A Synthetic Data Approach for Domain Generalization of NLI Models

Mohammad Javad Hosseini* **Andrey Petrov** **Alex Fabrikant** **Annie Louis***
Google Deepmind
{javadh,apetrov,fabrikant,annielouis}@google.com

Abstract

Natural Language Inference (NLI) remains an important benchmark task for LLMs. NLI datasets are a springboard for transfer learning to other semantic tasks, and NLI models are standard tools for identifying the faithfulness of model-generated text. There are several large scale NLI datasets today, and models have improved greatly by hill-climbing on these collections. Yet their realistic performance on out-of-distribution/domain data is less well-understood. We explore the opportunity for synthetic high-quality datasets to adapt NLI models for zero-shot use in downstream applications across new and unseen text domains. We demonstrate a new approach for generating NLI data in diverse domains and lengths, so far not covered by existing training sets. The resulting examples have meaningful premises, the hypotheses are formed in creative ways rather than simple edits to a few premise tokens, and the labels have high accuracy. We show that models trained on this data (685K synthetic examples) have the best generalization to completely new downstream test settings. On the TRUE benchmark, a T5-small model trained with our data improves around 7% on average compared to training on the best alternative dataset. The improvements are more pronounced for smaller models, while still meaningful on a T5 XXL model. We also demonstrate gains on test sets when in-domain training data is augmented with our domain-general synthetic data.

1 Introduction

Over the past decade, NLI tasks have been critical for benchmarking the representation strengths of our language models. Today, the accuracy on the oft-reported Multi-NLI (Williams et al., 2018) (MNLI) dataset has reached above 92%[1] and supersedes human-level performance (Nangia and Bowman, 2019). Simultaneously, the practical applications of NLI models has gathered immense attention in fact-checking and source attribution of LLM outputs (Honovich et al., 2022; Rashkin et al., 2023). For such downstream tasks, model generalization to new domains and data-distributions is critical. We present a method of synthetic data generation to create a *general* dataset with varied but balanced distribution of premise lengths, domains of text, and NLI labels (Table 1), and demonstrate improved accuracy with the new data.

Domain = essay
P: This book does a great job of putting all the different approaches under one roof, so that you can see what other researchers are doing and how they do it. H: The book covers different research approaches in a single place so you can compare them. L: entailment
Domain = reddit title
P: TIL the difference between "literally" and "figuratively". It was so easy to learn, I literally did a backflip. H: I did not bother learning the difference between "literally" and "figuratively". L: contradiction
Domain = story for kids
P: Once upon a time, there was a very special young lady named Cinderella. Her stepmother and stepsisters were very mean to her. But she continued to be kind and helpful. H: Cinderella was very kind to everyone. L: neutral

Table 1: NLI examples in our synthetic data.

MNLI was a first effort to create a multi-domain NLI *training* dataset with examples from 5 genres spanning fiction, formal texts and conversations, and with single-sentence premise/hypothesis texts. Today, the use of models trained from such datasets has expanded well past routine benchmarking into a variety of practical tasks. In downstream problems involving web-scale LLMs, such as fact-checking of social media text, the domains and texts are clearly more diverse. Yet, as a field, we have not fully explored the distribution-general abilities of

*Equal contribution.
[1] https://gluebenchmark.com/leaderboard

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2212–2226

our models for downstream semantic tasks beyond NLI itself. There are many other NLI training sets: ANLI (Nie et al., 2019) for harder reasoning going beyond stylistic elements, and WANLI (Liu et al., 2022) that generates synthetic examples through worker and AI collaboration and which replicate complex reasoning patterns.

To complement these efforts, we explore the opportunity for *synthetic high-quality datasets to adapt NLI models for zero-shot use in downstream applications across new and unseen text domains*.

Our generation covers nearly 40 realistic and distinct domains, ranging from reviews, social media comments, to legal texts, also with varying lengths of the premise text. Our technique employs a chain of LLM tasks tuned to generate high quality, creative premise-hypothesis pairs together with a 3-way NLI label (*entailment*, *contradiction*, or *neutral* (Williams et al., 2018)). A first step generates domain names, the second produces premises of different lengths in these domains, and the final LLM call produces hypotheses and labels conditioned on each premise. We demonstrate how this approaches creates data with a balanced distribution of domains, labels, and premise lengths.

We fine-tune NLI models on this synthetic data corpus, and present their accuracy on the TRUE factual consistency benchmark (Honovich et al., 2022), consisting of 11 tasks unseen by our data and other training sets. We show that our *general data*-trained models obtain state-of-the-art NLI performance and single-handedly outperform models trained on MNLI, ANLI, or WANLI with around 7% improvement over the best alternative for T5 (Raffel et al., 2020) small models. The gap is lower but still around 2% for T5 XXL model size.

Our main contribution is thus that the *general* synthetic data approach improves the generalization power of NLI models, especially when small models and fast inference is key. We further show that, while in-distribution performance is hard to beat for tasks with in-distribution training data, our synthetic data can still improve in-distribution performance when used to augment the training data for models with sufficient capacity.

2 Related Work

We describe various ways of creating NLI data, and why models should generalize beyond the data.

Human annotation of NLI examples. The major datasets available today were created via costly, time consuming annotation tasks, and significant human effort. Standardly, for datasets such as SNLI (Bowman et al., 2015a) and MNLI (Williams et al., 2018), the annotation starts with a premise sentence taken verbatim from different sources. Annotators are asked to write a hypothesis sentence that is entailed, contradicted, or is neutral to the premise. This writing task is complex for humans, and it is well-known that the collected examples often have undesired stylistic artifacts, for example, the hypotheses alone being highly predictive of the label (Gururangan et al., 2018). Later efforts (ANLI dataset) have focused on improving the quality of examples by including model adversaries into the annotation rounds (Nie et al., 2019).

The diversity in genre in these datasets depends on the sources from which the premises are drawn. The SNLI dataset contains image captions (Bowman et al., 2015b). The MNLI dataset (Williams et al., 2018) has 10 domains from the Open American National Corpus (OANC)[2], only 5 of which are used for training. The ANLI dataset (Nie et al., 2019) contains text from Wikipedia, news, fiction, formal spoken text, and causal or procedural text. In this way, their domain coverage is limited by the chosen corpora and licensing constraints. For example, none of these datasets contain reviews, forum discussions, and science texts, domains which are prevalent and important in applications.

Understandably, different sources and methods of data collection produces training examples of a certain style and distribution. The generalization of these NLI sets to fully new settings is an interesting problem (Adila and Kang, 2022), yet less explored. Our work aims to shed some insights here.

Domain generalization. This problem of training/test mismatch receives less importance during development since models are trained and evaluated on splits of the same dataset. But real world applications of these models cannot assume that the test data matches the training distribution. Models need to be adapted to individual test domains using domain adaptation, or alternatively one could train domain-general models which work well on multiple unseen domains or distributions. We focus on this problem of domain generalization.

There is a large body on work on how training and optimization of models can be adapted for better generalization (Muandet et al., 2013; Wang et al., 2021). Another well-known approach, espe-

[2] https://anc.org/

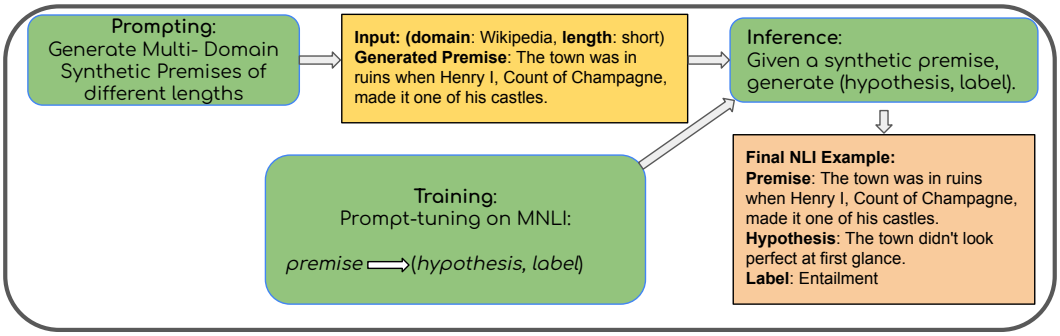

Figure 1: Generating the *General*-NLI examples.

cially in computer vision, is to augment the training data to increase its diversity and reduce model overfitting (Tobin et al., 2017; Rahman et al., 2019; Zhou et al., 2020). Liu et al. (2020) uses data augmentation to overcome multiple adverserial attacks. In this work, we explore the usefulness of synthetic data generation for the domain-generalization of NLI models.

Synthetic NLI data generation. Today's LLMs have opened up the possibility of synthetic data to aid many NLP tasks (Puri et al., 2020; He et al., 2022; Agrawal et al., 2023; Li et al., 2023). For NLI, synthetic data has been used for different goals: augmenting small training sets and adding *in-domain* examples for self-training (Vu et al., 2021; He et al., 2022), and increasing the size of harder examples (Liu et al., 2022). Most of these methods prompt LLMs with sequences of premise-hypothesis sentence pairs. The pairs are then labelled by a teacher model. Liu et al. (2022) specialize the generation towards complex linguistic and reasoning patterns in existing datasets. We employ synthetic data to improve the diversity and balance of training data along the dimensions of domain, premise length, and label skew.

3 Synthesizing a *General*-NLI Dataset

We aim to generate NLI examples in different domains, and with premises of varied lengths.[3]

Generating synthetic examples for NLI is in fact a challenging problem. The goal is to produce a pair of texts, which exemplify the reasoning behind different NLI labels. But for the data to be useful, these texts must have creative content and language, and require reasoning. Prior synthetic data generation approaches (He et al., 2022; Liu et al., 2022) generate the premise and hypothesis sentences as a single sequence, followed by annotating the label using a teacher model or human raters. Instead, to achieve maximum control over multiple dimensions: genre, length, and different NLI labels, we generate this dataset in two steps: (i) enumerate diverse domains and generate premises in those domains (Section 3.1), (ii) generate hypotheses and labels given the premises (Section 3.2). Figure 1 provides an overall view into the LLM tasks we use for generating our data.

3.1 Generating Premises with Varied Lengths in Multiple Domains

Going beyond distinctions based on the source of a text, it is hard to define the boundaries of a domain in a strict manner. Properties of a text differ along many dimensions: its genre (e.g., news, poetry, or fiction), topic (e.g. politics, science), and the platform or venue for the content, either spoken or written (e.g., reddit, email, image captions, or telephone conversations). We adopt a practical perspective, and consider all these distinctions as the latent features leading to differences between text collections. It is in fact well-known that stylistic variations in NLI datasets impact generalization (Belinkov et al., 2019; Adila and Kang, 2022).

So we do not start with predefined domains. Rather, we first build a text-generation model which generates triples of domain name, text length, and text in the domain. The resulting texts from different domains are collected into our premise set. We build this model using few-shot prompting of FLAN-PaLM2 L (Unicorn) model (Google and et al., 2023)[4] using texts from a few seed domains. We draw 18 in-prompt examples of varied lengths

[3]Longer hypotheses are not of interest typically. A hypothesis is entailed if hypothesis is true given the premise. Long hypotheses are less likely to contain precise entailment and contradiction relations, with some exceptions such as summaries.

[4]Available from https://developers.generativeai.google.

from 8 domains (*news headlines*, *news*, *shopping reviews*, *wikipedia*, *movie reviews*, *place reviews*, *twitter* and *reddit post*). The example texts are taken from public websites with a few edits if needed.[5] We provide the full list of seed examples in Appendix A. We also select these texts to be of different lengths.

Figure 2 shows our prompt.[6] The length category is set to either *short* for single sentences and *paragraph* for longer texts. We sample from this model, with a temperature of 1 to get creative *new domains* and texts. Ideally, these samples would directly be useful as premises (with domain and length labels). However, these samples were skewed towards certain domains, e.g., certain types of forums, which neither correspond to real word distributions of web-text, nor serve the purpose of a general model. So we first identify new domains generated by the model. We examined the *new* domains generated in about 1000 samples. Some generated domains were closely related to, or paraphrases of, each other; e.g. both *travel forums* and *US travel forums* were generated. Others were rare or noisy. So we manually selected 38 diverse domains including the seed domain names (Table 2).

We then generate balanced samples in these domains, and for the length labels (short, paragraph). We use the same prompt as in Figure 2, but substitute a new domain and length category of interest to it, to generate a text with those properties. This simple text generation model produced high quality and creative texts in different domains. We use these texts as the *premises* in our data.

3.2 Generating Hypotheses and Labels

We now discuss how we attach hypotheses and labels to our premises to generate complete NLI examples, i.e., (premise, hypothesis, label) triples.

We train LLMs to leverage existing NLI datasets, and learn the task of writing hypotheses for given premises.

Our model conditions on a premise to generate a (hypothesis, label) pair. We generate the label automatically (and accurately, details in next section), and do not need an additional human/teacher labelling step. This model is trained via prompt-tuning of FLAN-PaLM 540B (Chowdhery et al., 2023; Chung et al., 2022) on the training split of the MNLI dataset. Figure 3 shows our prompt which has definitions for the three NLI labels similar to the MNLI annotation guidelines (Williams et al., 2018).[7]

ads, blog post, book reviews, casual dialog, chat message, email, essay, fans forum, forum post, google play reviews, government documents, legal, legal document, medical, movie plot, movie reviews, news, news comments, news headlines, phone conversation, place reviews, quora, recipe, reddit comment, reddit title, research paper abstract, scientific article, shopping reviews, song lyrics, sports news, story for kids, student forum, student papers, support forum, travel guides, twitter, wikipedia, youtube comments

Table 2: Our final list of domains for data generation.

We used prompt-tuning (Lester et al., 2021) for training, in lieu of fine-tuning, for two reasons: a) With prompt-tuning, only a few embeddings are updated (100 in our experiments) leading to efficient training, and b) prompt-tuning provides regularization and avoids memorization of the training set details.[8] We note that in our preliminary experiments, we also tried just prompting LLMs (no training) to generate NLI examples; however, we did not obtain high quality examples.

Using the prompt-tuned model, we perform inference once on each of the premises obtained from Section 3.1.[9] We note that large models and regularization were important for creative generation of hypotheses. A T5 XL model (3B parameters) fine-tuned on the same task lead to examples with poor creativity and low utility for training. In many cases, the synthetic hypothesis was a subset of the premise (entailment) or had some simple modifications (e.g., negation) to introduce contradiction.

[5] This includes news websites (BBC) for *news* and *news headlines*, e-commerce and review websites (eBay and thetechoutlook.com) for *shopping reviews*, Wikipedia, Google Play for *movie reviews*, citymaps.uk and top-rated.online websites for *place reviews*, X (Twitter) and Reddit.

[6] We note that we also tried prompting without any instruction (just with few-shot examples) and the generated text had similar quality.

[7] We also tried another similar instruction for defining the task and labels (different wordings). However, we did not observe meaningful differences with the final prompt shown in the paper. This is probably because the LLM learns the task well after being prompt-tuned on a large set of examples (MNLI).

[8] See details of our prompt-tuning running time and hyperparameters in Appendix B.

[9] We discard examples if a) the generated text for hypothesis-premise pair is mis-formatted, or b) the generated labels are not among *entailment*, *neutral*, and *contradiction*. Such errors account for less than 1% of the generated data.

Generate a text of a given size in the domain.

domain: {place reviews}
length: {short}
text: {I waited an hour. The doctor was terribly stressed. She didn't answer questions.}

domain: {reddit post}
length: {paragraph}
text: {Hey there everyone! I often see people asking where to start when getting into prog metal, so I thought instead of answering every one of them individually I'd make a list. I'm not going into too much depth because otherwise this will become endless, but I'll try to give a brief explanation of all styles I'm going over. So let's get started!}

domain: {

Figure 2: The prompt used to generate new domains. For generating new text, we use the same prompt, add a domain and length category of interest (either *short* or *paragraph*), and add "text: {" at the end. We take the output up to the first "}" as the generated domain or text.

Given a sentence called premise, generate a related sentence called hypothesis. Then generate a label explaining the relationship between the sentences. The options for the label are 'entailment', 'contradiction', and 'neutral'. Entailment means that if the premise is true, the hypothesis is also true. Contradiction means that if the premise is true, the hypothesis is false. Neutral means that if the premise is true, we cannot say whether the hypothesis is true or false.

premise: {premise} hypothesis: {

Figure 3: The instruction used for training and inference of the (hypothesis, label) generator model.

3.3 The Final *General* Dataset

We generated (premise, hypothesis, label) triples in the 38 domains (from Section 3.1), and for two length categories (*short* and *paragraph*). The final dataset contains 684,929 examples. We hold-out 500 examples for creating a human annotated test set, and split the rest into training, development, and test splits.

Table 3 shows the data size and label distributions in each split. The number of examples from each label is relatively balanced (35.4% entailment, 31.1% contradiction, and 33.5% neutral). The balanced distribution of the GNLI dataset is predictable given how we generated the data. We use the MNLI data to prompt-tune our generator and this training data has a balanced class distribution. Therefore, after our prompt-tuning, the synthetic GNLI data also has a relatively balanced class distribution. This is in contrast to the previous WANLI dataset that used in-context learning to generate synthetic examples (Liu et al., 2022). Although they had balanced in-context examples (1/3 for each label), the final dataset had only 15% contradiction. This shows that prompt-tuning is effective in mirroring the training label distribution. We also note that if the original dataset was not balanced, we could perform sampling to obtain a balanced dataset. Alternatively, we could use a weighted loss function.

Our generation is also balanced with respect to the premise length and domain by design (we sample examples in a stratified manner). The average number of words per *short* premise is 21, and 60 for *paragraph* length. The average number of words in hypotheses is mostly uniform, 10 and 12 for *short* and *paragraph* length premises.

SPLIT	SIZE	# LABELS (E/C/N)
All	684,929	242,154 / 212,950 / 229,325
Train	670,739	237,325 / 208,676 / 224,738
Dev	6,845	2,453 / 2,146 / 2,246
Test	6,845	2,376 / 2,128 / 2,341
Human annotated test	490	181 / 155 / 154

Table 3: Different splits of our *general*-data. *Human annotated test* are 490 (out of 500) examples where at least 2 out of 3 annotators have agreed on the label.

Table 4 shows a few examples from our data. The premises come from different domains and are diverse in form and topic. Hypotheses are relevant to the premise and are creative in contrast to slightly modifying the premise and/or taking a subset of it. These attributes are unlike our observations with smaller and less powerful language models such as T5 XL (Section 3.2). In our experiments, we empirically demonstrate the impact of this data for training. We note that the idea here is to generate diverse data in different domains. It is possible that some of these examples are not factual, but the truth of the hypothesis is checked against the premise, not any background knowledge.

We also performed a human annotation experiment to a) to understand the accuracy of labels on our generated examples, and b) create a curated

DOMAIN AND LENGTH	PREMISE	HYPOTHESIS	LABEL
travel guides / short	**This charming boutique offers 43 rooms and suites in the heart of historic St John's**, and is the perfect base for exploring Antigua's rich history	The boutique is located right in the middle of the historic area.	entailment
support forum / short	**I'll be posting a video with the solution** once my phone finishes resetting.	I've already solved the problem.	neutral
legal document / paragraph	This Agreement will bind and inure to the benefit of both parties hereto and their respective personal representatives, heirs, successors, and permitted assigns. **Any attempt by any party hereto to assign, sell or otherwise transfer all or part of his or her rights or obligations under this Agreement, other than as provided herein, will be null and void**, notwithstanding the existence of any provision of law to the contrary.	This agreement allows for any party to reassign their rights and obligations.	contradiction
phone conversation / short	A. What's better for us for dinner tonight, Italian or Indian? B. Well, Italian is cheaper, but **Indian is quicker to order**.	Ordering Indian food takes a long time but it is better.	contradiction
essay /short	**The first three days of the trip were fantastic.** I had a blast with my friends.	The first three days of the trip were fantastic; the rest was horrible.	neutral
place reviews / short	**The food was fine** but there was only one couple serving that night and it was very busy.	The food tasted like it had been in the microwave for too long.	contradiction

Table 4: Synthetic examples from our data. We show examples with different domains, length categories, and labels. The most relevant part of the premise is bolded manually for ease of reading.

high-accuracy multi-domain test set for evaluation.

Each of the 500 generated examples was annotated with an NLI label by three of the paper's authors. Note that the examples were taken verbatim from the models, and are not revised by the annotators. The average Cohen's κ score between annotators is 67.97%, indicating substantial agreement. A majority label (2 out of 3 annotators) was obtained in 490 examples, and we use these as a human-annotated test set. We identified a subset, called *unanimous*, where all annotators agreed on the labels (344/500).

Annotators disagreed with examples that were ambiguous, e.g., the premise did not provide enough context. These cases are usually borderline and challenging. For the following example, one author annotated it as entailment and two authors annotated it as neutral (the person might still have not solved the problem).

premise: "I'll be posting a video with the solution once my phone finishes resetting."

hypothesis: "I've already solved the problem"

We also measured the accuracy of synthetic labels from our model against the *majority* and *unanimous* labels from human annotators. Model labels have high accuracy with 80.41% against majority and 89.53% on the *unanimous* examples. The κ coefficient between model labels and majority and unanimous subsets is also high (70.53% and 84.17% respectively).

4 Experiments

We explore the strengths of our *general*-dataset (GNLI for brevity), by examining the model predictions on data unseen during training.

We compare models trained on our data with those trained on other large NLI training sets. We choose three such sources: the MNLI dataset (392K training examples), ANLI (162K) with examples that are harder for MNLI trained models, and WANLI (102K), a dataset created by machine-human collaboration. We note that all of these datasets are collected with a similar methodology, i.e., given a premise (and optionally a label), annotators (or LLMs) write a hypothesis. The final label is then manually assigned to the example (if not given as input). In addition, WANLI and GNLI have used MNLI exemplars (few-shot or supervised learning) for data generation. So these datasets would have similar properties in theory.

For all these sources, we trained T5 models (Raffel et al., 2020) as a standard test bed, and explore models of different sizes: small (60M), large (770M), and XXL (11B). We trained on the respective training splits, and tune hyper-parameters on the corresponding validation sets. For WANLI which does not contain a validation split, we used the MNLI validation data. We also trained models on the combination of GNLI and other datasets. For these combined models, we tuned

hyper-parameters based on the classification accuracy on the development set of GNLI.[10]

We train two classifiers for each model size (e.g., T5 XXL) and training data (e.g., ANLI): a 3-way classifier with all the three labels, and a binary classifier. For the binary case, we convert each NLI dataset into a binary dataset with entailment and non-entailment (neutral and contradiction) labels. We use the binary classifiers and 3-way classifiers for factual consistency evaluation and NLI benchmarks, respectively.

4.1 Performance on Unseen Factual Consistency Benchmarks

We first test different models on data unseen by all of them. We use the TRUE benchmark, a collection of 11 evaluation datasets that contain human annotations for factual consistency in diverse tasks. The tasks include: A) **abstractive summarization**: FRANK (Pagnoni et al., 2021), SummEval (Fabbri et al., 2021), MNBM (Maynez et al., 2020), (Wang et al., 2020), QAGS-CNNDM (Wang et al., 2020), and QAGS-XSum (Wang et al., 2020). B) **dialogue generation**: BEGIN (Dziri et al., 2022), Q^2 (Honovich et al., 2021), and DialFact (Gupta et al., 2022). C) **fact verification**: FEVER (Thorne et al., 2018), and VitaminC (Schuster et al., 2021). D) **paraphrase detection**: PAWS (Zhang et al., 2019). The benchmark standardizes the above datasets by converting all annotations to binary labels corresponding to whether the entire text is factuality consistent w.r.t the grounding text or not. This task is a downstream application of NLI models and importantly, the data in this benchmark was not created using the same protocol as NLI benchmarks.

We train different sizes of T5 models on MNLI, ANLI, WANLI, and GNLI. In addition, we report results on models trained on the mixture of MNLI, ANLI, and WANLI (M + A + W). Table 5 shows the results. Following previous work, we report the Receiver Operating Characteristic Area Under the Curve (ROC AUC) for binary detection of inconsistent examples. GNLI outperforms MNLI on all datasets across model sizes showing that it has much stronger generalization. On average, GNLI outperforms MNLI for T5-small trained models by a 8.58% margin, T5-large trained models by 6.88%, and T5-XXL trained models by 2.96%.

To the best of our knowledge, the previously reported best NLI model on the TRUE benchmark was T5 XXL trained on ANLI (Honovich et al., 2022; Gekhman et al., 2023). GNLI obtains 6.85% improvement on average over the best alternative with a single dataset for T5 small (WANLI), 3.26% for T5 large (ANLI), and 2.03% for T5 XXL (WANLI). Therefore, GNLI obtains a new state-of-the-art result on TRUE, outperforming other models with large margins on average and on almost all of the individual test sets within TRUE. In addition, GNLI alone outperforms the mixture of MNLI, ANLI, and WANLI by a relatively large margin for T5 small and T5 large, and by a small margin for T5 XXL.

We note that Gekhman et al. (2023) have recently proposed a synthetic dataset, called TrueTeacher, for the task of factual consistency detection. They used summarization models to condense CNN/DM articles, and labelled the document-summary pair with FLAN-PaLM 540B (Chowdhery et al., 2023) based NLI model. Models trained on the TrueTeacher outperformed ANLI ones. We trained a T5 XXL on the TrueTeacher data and observed better performance compared to GNLI as well (88.06% vs 85.75% on average). This is expected since TrueTeacher is collected directly for the task of factual consistency detection and makes only the binary distinction.

4.2 Cross-Dataset Performance on NLI Benchmarks

We now examine the models on test sets available with large NLI collections (or validation sets in the absence of test data with labels (MNLI and ANLI)). In this case, at least one model has been trained on data from the same test distribution. Here we seek to understand how general our current datasets are and we also include GNLI in this analysis.

Table 6 shows the results. While all these datasets have been created with the aim of being domain-general, we see that generally the training data distribution makes a huge difference. The best test numbers are usually obtained by training on the corresponding training sets. For example, the best MNLI numbers are obtained with a model that is trained on MNLI. Note that the GNLI dataset (while including a component that is trained on MNLI) does not include the MNLI examples. These results indicate that the style and properties of different NLI test sets are still rather specific to the individual NLI dataset, and a large dataset with the same type of examples performs best on the corresponding test set. However, the GNLI dataset

[10] See hyper-parameter details in Appendix C.

	FRANK	QAGS C	QAGS X	MNBM	Summ Eval	BEGIN	Dial Fact	Q^2	PAWS	FEVER	Vitamin C	Avg
T5 SMALL												
MNLI	49.62	37.76	58.30	70.30	45.97	80.77	76.10	68.40	51.68	89.33	70.10	63.48
ANLI	50.95	54.64	44.70	53.99	51.34	57.66	55.39	45.02	47.09	55.24	53.50	51.77
WANLI	57.99	54.14	70.21	69.90	48.98	65.79	77.62	68.97	51.51	84.35	67.85	65.21
M + A + W	50.20	47.35	61.76	69.69	46.75	79.08	76.01	66.00	**59.15**	89.94	72.71	65.33
GNLI	**67.32**	**60.22**	**72.39**	**76.91**	**56.29**	**82.21**	**81.23**	**72.10**	57.33	**90.54**	**76.11**	**72.06**
T5 LARGE												
MNLI	79.15	58.13	79.56	79.27	61.59	82.13	87.65	77.32	75.82	93.97	81.60	77.84
ANLI	81.78	74.69	81.81	75.49	71.60	78.21	85.63	78.43	84.72	94.03	**89.63**	81.46
WANLI	80.31	74.46	70.11	67.70	72.86	80.37	**89.15**	**82.16**	83.17	93.82	82.79	79.72
M + A + W	83.57	72.28	82.27	78.28	72.61	81.13	87.25	79.81	**85.86**	94.63	86.48	82.20
GNLI	**90.14**	**81.33**	**84.02**	**79.49**	**79.75**	**83.45**	88.76	79.77	84.63	**94.73**	85.86	**84.72**
T5 XXL												
MNLI	88.18	79.03	83.07	78.31	72.35	81.76	88.32	76.84	83.11	95.13	84.58	82.79
ANLI	87.90	82.08	84.68	76.41	75.79	79.77	81.06	74.68	86.35	93.46	**90.59**	82.98
WANLI	88.59	72.18	82.85	73.29	74.61	82.47	**92.40**	**84.88**	87.29	94.97	87.38	83.72
M + A + W	90.60	**87.23**	86.73	79.49	**79.44**	83.56	85.56	76.45	**89.95**	95.07	88.55	85.69
GNLI	**91.38**	85.48	**87.03**	**79.97**	79.28	**84.00**	88.57	77.40	87.10	**95.34**	87.69	**85.75**

Table 5: Evaluation of multiple trained models on the TRUE benchmark. The rows M + A + W show the results of training models on the mixture of MNLI, ANLI, and WANLI. Results are split into three blocks based on the LM size (T5 small, T5 large, and T5 XXL). We report average AUC-ROC results on all the datasets (expressed as percentages). The best result for each model size and dataset is bolded.

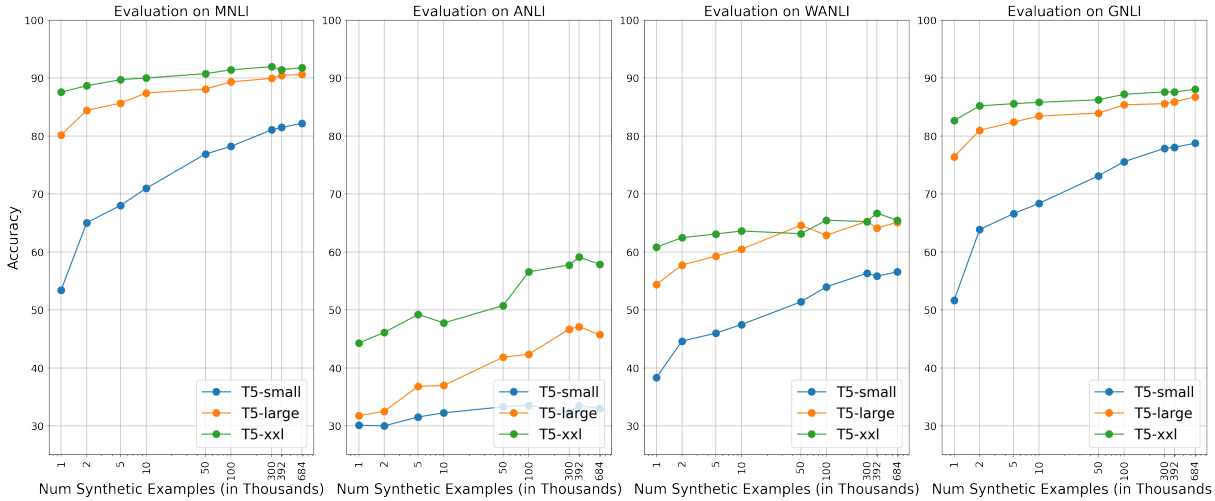

Figure 4: Accuracy of different T5 models when trained on different number of training examples from GNLI. Each plot has the results on one evaluation set.

is as accurate as MNLI on these test sets, even without the explicit addition of MNLI examples.

We also trained models on all datasets combined with GNLI. We note that in most cases (except for T5 small), the combined datasets have at least some modest improvements over the original datasets (underline numbers in the table). We speculate that T5 small's model capacity is not high enough to capture all the information in the combined datasets, but once the model capacity increases, we generally see improvements by adding GNLI to the other NLI datasets.

4.3 How Much Data is Needed for Successful Training?

GNLI is generated synthetically which is a more efficient and cheaper process compared to crowd-annotated data. It is possible to generate as many examples as necessary, and it is unknown in advance how many examples are needed to get a good performance. On the other hand, generating large sets of examples uses more computing resources. In this section, we study the effect of training data size on evaluation accuracy. We sample N thousand synthetic training examples from GNLI, where $N \in \{1, 2, 5, 10, 50, 100, 300, 392, 671\}$ (671K is the full GNLI dataset). We then train

Train / Eval	MNLI	ANLI	WANLI	GNLI Human
T5 SMALL				
MNLI	**83.37**	31.34	56.52	75.31
ANLI	70.35	**48.31**	52.70	67.14
WANLI	60.40	36.41	**72.60**	57.76
GNLI	82.18	33.00	56.56	**77.14**
MNLI + GNLI	82.66	30.94	55.82	77.76
ANLI + GNLI	72.89	37.94	48.65	69.80
WANLI + GNLI	78.02	34.87	86.69	76.53
T5 LARGE				
MNLI	**90.83**	40.22	63.58	82.24
ANLI	86.87	**63.62**	63.70	81.22
WANLI	81.10	48.03	**92.08**	77.14
GNLI	90.61	45.72	65.10	**83.67**
MNLI + GNLI	91.04	44.94	65.54	82.45
ANLI + GNLI	90.61	63.69	65.03	83.67
WANLI + GNLI	87.80	48.34	96.84	83.88
T5 XXL				
MNLI	**92.11**	55.44	65.92	**83.27**
ANLI	90.01	**73.37**	67.04	82.86
WANLI	84.61	60.00	**86.46**	81.84
GNLI	91.77	57.87	65.43	82.65
MNLI + GNLI	92.13	55.94	67.06	84.08
ANLI + GNLI	91.96	72.94	66.26	84.08
WANLI + GNLI	90.34	60.19	87.90	84.69

Table 6: Performance on NLI benchmarks (accuracy percentage). The models were trained on the respective datasets and tested on all their own and other datasets' test (or validation) sets. Results are split into three blocks based on the LM size (T5 small, large, T5 XXL). We also report the results of combining GNLI with MNLI, ANLI, and WANLI. We bold the highest accuracy per model size and evaluation dataset, for models trained on the single datasets. For each combined training set (GNLI + X) and model size, if the result is better than the original dataset (X), the number is underlined.

T5 models on all these sample sizes. We then evaluate the trained models on different NLI datasets. The evaluation is on validation sets from MNLI and ANLI, and WANLI and GNLI (synthetic) test sets.

Figure 4 shows the results. We observe that in most cases (model sizes and evaluation sets), at least around 300K examples is needed to get a decent performance. We also explicitly tested GNLI with 392K which is the same size as MNLI. In all cases, GNLI 392K has a very similar accuracy to the full dataset. We also observed similar trends for the TRUE benchmark.

5 Conclusion

A decade of increasingly useful NLI benchmarks and datasets have been instrumental in improving LLMs for various tasks. We have presented a new exploration of how the data distribution of each data source still impacts downstream performance on new examples. We proposed a synthetic data approach to mitigate these effects with examples balanced for domain, length and labels. We show that, by drawing on an LLM's parametric knowledge of a broad range of domains, such synthetic data enables us to both train significantly more domain-general NLI models, and to improve intrinsic NLI model performance on in-domain test data by augmenting in-domain training data.

6 Limitations

We do not release the synthetic *general* NLI data with our paper. However, our method for generating them can be replicated with access to an LLM, either to directly reproduce our results or to apply our approach to other domains, text lengths, and/or training set sizes of interest. Our process for generation requires multiple LLM tasks which uses more compute than a single stage one. But we note that the data is of high linguistic quality with this method. At the same time, the generated premises could potentially contain fictional information, and should not be used for training models that learn facts from data. We have applied our approach to generalize only one dataset (MNLI), which has examples in English. While we obtain positive results, the results on other datasets, and for other languages remain an empirical question. In addition, we performed experiments with FLAN-PaLM 540B (for prompt-tuning) and FLAN-PaLM2 L (for prompting). However, our method is straightforward and can be easily replicated by other LLMs. We expect comparable LLMs should lead to similar results.

References

Dyah Adila and Dongyeop Kang. 2022. Understanding out-of-distribution: A perspective of data dynamics. In *I (Still) Can't Believe It's Not Better! Workshop at NeurIPS 2021*, pages 1–8. PMLR.

Priyanka Agrawal, Chris Alberti, Fantine Huot, Joshua Maynez, Ji Ma, Sebastian Ruder, Kuzman Ganchev, Dipanjan Das, and Mirella Lapata. 2023. Qameleon: Multilingual qa with only 5 examples. *Transactions of the Association for Computational Linguistics*, 11:1754.

Yonatan Belinkov, Adam Poliak, Stuart M. Shieber, Benjamin Van Durme, and Alexander M. Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence,*

Italy, July 28- August 2, 2019, Volume 1: Long Papers, pages 877–891.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015a. 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.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015b. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 632–642. Association for Computational Linguistics (ACL).

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv e-prints*, pages arXiv–2210.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2022. Evaluating attribution in dialogue systems: The begin benchmark. *Transactions of the Association for Computational Linguistics*, 10:1066–1083.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models. *arXiv preprint arXiv:2305.11171*.

Rohan Anil Google and, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Prakhar Gupta, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Dialfact: A benchmark for fact-checking in dialogue. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3785–3801.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. 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 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, Annotate, and Learn: NLP with Synthetic Text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461.

Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847.

Tianyu Liu, Zheng Xin, Xiaoan Ding, Baobao Chang, and Zhifang Sui. 2020. An empirical study on model-agnostic debiasing strategies for robust natural language inference. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 596–608.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919.

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. 2013. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 10–18. PMLR.

Nikita Nangia and Samuel Bowman. 2019. Human vs. muppet: A conservative estimate of human performance on the glue benchmark. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4566–4575.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2019. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Raul Puri, Ryan Spring, Mohammad Shoeybi, Mostofa Patwary, and Bryan Catanzaro. 2020. Training question answering models from synthetic data. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5811–5826, Online. Association for Computational Linguistics.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. 2019. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4):777–840.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30.

Tu Vu, Minh-Thang Luong, Quoc Le, Grady Simon, and Mohit Iyyer. 2021. STraTA: Self-training with task augmentation for better few-shot learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5715–5731, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Tao Qin. 2021. Generalizing to unseen domains: A survey on domain generalization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4627–4635. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

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.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. 2020. Learning to generate novel domains for domain generalization. In *Computer Vision – ECCV 2020*, pages 561–578, Cham. Springer International Publishing.

A Seed Examples for Prompting

We provide the seed examples for prompting FLAN-PaLM2 L (Unicorn) to generate premises in Table 7.

B Running time and Hyper-parameter Details of Prompt Tuning Experiments

For prompt-tuning of FLAN-PaLM 540B, we used an input length and output length of 512. We tuned 100 prompt embeddings and used them during inference. We used a learning rate of 0.3 and did not use dropout. We trained with a batch size of 16 for $24,544$ steps that is equivalent to one epoch on the MNLI dataset with $392,702$ training examples. The prompt-tuning took around 110 hours to complete with 256 Cloud TPU v4 chips.

C Hyper-parameter Details of T5 Fine-tuning Experiments

For training T5 models (both binary and 3-way), we tuned learning rates $\in \{5e-4, 1e-4, 5e-5\}$ and fine-tuned with batch size of 32 for 50K steps. We checkpoint every $1K$ steps for early stopping. We use a dropout rate of 0.1. We trained with an input length of 512. During inference for factual consistency evaluation (Section 4.1) and NLI benchmarks (Section 4.2), we used an input length of 1024 and 512 respectively.

We report the best selected hyper-parameters for T5 binary and 3-way models in Table 8 and Table 9.

Domain	Length	Text
news headlines	short	Congress approves debt deal, averting a US default
news headlines	short	Man airlifted to hospital from Skye beauty spot
news	short	Expectations were set high by the WSC concerning what the event would do for upcoming Indian entrepreneurs.
news	short	But despite high promises, it didn't take long for the first day of the convention to be plunged into chaos.
shopping reviews	paragraph	Good value for the seventy eight dollars that I paid for it. easy to change the filter. Quite on high. Haven't had it long enough to say how well it filters the air but I can see lint and dust on the filter pre screen. And I've only had it nine days I think. Love that I can turn the lights off.
shopping reviews	short	my first impressions are that's the Google Pixel 7 is a nice phone, BUT not as good as the moto g power in terms of ease of use and functionality.
shopping reviews	short	Battery has yet to be determined on the Pixel, but from a full charge, I'm down to 56% after 2 hours of use.
wikipedia	paragraph	Alfred was baptised by Frederick Cornwallis, Archbishop of Canterbury, in the Great Council Chamber at St James's Palace on 21 October 1780. His godparents were his elder siblings George, Prince of Wales; Prince Frederick; and Charlotte, Princess Royal. Alfred was a delicate child.
wikipedia	short	The premise of Two Hundred Rabbits was based on a dream that author Lonzo Anderson had after reading a French folk tale.
movie reviews	paragraph	As usual, James Cameron shows us his creative genius. The story is very different from the first, and I don't want to give out any story until you've seen it. It is worth watching, and if you own the first it is also worth buying. My only complaint, and it is BIG, is it turns out to only be in 480p resolution...not even 1080p or 4K. It looks good if you play it in YouTube, but still. It should be in 4K.
movie reviews	short	The actor portraying Mr. Darcy had no concept of the kind of man Darcy is or his nature.
place reviews	paragraph	Beautiful space which is nicely a bit secluded from the hussle at coal drop but still easy to reach. Wines were excellent, cheeses delicious, food great, and cocktails outstanding. Folks were kind and professional. Crowd was elegant but relaxed. Amazed they just opened three days ago, they operate like they have been at it forever. Loved every minute!
place reviews	short	The steep stairs need to be negotiated with caution especially after indulging in bout of revelry.
place reviews	short	I waited an hour. The doctor was terribly stressed. She didn't answer questions.
twitter	short	Sevilla is **Red and White** ♡
twitter	short	Lil X just asked if there are police cats, since there are police dogs :))
reddit post	paragraph	Hey there everyone! I often see people asking where to start when getting into prog metal, so I thought instead of answering every one of them individually I'd make a list. I'm not going into too much depth because otherwise this will become endless, but I'll try to give a brief explanation of all styles I'm going over. So let's get started!
reddit post	short	I am someone who hates doing laundry.

Table 7: Seed Examples for Prompting.

Dataset/Model	T5 small	T5 large	T5 XXL
MNLI	lr=5e-4, steps=40K	lr=5e-4, steps=10K	lr=5e-4, steps=15K
ANLI	lr=5e-4, steps=5K	lr=5e-4, steps=20K	lr=5e-4, steps=40K
WANLI	lr=5e-4, steps=15K	lr=5e-4, steps=10K	lr=1e-4, steps=50K
GNLI	lr=5e-4, steps=40K	lr=5e-4, steps=20K	lr=5e-5, steps=40K
MNLI + GNLI	lr=5e-4, steps=20K	lr=5e-4, steps=35K	lr=5e-5, steps=25K
ANLI + GNLI	lr=5e-4, steps=50K	lr=5e-4, steps=20K	lr=5e-4, steps=50K
WANLI + GNLI	lr=5e-4, steps=45K	lr=5e-4, steps=30K	lr=5e-5, steps=50K

Table 8: Best selected hyper-parameters for T5 binary models. We report learning rates (lr) and the number of steps.

	T5 small	T5 large	T5 XXL
MNLI	lr=5e-4, steps=40K	lr=5e-4, steps=10K	lr=5e-5, steps=35K
ANLI	lr=5e-4, steps=45K	lr=5e-4, steps=10K	lr=5e-4, steps=15K
WANLI	lr=5e-4, steps=10K	lr=5e-4, steps=10K	lr=5e-5, steps=15K
GNLI	lr=5e-4, steps=50K	lr=5e-4, steps=15K	lr=5e-5, steps=35K
MNLI + GNLI	lr=5e-4, steps=25K	lr=5e-4, steps=45K	lr=5e-4, steps=45K
ANLI + GNLI	lr=5e-4, steps=05K	lr=5e-4, steps=45K	lr=5e-5, steps=45K
WANLI + GNLI	lr=5e-4, steps=50K	lr=5e-4, steps=35K	lr=5e-5, steps=35K

Table 9: Best selected hyper-parameters for T5 3-way classification models. We report learning rates (lr) and the number of steps.