	24.84	11.99	20.54
Impossible-T5	**16.40**	**27.22**	**13.15**	**22.75**

Table 8: Experimental results of Impossible-T5 and unsupervised paraphrase generation methods on Quora and MSCOCO. Impossible-T5 consistently outperforms all unsupervised baselines across both benchmarks, in both metrics.

B Comparison with Unsupervised Paraphrase Generation Methods

To better understand the effectiveness of IMPOSSIBLE DISTILLATION, we conduct additional experiments that compare Impossible-T5 against state-of-the-art unsupervised methods for paraphrase generation (*i.e.* trained without human-written reference). Following prior works, we use Quora (Sharma et al., 2019) and MSCOCO (Lin et al., 2015) datasets repurposed for paraphrase generation. For baselines, we compare against Lag VAE (He et al., 2019), CGMH (Miao et al., 2019), UPSA (Liu et al., 2020), BT (Wieting and Gimpel, 2018), Corruption (Hegde and Patil, 2020), ConRPG (Meng et al., 2021), and MCPG (Chen et al., 2022b). Following past works, we compute and report iBLEU and 4-gram BLEU of each system.

The results are as shown in Table 8. Impossible-T5 consistently outperforms all unsupervised baselines across both benchmarks, in both metrics.

C Generalization to Sentence Summarization

C.1 Critic Models for Summarization

In IMPOSSIBLE DISTILLATION, data generation can easily be adapted to sentence summarization, by redefining the filters for the target task (§3.6).

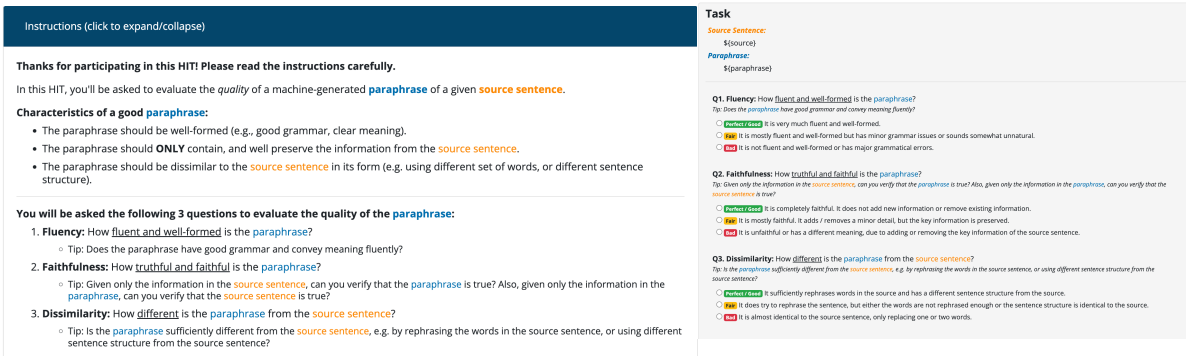

Figure 4: Screenshot of MTurk interface used for the human evaluation of model generated paraphrases.

Here, we explain the details of the critic models used for sentence summarization. First, a good summary should be entailed by the original statement without hallucination. Unlike paraphrases, however, summaries allow omitting less important details in the original statement. Therefore, we modify the semantic equivalence filter in paraphrase generation to consider only the unidirectional entailment between x and y:

$$f_{semantic}(x, y) = \mathbb{1}\Big\{p_{NLI}(x \Rightarrow y) \geq \tau_{semantic}\Big\}$$

We use the same threshold value as for paraphrase generation, $\tau_{semantic} = 0.75$. Next, a desirable summary should be a concise representation of the original statement. We therefore discard all pairs whose compression ratio (*i.e.* the sequence length ratio of y to x) is larger than a threshold τ_{comp_ratio}:

$$f_{comp\_ratio}(x, y) = \mathbb{1}\Big\{|y| < |x| \cdot \tau_{comp\_ratio}\Big\}$$

For our experiments, $\tau_{comp\_ratio} = 0.8$. Finally, we employ diversity filter as for paraphrase generation, removing all duplicate *(source, summary)* pairs from the generated dataset:

$$\mathcal{D}_T = \{(x, y) | (x, y) \in \mathcal{C}_T, \\ f_{semantic} \wedge f_{comp\_ratio} \wedge f_{diversity}(x, y) = 1\}$$

C.2 DIMSUM and Impossible-T5

Other than the re-defined filters, we use the same settings as paraphrase generation throughout the distillation pipeline. After self-distillation, we yield a high-quality dataset for sentence summarization (**D**ataset of **Im**possible **Sum**maries, or **DIMSUM**), with 1.5M sentence-summary pairs across news and biomedical domains. During this process, we also train T5-large into a specialized model for sentence summarization, which we consistently call **Impossible-T5** as for paraphrase generation.

D Human Evaluation Details

For human evaluation, we recruit annotators from Amazon Mechanical Turk (MTurk) with an IRB approval, and ensure that all paraphrases are annotated by 6 distinct evaluators with Hit Rate over 99%. To minimize subjectivity, we use 3-point Likert scale where annotators evaluate the fluency (whether the paraphrase exhibits fleunt language), faithfulness (whether the paraphrase well preserves the content of the original sentence and does not hallucinate), and dissimilarity (whether the paraphrase is sufficiently different from the original statment) of each output. We compensate workers with the hourly wage of $15. Figure 4 shows the actual MTurk interface used for paraphrase evaluation.

E Limitations and Future Work

In this work, we limit our experiments to sentential paraphrasing and summarization tasks. In future works, IMPOSSIBLE DISTILLATION could be applied to a broader range of tasks, *e.g.,* translation. To generate a parallel corpus for translation without human supervision, IMPOSSIBLE DISTILLATION could leverage the strong capability of multilingual LMs (Lample and Conneau, 2019; BigScience, 2023) and cross-lingual filters (Conneau et al., 2018).

IMPOSSIBLE DISTILLATION makes use of a fixed set of filters (*e.g.,* off-the-shelf NLI model) to determine which pair qualifies as a high-quality sample. Throughout the distillation pipeline, these filters remain frozen. Although our experiments show that the frozen filters are strong enough to distill a high-quality dataset than state-of-the-art paraphrase corpora, such filters may not always be accessible in wider range of tasks. Hence, future works could improve the framework by learning not

only the task model that generates candidate pairs, but also the filter model that scores the plausibility of a given pair. We envision that by co-evolving the task model and filter model throughout the distillation stages, our framework could generalize to more complex problems such as commonsense reasoning, where it is non-trivial to define which pairs qualify as good task example.

As with any distillation technique, IMPOSSIBLE DISTILLATION carries potential risk of amplifying undesirable properties of language models. While we focus on conditional generation tasks where the output is closely bound to the input, the trained model could inherit the bias and toxicity of its teacher in a more open-ended setting. Nonetheless, IMPOSSIBLE DISTILLATION distills knowledge into a symbolic, textual dataset – which can be interpreted and evaluated, allowing users to intervene in the distillation process and selectively filter which knowledge to be amplified. The inherent transparency of IMPOSSIBLE DISTILLATION, when incorporated with recent techniques for automatic bias detection and reduction, could empower safer knowledge transfer between language models.