Compositional Zero-Shot Domain Transfer with Text-to-Text Models

Fangyu Liu[1]*, **Qianchu Liu**[2], **Shruthi Bannur**[2], **Fernando Pérez-García**[2], **Naoto Usuyama**[3], **Sheng Zhang**[3], **Tristan Naumann**[3], **Aditya Nori**[2], **Hoifung Poon**[3], **Javier Alvarez-Valle**[2], **Ozan Oktay**[2], **Stephanie L. Hyland**[2]

[1] University of Cambridge, UK [2] Microsoft Health Futures, UK [3] Microsoft Health Futures, USA

fl399@cam.ac.uk, {t-floraliu,stephanie.hyland}@microsoft.com

Abstract

Label scarcity is a bottleneck for improving task performance in specialized domains. We propose a novel compositional transfer learning framework (DoT5[1]) for zero-shot domain transfer. Without access to in-domain labels, DoT5 jointly learns domain knowledge (from masked language modelling of unlabelled in-domain free text) and task knowledge (from task training on more readily available general-domain data) in a multi-task manner. To improve the transferability of task training, we design a strategy named NLGU: We simultaneously train natural language generation (NLG) for in-domain label-to-data generation, which enables data augmentation for self-finetuning and natural language understanding (NLU) for label prediction. We evaluate DoT5 on the biomedical domain and the resource-lean subdomain of radiology, focusing on natural language inference, text summarization, and embedding learning. DoT5 demonstrates the effectiveness of compositional transfer learning through multi-task learning. In particular, DoT5 outperforms the current state-of-the-art in zero-shot transfer by over 7 absolute points in accuracy on RadNLI. We validate DoT5 with ablations and a case study demonstrating its ability to solve challenging NLI examples requiring in-domain expertise.

1 Introduction

While pretrained language models demonstrate massive improvements on a wide range of natural language processing (NLP) tasks, it remains challenging to apply them to specialized domains (Ramponi and Plank, 2020). To acquire domain-specific task knowledge, a conventional approach is to perform domain-specific pretraining—usually masked language modelling (MLM) on in-domain raw text—followed by finetuning with in-domain task-annotated data (Lee et al., 2020; Gu et al., 2021; Boecking et al., 2022). However, this approach requires in-domain task labels that can be expensive to acquire. Another approach is to train a model with the usually abundant general-domain task labels and directly transfer to the new domain (Romanov and Shivade, 2018; Ma et al., 2021), but the transfer performance is often limited by the domain gap. Past studies on zero-shot domain transfer or unsupervised domain adaptation have explored methods to transfer task knowledge from a source domain to an unseen target domain (Ramponi and Plank, 2020; Ganin and Lempitsky, 2015), but they usually require external modules to perform feature or domain alignment and are not always easily applicable to pretrained language models. In particular, there is little understanding of how we can leverage and combine domain-specific knowledge and general-domain task knowledge in the context of the recent success of text-to-text architectures in transfer learning.

To close this gap, we propose DoT5, a novel compositional zero-shot domain-transfer framework based on the state-of-the-art (SOTA) transfer learning model transfer learning model Text-to-Text Transfer Transformer (T5) (Raffel et al., 2020). Throughout, the 'zero-shot' setup refers to zero-shot *domain* transfer with no access to labelled *in-domain* data.[2] By ''compositional'' we mean that DoT5 is able to combine seen task

*Work done at Microsoft Health Futures.

[1]DoT5 (read as ''dot five''): **D**omain Compositional Zer**O**-shot **T5**.

[2]The definition of 'zero-shot' in this paper follows recent studies (Pan et al., 2022; Zhao et al., 2022), and is similar to unsupervised domain adaptation, as discussed in §2. Another similar usage of 'zero-shot' is found in cross-lingual setups where no task labels are accessible in the target test language but labels in the same task are available in a source language. Note that this definition is different from 'zero-shot learning' traditionally used to refer to the prediction of unseen classes.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1097–1113, 2023. https://doi.org/10.1162/tacl_a_00585
Action Editor: Alexander Rush. Submission batch: 2/2023; Revision batch: 2/2023; Published 9/2023.

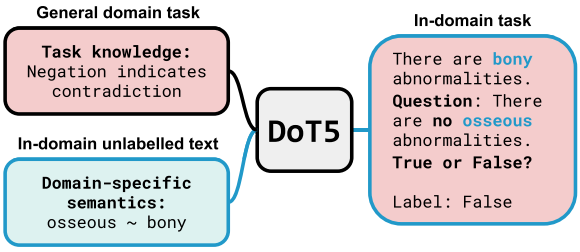

Figure 1: By combining task knowledge from general domain data and domain knowledge from in-domain unlabelled text, our text-to-text model DoT5 learns to solve in-domain tasks.

labels and domain text to acquire an unseen combination of task domain knowledge.

As shown in Figure 1, DoT5 combines domain knowledge and task knowledge by making the best use of in-domain free text and general-domain task labels, which are typically accessible and abundant. For example, in the context of natural language inference (NLI), DoT5 can learn domain-specific semantics (e.g., ‘‘bony abnormalities’’ is a synonym of ‘‘osseous abnormalities’’) from in-domain free text and transferable task knowledge from general-domain task labels (e.g., negation indicates contradiction) to infer domain-specific task knowledge (e.g., ‘‘There are no bony abnormalities’’ contradicts ‘‘There are osseous abnormalities’’).

We apply DoT5 to NLI, summarization, and text embedding learning, which are fundamental applications across many domains, and we explore zero-shot domain transfer to the high-value and highly specialized domain of biomedicine and its extremely low-resource subdomain of radiology. Due to their specialization, obtaining labelled data in these domains is expensive and time-consuming. For example, the radiology-specific NLI dataset (RadNLI) (Miura et al., 2021) contains only 960 manually labelled examples as development and test data and no training data is available.

The key to DoT5’s compositional transfer is *continual multi-task pretraining to simultaneously acquire domain and task knowledge*: We jointly train T5 with MLM on in-domain unlabelled data and general-domain tasks (NLI and summarization). To better acquire the transferable task knowledge from the general-domain task labels, we propose a multi-task setup we call NLGU. As depicted in Figure 2, NLGU gives each task two formulations: natural language generation (NLG) (label-to-data generation), and natural language understanding (NLU) (data-to-label prediction). NLU enables label prediction when tested in an unseen domain and forces model sensitivity to the conditioned label, assisting NLG. Meanwhile, NLG enables downstream tasks such as summarization or data augmentation. This enables DoT5 to generate its own NLI in-domain task data for further finetuning (a process we call self-finetuning), or to generate positive and negative examples for improving text embeddings by contrastive learning (Oord et al., 2018).

Our experiments show the effectiveness of DoT5 in zero-shot domain transfer, and our proposed multi-task compositional approach achieves large gains compared with sequential training with T5 across all tasks. In particular, we achieve SOTA zero-shot domain transfer performance on RadNLI (Romanov and Shivade, 2018), outperforming baselines including large language models (LLMs), sequential training approaches, and task-specific baselines by large margins. We also identify several key insights through extensive analysis: 1) All three key components (in-domain MLM, NLGU, self-finetuning) in DoT5 are important for transfer success while multi-task learning with in-domain MLM is the key for combining domain and task knowledge. 2) Scaling up model size significantly improves transfer performance. 3) DoT5 is able to solve challenging domain-specific task examples, indicating it acquires domain-specific task knowledge through compositional transfer.

To summarize, we present the following major contributions: 1) We propose DoT5, a general framework for compositional transfer learning with text-to-text models, and show that multi-task training is superior to sequential training in the models’ domain transfer. 2) With a novel NLGU training strategy combining generation and understanding, DoT5 can be used for both classification and generation tasks.[3] With the latter, DoT5 can perform self-finetuning to further improve transfer performance. 3) We show the effectiveness of DoT5 in zero-shot domain transfer, achieving SOTA zero-shot performance in radiology NLI. 4) Comprehensive analysis demonstrates the inner workings of DoT5’s compositional transfer.

[3]Notice that the tasks are limited to those that can have pairwise input instead of single sentence input.

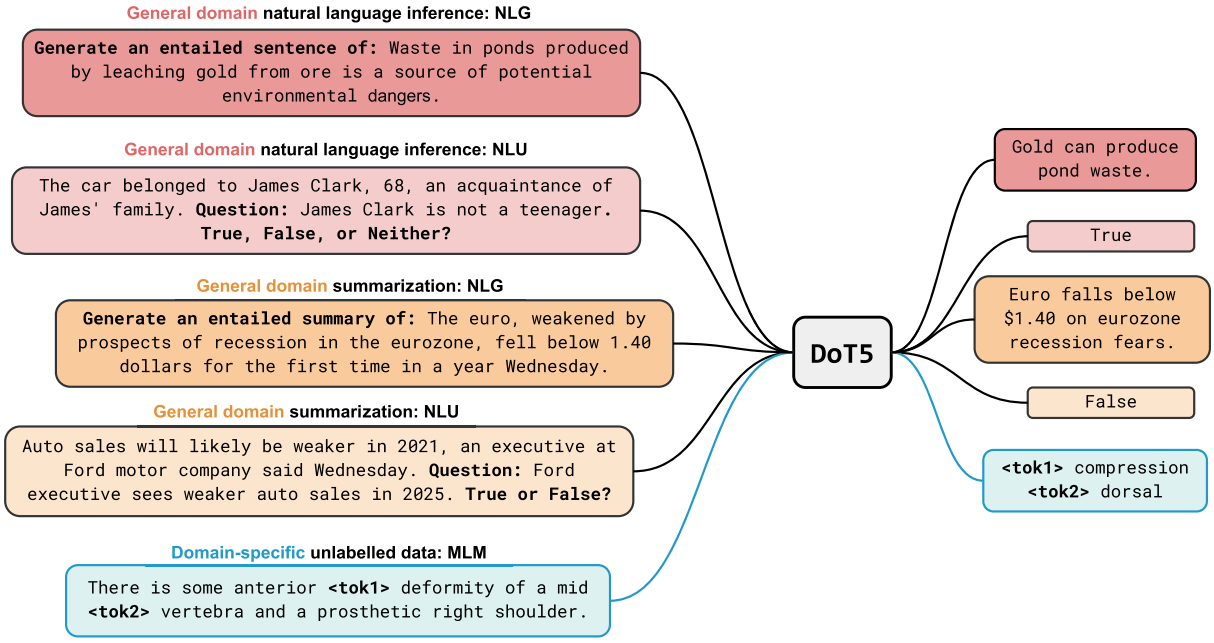

Figure 2: Continual pretraining of DoT5 on general-domain tasks (warm colors) and in-domain unlabelled text (blue). For task training, we form both NLG and NLU variants of NLI and summarization. All training is performed simultaneously, exploiting the unified text-to-text framework of T5.

2 Related Work

Cross-task Transfer with Text-to-text Models T5 (Raffel et al., 2020) unifies NLP tasks under a seq-to-seq framework and solves them using a single model. T0 (Sanh et al., 2022), FLAN (Wei et al., 2022), MetaICL (Min et al., 2022), and ExT5 (Aribandi et al., 2022) build on top of this idea and explore pretraining T5 with a massive collection of NLP datasets with diverse natural language prompts. Among them, T0, FLAN, and MetaICL investigate pretraining on a set of tasks, and then zero-shot transfer to another set of unseen tasks.

Domain-specific Pretraining Gururangan et al. (2020) show continual training on domain and task data can adapt pretrained models for new domains and tasks. Both BioBERT (Lee et al., 2020) and BlueBERT (Peng et al., 2019) apply the BERT pretraining protocol (i.e., masked language modelling and next sentence prediction) on PubMed Central (PMC) or PubMed articles. They continue pretraining BERT checkpoints instead of training from scratch. Gu et al. (2021) demonstrate the importance of domain-specific vocabulary and pretraining from scratch when in-domain text is abundant, and produces PubMedBERT by pretraining on PubMed articles. Similar to PubMedBERT, SciBERT (Beltagy et al., 2019) pretrains from scratch on a mix of both PMC and computer science publications. Boecking et al. (2022) introduce CXR-BERT, which is pretrained on biomedical and radiology corpora. SciFive (Phan et al., 2021) continually pretrains T5 checkpoints on PubMed abstracts with seq-to-seq MLM. We compare to finetuned versions of SciFive, PubMedBERT, and CXR-BERT in §4.3.

Zero-shot Domain Transfer Learning Training in one domain and directly testing on another domain has been a prevalent paradigm in zero-shot cross-domain transfer (Miura et al., 2021; Boecking et al., 2022; Agrawal et al., 2022). A similar zero-shot setup is also frequently seen in other transfer learning scenarios such as cross-lingual zero-shot learning (Conneau et al., 2018, 2020). Our summarization experiment is most similar to such a direct zero-shot setup. Concurrently, Pan et al. (2022) also propose to combine in-domain training and out-of-domain task knowledge. They proposed a zero-shot in-domain question answering model by finetuning a general-domain RoBERTa model with first domain-specific NER and then general-domain question answering. This study is the closest to our approach, with several key differences: Their method requires in-domain labels (in-domain

NER) whereas we do not require any in-domain task labels. They only test on question answering whereas we show a more diverse range of evaluation datasets. Additionally, they do sequential training whereas we perform multi-task training. Finally, their model is not generative and therefore it cannot perform NLGU and self-finetuning as we did in our approach (see §3).

Our proposed NLGU and self-finetuning strategies are closely related to cross-domain data augmentation. A line of work in information retrieval generates ''in-domain'' pseudo training data leveraging unlabelled in-domain texts. As an example, Ma et al. (2021) and Wang et al. (2022) train a passage-to-query generator for synthesizing in-domain queries for the task of zero-shot passage retrieval. Similarly, The NLG component in our proposed NLGU strategy can also perform data augmentation but with better granularity and diversity as we can generate label-conditioned task data to create both positive and negative examples.

Besides zero-shot transfer in NLP, unsupervised domain adaptation (which also assumes labels in current domain and unlabelled data in the target domain) is a long-standing research topic in machine learning in general (Huang et al., 2006; Pan et al., 2010b; Ganin and Lempitsky, 2015; Ramponi and Plank, 2020). Many conventional unsupervised domain adaptation methods require external components to align domains on the feature/embedding level. For example, Pan et al. (2010a) propose applying spectral feature alignment to align domain-specific words across domains into unified clusters. Ganin and Lempitsky (2015) add a domain classifier that promotes domain-invariant features via a gradient reversal layer. These methods are not always immediately suitable for the recent pretrained language models, especially the text-to-text models. In comparison, our approach exploits the task unifying nature of text-to-text models, which contain the inherent transfer learning abilities and requires minimal architecture changes.

3 Method

To achieve compositional transfer, DoT5 acquires domain knowledge and task knowledge via continual pretraining (see Figure 2). Specifically, we optimize a joint loss function composed of an in-domain masked language model loss (''domain-MLM'') and a general-domain task-specific loss:

$$\mathcal{L}_{\text{joint}} = \lambda \mathcal{L}_{\text{domain-MLM}} + (1-\lambda)\mathcal{L}_{\text{task}} \quad (1)$$

We set $\lambda = 0.5$ but explore tuning it in §4.1.

We use T5, an encoder-decoder generative language modelling framework (Raffel et al., 2020), to learn a conditional sequence generator $P(\text{output}|\text{input})$. T5 is chosen for two reasons: 1) It is a strong transfer learning model, and 2) it can unify classification and generation, which has potential to further boost transfer performance (see NLGU discussion in §3.2). We use the same pretraining objective (cross-entropy with teacher-forcing) as in T5.

We detail the two loss components for continual pretraining in §3.1 and §3.2. Once the model has been continually pretrained, it can be used to perform zero-shot domain transfer on a task. Task-specific designs for inference are given in §3.3.

3.1 Continual Pretraining with In-domain MLM

For $\mathcal{L}_{\text{domain-MLM}}$ we use the MLM loss (Devlin et al., 2019) to continually pretrain a T5 on in-domain free text: Given a piece of sampled radiology or biomedical text, we randomly mask 15% of its tokens and ask the model to denoise the masked input sequence, i.e., generate the masked tokens.

3.2 Continual Pretraining on General-domain Tasks

For $\mathcal{L}_{\text{task}}$, we define (x_1, x_2) as a text pair that denotes (*premise*, *hypothesis*) for NLI, and (*document*, *summary*) for summarization. The standard NLI task assigns labels from y: {entailment, neutral, contradiction}, and the task is $(x_1, x_2) \rightarrow y$. For summarization, the task is usually cast as $x_1 \rightarrow x_2$. We follow Sanh et al. (2022) to adopt a multi-task learning strategy to train summarization and NLI simultaneously. Hence, the basic setup of task learning would be: NLI as a discriminative NLU task plus summarization as an NLG task.

NLGU: Simultaneous NLG and NLU One immediate question is whether we can turn each task into both NLG and NLU (i.e., adding NLG for NLI and NLU for summarization). For NLI, we can add label-to-data NLG to generate pseudo

	Setting	Prompt (Input)	Output
NLI	NLG: $(x_1, y) \rightarrow x_2$	**Generate a** {label} **sentence of:** {premise}	{hypothesis}
	NLU: $(x_1, x_2) \rightarrow y$	{premise} **Question:** {hypothesis} **True, False or Neither?**	{True | False | Neither}
Sum.	NLG: $(x_1, y) \rightarrow x_2$	**Generate a** {label} **summary of:** {document}	{summary}
	NLU: $(x_1, x_2) \rightarrow y$	{document} **Question:** {summary} **True or False?**	{True | False}

Table 1: Prompts used for task-specific training with NLGU for both NLI and summarization (Sum). For NLI x_1: premise, x_2: hypothesis, and the label (y) is one of {entailed, neutral, contradictory}. For summarization x_1: document, x_2: summary, and the label (y) is one of {entailed, contradictory}.

in-domain text for data augmentation, performing $(x_1, y) \rightarrow x_2$ (the label y is used as control code). For summarization, we can also follow NLI to add a NLU task that predicts whether a document-summary pair is entailed (the correct match) or contradictory (a counterfactual summary) (§4.1). This NLU component aims to improve the factuality of generated text as it encourages the model to distinguish counterfactuals and true summaries. With the hypothesis that performing NLG and NLU simultaneously will mutually benefit each other, we propose NLGU, meaning joint training of NLG and NLU. With NLGU, we unify both summarization and NLI into $(x_1, x_2) \rightarrow y$ for NLU and $(x_1, y) \rightarrow x_2$ for NLG. The conditional generator then simultaneously optimizes two losses:

$$\mathcal{L}_{\text{task}} = \gamma \mathcal{L}_{(x_1,x_2)\rightarrow y} + \mathcal{L}_{(x_1,y)\rightarrow x_2} \quad (2)$$

We set $\gamma = 10$ to balance the two losses since x_2 is usually much longer than y (the classification label). NLU and NLG are both trained with sequence-to-sequence generation, and differ only in the input prompt and the expected output (Table 1). The prompt for $\mathcal{L}_{(x_1,x_2)\rightarrow y}$ is from Brown et al. (2020). The prompts for summarization are akin to those for NLI, with `premise` and `hypothesis` replaced with `document` and `summary`, respectively, and we only use {entailment, contradiction} relations.

3.3 Task-specific Designs for In-domain Zero-shot Inference

After continual pretraining, we zero-shot-transfer the trained model to three applications in specialized domains without requiring labels from these domains: 1) NLI, 2) summarization, and 3) text embedding learning.

NLI (with Self-finetuning) While the model is capable of directly performing NLI after training on general-domain NLI task labels with $(x_1, x_2) \rightarrow y$, we propose an additional step, self-finetuning, to boost transfer performance (§5.1). We first use the model's NLG capabilities to generate pseudo in-domain NLI data: We sample a set of sentences from the target domain as premises, and prompt the pretrained model to generate hypotheses (the NLG task) with each of the three control codes (labels). This pseudo-in-domain NLI dataset is then used as additional training data to finetune the same model to perform the NLU task: $(x_1, x_2) \rightarrow y$. The resulting finetuned model is then used for zero-shot NLI transfer.

Text Summarization We directly prompt the model after continual pretraining to summarize in-domain documents. We use the same prompt as pretraining: ''Generate an entailed summary of: {document}''. The output summary is then compared against the gold summary. Since this is already a task of text generation, i.e., $(x_1, y) \rightarrow x_2$, we cannot exploit self-finetuning as for NLI since we cannot improve generation from training on the model's own generated pseudo data.

Text Embedding Learning DoT5 can be directly used as a generator for data augmentation. Apart from creating more pseudo NLI task data to improve NLI, DoT5 can improve domain-specific embedding learning in general. To do so, we sample a set of in-domain sentences as anchors, and prompt the trained model to generate entailed and contradictory sentences to form positive and negative pairs for each anchor. With beam search size of 5, we sample the top-k most probable sequences as the entailed (positives) and contradictory (negatives) sentences of the anchor.[4] Given the collected anchors and positive/negative

[4]We experimented generating one, three, and five pairs of positives and negatives and found three to be the best in our setup. We thus use three across all models.

	Dataset	Task	Used for	# Examples
General	SNLI – Bowman et al.	NLI	Task pretrain.	550K
	MultiNLI – Williams et al.	NLI	Task pretrain.	392K
	AdversarialNLI – Nie et al.	NLI	Task pretrain.	162K
	Gigaword – Graff et al.	Summ.	Task pretrain.	1M
Radiology	MIMIC-CXR – Johnson et al.	MLM	Domain pretrain.	227K
	RadNLI – Miura et al.	NLI	Evaluation	480
	Open-I – Demner-Fushman et al.	Summ.	Evaluation	683
Biomedical	PubMed Abstracts	MLM	Domain pretrain.	4.2M
	MedNLI – Romanov and Shivade	NLI	Evaluation	1.4K
	PubMed 'ShortSum'	Summ.	Evaluation	5K
	MedSTS – Yanshan et al.	Similarity	Evaluation	371

Table 2: Datasets used in the study for task/domain pretraining ('pretrain.') and evaluation. The tasks are NLI, text summarization ('Summ.'), and document retrieval based on text embedding similarity.

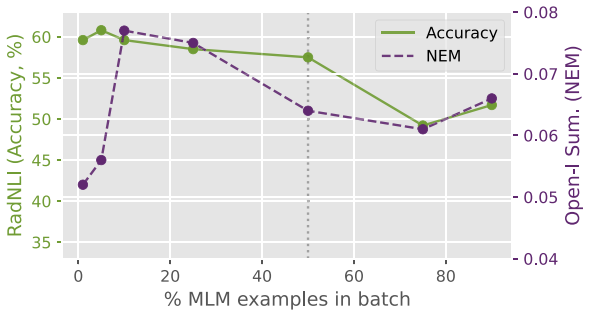

Figure 3: Varying the prevalence of in-domain MLM and task data in $\text{DoT5}_{\text{small}}$ training.

sentences, we finetune a SOTA sentence embedding model with a contrastive loss. Specifically, we continually finetune the all-mpnet-base-v2[5] model with a variant of InfoNCE (Oord et al., 2018) modified to handle multiple positives (Miech et al., 2020). The learned embedding space is then used for query-document retrieval or for computing text similarity.

4 Experiment

We introduce our experimental setup in §4.1, briefly discuss baseline approaches in §4.2, and present results in §4.3.

4.1 Experimental Setup

Details of the datasets used for training and evaluation are given in Table 2.

Pretraining Datasets As our continual pretraining is a multi-task process, we balance the in-domain and general-domain datasets in each batch via up/downsampling as needed: For radiology, we upsample MIMIC-CXR samples via duplication, whereas for biomedicine we downsample PubMed abstracts, in each case matching the general-domain task dataset size. We also balance the number of samples coming from each task, downsampling the summarization dataset to roughly match that of NLI ('# Examples' in Table 2). Experiments with $\text{DoT5}_{\text{small}}$ (Figure 3) indicate that downstream task performance could be boosted by tuning the relative prevalence of the data sources, with a task-dependent optimal value. In this proof-of-concept study, we fix a ratio of 1:1.

We generate counterfactual summaries of Gigaword based on Rajagopal et al. (2022). Specifically, we run a named entity recognition model on the documents from the Gigaword summarization training data, specifically the ''`en_core_web_sm`'' trained SpaCy pipeline (Honnibal et al., 2020). For each document that contains a named entity, we randomly sample an entity and replace it with a different named entity of the same category from the training corpus. This is our 'counterfactual' example.[6] We also filter out noisy data when the generated counterfactual contains UNK or #. The resulting dataset, as listed in Table 2, consists of 50% document-'wrong summary' pairs (i.e., 500k pairs), one for each true document-summary pair. To create pseudo NLI data for the self-finetuning process, we use all premises from the RadNLI/MedNLI development set and generate one entailed, one neutral, and one contradictory hypothesis for each premise. In total, we have 1440 and 4185 pseudo examples for RadNLI and MedNLI, respectively.

Evaluation Datasets and Metrics All the evaluation datasets are from domain-specific tasks (Table 2). For NLI, we report accuracy and

[5] `https://discuss.huggingface.co/t/train-the-best-sentence-embedding-model-ever-with-1b-training-pairs/7354`. `https://huggingface.co/sentence-transformers/all-mpnet-base-v2`.

[6] Note that a counterfactual is not always a contradiction. We approximate contradiction this way and use the 'contradictory' control code in our experiments for consistency.

macro-F_1 (out of 100, for legibility) on the test set of RadNLI and MedNLI. For summarization in radiology, we evaluate on findings-to-impression[7] summarization on the test split of the Open-I dataset (Demner-Fushman et al., 2016). For biomedical summarization, we create an abstract-to-title summarization dataset, 'PubMed ShortSum'. The data for this task is sampled from PubMed and filtered to abstracts shorter than 1000 characters. Compared with the traditional article-to-abstract PubMed summarization task, which evaluates long summary generation for long text (Cohan et al., 2018), PubMed ShortSum evaluates extreme summarization for short text and is a more comparable task to our general domain Gigaword summarization. For summarization evaluation, we use standard lexical metrics (BLEU-4, ROUGE-L) and domain-specific factuality metrics: named entity matching (NEM) for both radiology (Miura et al., 2021) and biomedical (Alambo et al., 2022) summarization, and CheXbert (Smit et al., 2020)[8] for radiology.

We evaluate embeddings trained for the biomedical domain on MedSTS (Yanshan et al., 2020), a clinical text similarity benchmark. Since the radiology domain has no text similarity datasets available, we design an impression-to-findings retrieval task on the Open-I test set, and report Accuracy@1/5/10. This retrieval task can also evaluate embedding quality as it requires the model to differentiate text from same/different reports by encoding texts from matching findings-impression pairs (from the same report) with similar representations.

Training Details Models are trained for 10 epochs with validation loss used for checkpoint selection. We use distributed data parallelism on eight GPUs with the largest batch size permissible given computational constraints, resulting in batch sizes of 1024, 512, and 128 for small, base, and large models. With a dataset of ∼8M samples, we thus train the large model for ∼64,000 steps per epoch. We use AdaFactor (Shazeer and Stern, 2018) with learning rates of 10^{-3} for MIMIC-CXR and 2×10^{-5} for PubMed pretraining.

[7]In a radiology report, the ''findings'' section is a detailed description and the ''impression'' section is a summary of the findings with follow-up recommendation.

[8]The average of the weighted-F_1 score across 14 pathological observations labelled by CheXbert.

Baselines	In-domain Text	General domain NLI/Summ.
BERT – Miura et al.	✓	∼ / –
CXR-BERT – Boecking et al.	✓	∼ / –
ESIM – Chen et al.	✗	✓ / –
T0 & T0++	✗	✗ / ✓
GPT-3	✗	✗ / ✗
GPT-3-{NLI, GW}	✗	✓ / ✓
CXR-BERT-NLI	✓	✓ / –
PubMedBERT-NLI	✓	✓ / –
$\text{SciFive}_{\text{large}}$-{NLI, GW}	✓	✓ / ✓
T5_{large}-MLM → Task	✓	✓ / ✓
DoT5	✓	✓ / ✓

Table 3: Baseline comparisons grouped into three categories: (1) task-specific zero-shot baselines (green), (2) large language models (grey), and (3) sequential training on in-domain text and general-domain task labels (pink). ''✓'' and ''✗'' specify whether the given data source was used for training. ''Summ.'' means summarization. Models only evaluated on NLI do not require summarization data, hence ''–''. ''∼'' indicates that BERT and CXR-BERT were finetuned on MedNLI, a 'near-domain' NLI dataset.

4.2 Baselines

We have three categories of baselines: (1) task-specific zero-shot baseline models reported from the literature (where applicable); (2) LLMs including T0 and GPT-3; and (3) sequential training first on in-domain unlabelled data and then on general-domain task labels. All the baseline models in our study must satisfy one constraint: not using any in-domain labels for the task, but they may differ in the required training resources (detailed comparison is found in Table 3). We compare with (2) as LLMs are known to be excellent zero-shot and few-shot learners for an unseen task, and should serve as a reasonable baseline for domain transfer. We provide (3) as a straightforward baseline to *sequentially* combine in-domain MLM training and general-domain task training as opposed to our proposed multi-task training.

Task-specific Zero-shot Baselines We compare with the strongest task-specific zero-shot models from the literature. For the NLI task, we compare with Miura et al. (2021) and Boecking et al. (2022), which both finetune a BERT model with MedNLI training data and then test on

RadNLI. Boecking et al. (2022) perform better as they use radiology-specific BERT model. Note that MedNLI is a *nearby-domain* corpus rather than general-domain task data, and in fact there has not been successful attempts in the literature to transfer general-domain NLI to RadNLI. Note that in the later sequential training section we will establish such baselines from finetuning CXR-BERT on general-domain NLI. For MedNLI, we compare with the best transfer learning results so far, ESIM (MultiNLI) which was trained on MultiNLI datasets (Romanov and Shivade, 2018). For radiology summarization, to our knowledge, we are the first to report results on direct transfer from general-domain summarization. For biomedical summarization, since we use a new dataset (PubMed ShortSum), there is no prior comparison.

Large Language Models T0 (Sanh et al., 2022) and GPT-3 (Brown et al., 2020) are massively pretrained language models that can be used off-the-shelf for zero-shot or few-shot inference. T0 is pretrained with multiple tasks including general-domain summarization datasets (but *not* NLI), and shows strong transfer ability (Sanh et al., 2022). T0 can be seen as a strong general-domain summarization model and also strong zero-shot domain transfer baseline on summarization. T0 is also particularly effective in transferring to unseen tasks. Therefore, we include T0 as a zero-shot baseline for NLI even though it has not been trained with any NLI data. We test T0 (3B) and the most powerful T0++ (11B) model. GPT-3 (Brown et al., 2020) (`davinci`) is a massive language model with 175B parameters, pretrained on raw text with an autoregressive language modelling objective.

In the general domain, both models are shown to have performed reasonably well on NLI and summarization with prompting. We test their zero-shot-inference capabilities in our experiments, following the original papers for prompt design. For the NLI task, both T0 models and GPT-3 use the ANLI prompt template described in Brown et al. (2020): ''<premise> Question: <hypothesis> True, False or Neither?''. For the summarization task, T0 used the prompt: ''<document> \n === \n Generate a title for this article:''. For GPT-3 summarization, we used the prompt (''<document>\n\n Tl;dr:'') as recommended in the OpenAI GPT-3 playground example.[9] Since GPT-3 benefits when few-shot examples are incorporated in the prompt, we create two additional baselines (GPT-3-NLI and GPT-3-GW[10]) that perform in-context learning of the task from general-domain NLI training data (30 examples, randomly selected) and Gigaword summarization training data (20 examples, randomly selected) respectively (Table 2).

Sequential Training The most straightforward way to exploit both in-domain unlabelled data and task labels is to first train on in-domain MLM and then further finetune on general-domain task labels.[11] We provide two variants of this baseline. The first type performs continual training with general-domain task labels from SOTA domain-specific pretrained models. We adopt SciFive (Phan et al., 2021), a T5 model pretrained on large biomedical corpora, CXR-BERT-General (Boecking et al., 2022), a radiology-specialized BERT model, and the PubMed-specific PubMedBERT (Gu et al., 2021). For finetuning these models we use the same general-domain task data as provided to DoT5, where for the BERT models we only do finetuning on NLI. This results in baseline models $\text{SciFive}_{\text{large}}$-NLI, $\text{SciFive}_{\text{large}}$-GW (summarization), CXR-BERT-NLI, and PubMedBERT-NLI. We further improve $\text{SciFive}_{\text{large}}$-NLI by including our proposed self-finetuning stage ($\text{SciFive}_{\text{large}}$-NLI + SFT). Since there is no radiology-pretrained T5 model, we compare with SciFive on both domains.

The second baseline type strictly compares multi-task training (DoT5) and sequential training. Here, we first pretrain T5 with in-domain MLM, and then continually pretrain on the general-domain task data, ensuring other factors remain the same including the training duration, use of NLGU, and use of self-finetuning where appropriate. We call this setting T5_{large}-MLM $\rightarrow$ Task.

[9] `https://beta.openai.com/examples/default-tldr-summary`.

[10] These are still zero-shot baselines as they do not use in-domain task examples.

[11] This baseline category is similar to contemporaneous work (Pan et al., 2022) where domain-task transfer is achieved through sequential in-domain *off-task* training followed by general-domain *in-task* training. Here we do not use in-domain task data of any kind.

Model	Accuracy	F_1-score
Radiology (RadNLI)		
BERT (Miura et al., 2021)	53.3	–
CXR-BERT (Boecking et al., 2022)	65.2	–
T0 (3B)	24.2	21.2
T0++ (11B)	35.4	33.3
GPT-3	22.1	18.9
GPT-3-NLI	26.7	25.6
CXR-BERT-NLI	75.0	73.5
$\text{SciFive}_{\text{large}}$-NLI	47.5	35.4
$\text{SciFive}_{\text{large}}$-NLI + SFT	70.2	66.3
T5_{large}-MLM $\rightarrow$ Task	78.3	75.6
$\text{DoT5}_{\text{large}}$	**82.1**	**79.8**
Biomedicine (MedNLI)		
ESIM (MultiNLI) (Chen et al., 2017)	51.7	–
T0 (3B)	37.0	23.9
T0++ (11B)	55.2	44.6
GPT-3	39.9	38.5
GPT-3-NLI	39.2	28.8
PubMedBERT-NLI	**75.7**	**75.8**
$\text{SciFive}_{\text{large}}$-NLI	50.1	41.4
$\text{SciFive}_{\text{large}}$-NLI + SFT	67.3	65.8
T5_{large}-MLM $\rightarrow$ Task	71.4	71.5
$\text{DoT5}_{\text{large}}$	71.2	69.9

Table 4: Zero-shot NLI results, showing micro accuracy and macro F_1. BERT and CXR-BERT are trained on MedNLI, we reproduce numbers from Miura et al. (2021) and Boecking et al. (2022) respectively. ESIM (Chen et al., 2017) is the highest-performing directly transferred model reported by Romanov and Shivade (2018). T0 (Sanh et al., 2022) and GPT-3 (Brown et al., 2020) baselines were conducted by us, matching stated hyperparameters where possible. Models with ''-NLI'' are finetuned or prompted baselines. 'SFT' means with self-finetuning.

4.3 Main Results

NLI (Table 4) $\text{DoT5}_{\text{large}}$ establishes new SOTA for zero-shot domain transfer on RadNLI and competitive results on MedNLI (Table 4). On RadNLI, $\text{DoT5}_{\text{large}}$ reaches an impressive 82.1% on accuracy and is the best performing model. It outperforms the strongest reported number from the literature (CXR-BERT) by more than 15%, and our baseline CXR-BERT-NLI by almost 7%. Comparing DoT5 to T5_{large}-MLM $\rightarrow$ Task on RadNLI reveals the benefit of *multitask* training for compositional transfer.

On MedNLI, $\text{DoT5}_{\text{large}}$ outperforms ESIM (MultiNLI) by almost 20% (accuracy), but does not quite reach the 75.7% accuracy achieved by PubMedBERT-NLI, which establishes a new SOTA in zero-shot domain transfer on MedNLI—supervised SOTA is 86.6% (Phan et al., 2021). Although factors such as tokenization and pretraining strategies may contribute, we speculate that the domain gap between MedNLI and our biomedical pretraining corpus explains the weaker performance of DoT5 on MedNLI. MedNLI was sourced from *clinical* notes in MIMIC-III, which differ distributionally from biomedical articles in PubMed. Supporting this hypothesis, we observed that DoT5 pretrained on radiology text, and the *sequential* baseline T5_{large}-MLM $\rightarrow$ Task achieved similar performance on MedNLI (70% accuracy), indicating that results on MedNLI may not fully reflect compositional domain knowledge transfer in our setup. In this case, a strong NLI-specific model is most performant, while lacking potentially advantageous versatile text generation/summarization capabilities.

Model	NEM	CheXbert	BLEU-4	ROUGE-L
Radiology (Open-I Summarization)				
T0 (3B)	.054	.243	.027	.088
T0++ (11B)	.019	.145	.012	.061
GPT-3	.050	.219	.006	.063
GPT-3-GW	**.093**	**.304**	.019	**.127**
$\text{SciFive}_{\text{large}}$-GW	.019	.124	.002	.036
T5_{large}-MLM $\rightarrow$ Task	.050	.256	.015	.077
$\text{DoT5}_{\text{large}}$	.082	.258	**.038**	.117
Biomedicine (PubMed ShortSum)				
T0 (3B)	**.293**	–	.053	.291
T0++ (11B)	.290	–	**.066**	**.341**
GPT-3	.197	–	.017	.184
GPT-3-GW	.272	–	.046	.266
$\text{SciFive}_{\text{large}}$-GW	.109	–	.010	.149
T5_{large}-MLM $\rightarrow$ Task	.230	–	.044	.232
$\text{DoT5}_{\text{large}}$	.263	–	.047	.260

Table 5: Zero-shot summarization results. NEM (named entity matching) and CheXbert (radiology-specific) assess domain-specific factuality, while BLEU and ROUGE are standard lexical metrics. In all cases higher is better. GW = *Gigaword*. T0 (Sanh et al., 2022) and GPT-3 (Brown et al., 2020) baselines were conducted by us.

Summarization (Table 5) $\text{DoT5}_{\text{large}}$ achieves competitive performance compared with the best model in radiology (GPT-3-GW) and biomedical domains (T0 models) (Table 5). In radiology, $\text{DoT5}_{\text{large}}$ is the second-best model. That the strongest performing models on summarization are LLMs with substantially many parameters is not surprising; we observe in §5.2 that DoT5 too

Radiology (Open-I Retrieval)			
Model	$\text{Acc}_{@1}$	$\text{Acc}_{@5}$	$\text{Acc}_{@10}$
all-mpnet-base-v2	8.3	15.1	20.2
+ $\text{DoT5}_{\text{large}}$ (no-MLM)	12.0	19.9	22.8
+ $\text{DoT5}_{\text{large}}$	**13.3**	**20.4**	**25.5**
Biomedicine (MedSTS)			
Model	r	ρ	
all-mpnet-base-v2	72.8	64.6	
+ $\text{DoT5}_{\text{large}}$ (no-MLM)	$76.4_{\pm 0.04}$	$67.1_{\pm 0.06}$	
+ $\text{DoT5}_{\text{large}}$	$\mathbf{76.9}_{\pm 0.00}$	$\mathbf{67.9}_{\pm 0.09}$	

Table 6: Text embedding learning results. Starting from a state-of-the-art embedding model (all-mpnet-base-v2), we finetune with DoT5-generated data (indicated by '+'). Radiology evaluation is retrieval: given the `impression` section of a report, find the corresponding `findings` section. For biomedicine, we report similarity on MedSTS (Yanshan et al., 2020), where r and ρ refer to Pearson's r and ρ (scaled by 100 for legibility).

enjoys scaling effects. Most importantly, we again demonstrate the benefit from multi-task compositional transfer as $\text{DoT5}_{\text{large}}$ significantly outperforms both T5_{large}-MLM→Task and SciFive-GW across all metrics in both domains. This further verifies that a naïve sequential training on these two sources does not lead to effective compositional knowledge transfer. We also acknowledge it is more difficult to perform domain transfer for generation tasks in general: We cannot perform the data augmentation NLG and self-finetuning pipeline as it amounts to training the model to generate its own outputs.

Text Embedding Learning (Table 6) The DoT5-generated examples greatly improve the SOTA sentence embedding model's capability on both impression-to-findings retrieval in radiology and semantic textual similarity (MedSTS) in the biomedicine domain (Table 6). This is evidence that DoT5-generated sentences are of high quality and have captured semantic similarity and contradiction required for learning good embedding model. We also compare with an ablated version of DoT5 without in-domain MLM to generate data and find that the full model performs better across the board. his shows the importance of domain training for generating good in-domain examples. We explore this further in §5.1.

Setting	RadNLI (acc.)	Sum. (NEM)
$\text{DoT5}_{\text{large}}$ (full model)	**82.1**	**.082**
(1) no in-domain MLM	63.5	.015
(2) no NLGU & (3)	59.0	.052
(3) no self-finetuning	49.6	–

Table 7: Ablation study on DoT5 components, evaluated on radiology. Removing MLM removes in-domain text during pretraining. Removing NLGU reduces NLI to purely discriminative (thus also disabling self-finetuning) and summarization to purely generative tasks. Self-finetuning is only used for NLI tasks. Sum. = summarization, NEM = named-entity matching metric. Note that the component is removed one by one but not incrementally.

5 Further Analysis

In this section, we demonstrate the importance of individual components of DoT5 (§5.1) and explore the role of model size (§5.2). Finally, we provide fine-grained analysis on RadNLI to verify whether DoT5 has indeed acquired domain-specific task knowledge from compositional transfer (§5.3).

5.1 Ablation Study

Through ablations, we probe the contributions of key components of DoT5: 1) In-domain MLM, 2) NLGU (combining NLU and NLG) (§3.1), and 3) self-finetuning for zero-shot NLI (§3.3). We conduct these ablations on the radiology domain on $\text{DoT5}_{\text{large}}$. The results are shown in Table 7.

We observe that all components are essential to the success of the model. In-domain MLM is especially important for summarization, without which the model fails in zero-shot transfer as it often just extracts a random subsequence of the document. Removing NLGU harms both NLI and summarization. Training without NLGU removes the NLG component from NLI and therefore disables self-finetuning. Self-finetuning is the most important component for boosting NLI performance, without which the model's accuracy drops more than 30%. As shown in Table 4, SciFive also benefits from self-finetuning in this way. This indicates that the pseudo in-domain NLI task data generated by NLGU is crucial. Training without

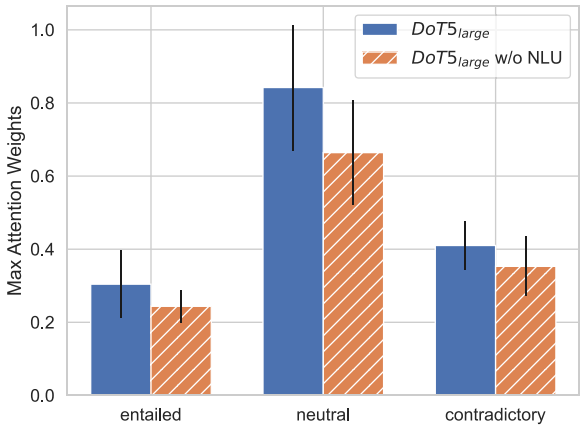

Figure 4: Maximum attention weights assigned to control code {label} (‘‘entailed’’, ‘‘neutral’’, ‘‘contradictory’’) in the prompt in NLI hypothesis generation, averaged over 100 randomly sampled examples from the RadNLI dev set. Error bars represent standard deviation.

Prompt: Generate a {label} **sentence of** {premise}:		
premise:	*‘‘Low lung volumes but no evidence of pneumonia.’’*	
label	Entailment:	*‘‘There is no evidence of pneumonia.’’*
	Neutral:	*‘‘The patient has a history of smoking.’’*
	Contradiction:	*‘‘The patient has pneumonia.’’*
premise:	*‘‘The patient is rotated slightly which limits assessment.’’*	
label	Entailment:	*‘‘Assessment is limited due to patient rotation.’’*
	Neutral:	*‘‘The patient is rotated to the left.’’*
	Contradiction:	*‘‘The patient is oriented perfectly. ’’*

Table 8: Pseudo-NLI data in the radiology domain generated by DoT5$_{\text{large}}$ for a given input premise and label. Premises are taken from the development split of the RadNLI dataset.

NLGU also removes the NLU task for summarization and brings down the performance, indicating that having an NLU task can also benefit generation.

We hypothesize that NLU improves NLG by forcing the model to be more sensitive to the control code in the prompt, leading to improved pseudo-data generation and better summarization. To test this, following Tang et al. (2018), we compute the maximum attention weights across all attention heads to the control codes in the prompt when generating an NLI hypothesis (Figure 4). We compare DoT5$_{\text{large}}$ trained with or without NLU. We see that the full model attends more on the control codes, suggesting that NLU is increasing label conditionality during generation. Table 8 shows some examples: When required to generate an entailment, the model can usually correctly paraphrase the original sentence; for negation,

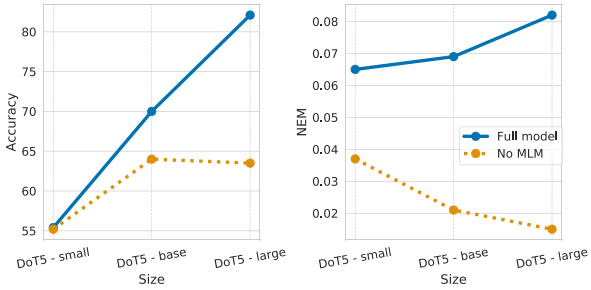

Figure 5: Scaling-up effect on RadNLI (left) and Open-I summarization (right). Both the full model and its ablated versions are compared. Note that self-finetuning is only applicable for the NLI tasks.

the model is usually able to correctly identify modifiers to flip the logic (e.g., change ‘‘increase’’ to ‘‘decrease’’ and adding or removing ‘‘no’’); for neutral, the model generates a thematically related sentence but not directly negating or agreeing with the original sentence.

5.2 Effect of Scaling Up

We have so far reported results on a large T5 model (770M parameters). In Figure 5, we plot the performance of small (70M) and base (220M) DoT5 models with their ablated versions for RadNLI and radiology summarization, showing a clear trend of increasing performance as the model size grows. Interestingly, this scaling effect disappears when we remove in-domain MLM, revealing the importance of domain training for larger models, especially for summarization. This is possibly because, without domain training, scaling up the model leads to overfitting to the general-domain task data. The compositional transfer framework from DoT5 however regularises the model for more complex knowledge acquisition, and thus is able to harness the power from larger models.

5.3 Evidence of Compositional Transfer in DoT5: A Case Study on RadNLI

Although RadNLI is a radiology-specific NLI dataset, we observe that some examples may be solvable using general-domain task knowledge (e.g., syntactic cues) alone. A general-purpose NLI model will likely detect that ‘There is no pneumothorax’ contradicts ‘There is a pneumothorax’ without requiring radiology-specific

Does not require radiology expertise	
Premise	*There is a small left pleural effusion.*
Hypothesis	*No pleural effusion or pneumothorax is seen.*
Label	Contradiction
Requires radiology expertise	
Premise	*The cardiac silhouette is top normal.*
Hypothesis	*The heart is not enlarged.*
Label	Entailment

Table 9: Examples from RadNLI that do or do not require radiology-specific knowledge to solve. While all models listed in Table 10 correctly solved the top example, only DoT5$_{\text{large}}$ solved the more challenging second example.

knowledge such as an understanding of pneumothorax. Therefore, higher performance on RadNLI may not strictly guarantee the model has acquired in-domain knowledge. To quantify how much of DoT5's transfer success is due to the acquisition of the previously unseen domain-specific task knowledge versus from direct application of the general-domain task knowledge, we manually annotated each of the 480 sentence pairs in the RadNLI test set by whether it could be solved without particular medical expertise.[12] Examples are shown in Table 9.

Table 10 compares three models on these subsets: DoT5$_{\text{large}}$ (a), T5$_{\text{large}}$-MLM $\rightarrow$ Task (b), and DoT5$_{\text{large}}$ *without* in-domain MLM (c) (equivalent to 'T5$_{\text{large}}$ $\rightarrow$ Task'). We further test with and without self-finetuning to probe its capacity to strengthen domain-specific competence.

While DoT5$_{\text{large}}$ achieves the best performance overall, it is specifically on challenging domain-specific cases that it outperforms T5$_{\text{large}}$-MLM $\rightarrow$ Task, an increase of 15 points in F_1. For example, in Table 9, only DoT5$_{\text{large}}$ is able to solve the second example which requires radiology-specific knowledge (the model should know cardiac silhouette includes heart size; and if the heart is top normal, then it should not be enlarged). This demonstrates the role of compositional transfer for inferring the otherwise unseen in-domain task knowledge (in this case, radiology NLI knowledge) solving challenging cases that require expertise.

[12]We determined 228 (47%) pairs could be solved without medical/radiological expertise, 177 (37%) could not, and the remaining 75 (16%) were ambiguous. Ambiguous cases were excluded from the analysis.

		Expertise required	
Model	**All cases**	**Yes**	**No**
a) DoT5$_{\text{large}}$	**80.7**	**70.1**	**86.4**
No self-finetuning	51.0	43.3	50.8
b) T5$_{\text{large}}$-MLM $\rightarrow$ Task	75.6	54.8	**86.5**
No self-finetuning	35.6	36.2	35.6
c) T5$_{\text{large}}$ $\rightarrow$ Task	59.5	35.3	70.1
No self-finetuning	37.5	36.1	35.1
Zero-rule baseline	24.6	29.0	20.0

Table 10: Macro F_1 of DoT5 with and without in-domain data during pretraining, on subsets of RadNLI requiring radiology-specific expertise or not. The zero-rule baseline always outputs the most common class (for RadNLI, this is 'Neither'). We report macro F_1 to account for differing label distributions. Note that T5$_{\text{large}}$ $\rightarrow$ Task is equivalent to DoT5$_{\text{large}}$ without in-domain MLM training.

The two ablated versions help understand where this domain-specific task knowledge is acquired. In-domain MLM training is key as removing it (c) significantly decreases the performance on domain-expert cases in particular, producing a model which cannot benefit from self-finetuning at all for such cases. This is because without in-domain MLM, the model is not able to generate good-quality pseudo in-domain labels in the first place, and therefore self-finetuning has little effect on the expert cases. Introducing in-domain data sequentially (b) resolves the performance gap on non-expert cases, but still underperforms on domain-specific cases relative to multi-task training (a). We conclude that the compositional fusion of task and domain knowledge happens during DoT5's multi-task pretraining phase with in-domain MLM as the key element, and that domain-specific competence is elicited through self-finetuning.

6 Conclusion and Discussion

We propose DoT5, a compositional transfer learning framework to solve domain-specific NLP tasks without requiring in-domain task labels. We show the effectiveness of DoT5 on zero-shot transfer to multiple tasks in the biomedicine and radiology domains. DoT5 significantly outperforms T5 sequential training across all tasks,

and achieves zero-shot SOTA in radiology NLI with massive gains. We also conduct extensive analyses to identify the contribution from each model component and the benefits from scaling up the model size, and demonstrate direct evidence of domain-specific task knowledge learned from DoT5's compositional transfer.

Limitations of this work include the challenge of drawing clear boundaries between domains and the necessarily incomplete exploration of hyperparameters and configurations. For example, general domain texts may contain biomedical or radiology sources, and our 'biomedical' NLI evaluation set leans strongly clinical, introducing a degree of domain shift. Investigation of the weighting of terms in the loss reveals the potential to improve performance through more exhaustive hyperparameter search—we emphasise that this was a proof-of-concept study and although DoT5 performs favourably, zero-shot domain transfer could be further pushed, especially if only a single downstream task is required.

The proposed NLGU method and subsequent self-finetuning was critical for improving downstream task performance. However, we observed an intermittent negative effect wherein the model would attempt to solve the NLU task when presented with an unusually long prompt. Further work can be done to refine this approach. For example, the benefit of NLGU in resource-rich domains is unclear. As our focus is on domain transfer and we do not evaluate on general-domain tasks, we leave such experimentation to future study.

Finally, we acknowledge that it is non-trivial to apply our full framework to single-sentence/paragraph classification tasks. While our most basic setup (compositional training of in-domain MLM and vanilla task training) can still be transferable to any task format, NLGU and self-finetuning would currently only work for tasks that involve pairs of texts. Nonetheless, we believe DoT5 proves to be a highly effective zero-shot domain transfer framework which will be beneficial to domain-specific applications beyond radiology and biomedicine.

Acknowledgments

The authors would like to thank the anonymous TACL reviewers and editors for their detailed feedback and helpful suggestions.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022.

Amanuel Alambo, Tanvi Banerjee, Krishnaprasad Thirunarayan, and Michael Raymer. 2022. Entity-driven fact-aware abstractive summarization of biomedical literature. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 613–620. IEEE.

Vamsi Aribandi, Yi Tay, Tal Schuster, Jinfeng Rao, Huaixiu Steven Zheng, Sanket Vaibhav Mehta, Honglei Zhuang, Vinh Q. Tran, Dara Bahri, Jianmo Ni, Jai Gupta, Kai Hui, Sebastian Ruder, and Donald Metzler. 2022. ExT5: Towards extreme multi-task scaling for transfer learning. In *International Conference on Learning Representations*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1371

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022*, pages 1–21, Cham. Springer Nature Switzerland.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1152

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-2097

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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. https://doi.org/10.18653/v1/D18-1269

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310. https://doi.org/10.1093/jamia/ocv080, PubMed: 26133894

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2003. English Gigaword. *Linguistic Data Consortium, Philadelphia*, 4(1):34.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23. https://doi.org/10.1145/3458754

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.740

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. 2006. Correcting sample selection bias by unlabeled data. *Advances in Neural Information Processing Systems*, 19.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317. https://doi.org/10.1038/s41597-019-0322-0, PubMed: 31831740

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682, PubMed: 31501885

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2021. Zero-shot neural passage retrieval via domain-targeted synthetic question generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1075–1088, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.92

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889. https://doi.org/10.1109/CVPR42600.2020.00990

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.201

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.416

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.441

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010a. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*, pages 751–760. https://doi.org/10.1109/TNN.2010.2091281

Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. 2010b. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210.

Xiang Pan, Alex Sheng, David Shimshoni, Aditya Singhal, Sara Rosenthal, and Avirup Sil. 2022. Task transfer and domain adaptation for zero-shot question answering. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 110–116, Hybrid. Association for Computational Linguistics.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT

and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5006

Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: A text-to-text transformer model for biomedical literature. *arXiv preprint arXiv: 2106.03598*.

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. *Journal of Machine Learning Research*, 21(140):1–67.

Dheeraj Rajagopal, Siamak Shakeri, Cicero Nogueira dos Santos, Eduard Hovy, and Chung-Ching Chang. 2022. Counterfactual data augmentation improves factuality of abstractive summarization. *arXiv preprint arXiv: 2205.12416*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M. Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2018. An analysis of attention mechanisms: The case of word sense disambiguation in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 26–35, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6304

Kexin Wang, Nandan Thakur, Nils Reimers, and Iryna Gurevych. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2345–2360, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.168

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge

corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/N18-1101

Wang Yanshan, Afzal Naveed, Fu Sunyang, Liwei Wang, Shen Feichen, Rastegar-Mojarad Majid, and Hongfang Liu. 2020. MedSTS: A resource for clinical semantic textual similarity. *Language Resources and Evaluation*, 54(1):57–72. https://doi.org/10.1007/s10579-018-9431-1

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Weiran Xu, Huixing Jiang, Wei Wu, and Yanan Wu. 2022. Domain-oriented prefix-tuning: Towards efficient and generalizable finetuning for zero-shot dialogue summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4848–4862, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.357