Evaluating Prompts Across Multiple Choice Tasks In a Zero-Shot Setting

Gabriel Orlanski
go533@nyu.edu

Abstract

Large language models have shown that impressive zero-shot performance can be achieved through natural language prompts (Radford et al., 2019; Brown et al., 2020; Sanh et al., 2021). Creating an effective prompt, however, requires significant trial and error. That *prompts* the question: how do the qualities of a prompt effects its performance? To this end, we collect and standardize prompts from a diverse range of tasks for use with tasks they were not designed for. We then evaluate these prompts across fixed multiple choice datasets for a quantitative analysis of how certain attributes of a prompt affect performance. We find that including the choices and using prompts not used during pre-training provide significant improvements. All experiments and code can be found here.

1 Introduction

Recent work has shown that using a natural language (NL) prompt with pre-trained language models (LM) significantly improves performance in few-shot and zero-shot settings (Brown et al., 2020; Schick and Schütze, 2021b), to the point where they can be worth 100s of data points (Scao and Rush, 2021). Further, T5 (Raffel et al., 2020) showed that simple prompts and reformulating NLP tasks as text-to-text performs well on a wide range of tasks. Recent models such as FLAN (Wei et al., 2021b) and T0 (Sanh et al., 2021) demonstrate that multi-task training a large LM with prompts results in improved zero-shot performance on a wide range of tasks. However, manually designing a prompt is a non-trivial task due to the trial-and-error nature of the task (Jiang et al., 2020; Shin et al., 2021). Some works focus on "prompt programming"– best practices for designing prompts (Reynolds and McDonell, 2021; Liu et al., 2021a). While others have looked towards continuous "soft-prompts"– random vectors added to the input sequence, which are then fine-tuned (Zhong et al., 2021; Qin and Eisner, 2021). However, recent work has revealed that these prompts are susceptible to minor perturbations (Mishra et al., 2021).

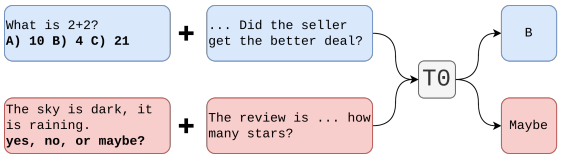

Figure 1: Overview of the approach. For a given example from a dataset (leftmost box) we use a prompt from a different task to perform zero shot predictions with T0-3B (Sanh et al., 2021). The **bolded text** in the example represents its choices.

Motivated by this, we aim to conduct a quantitative analysis of what affects a prompt's performance. We evaluate T0-3B (Sanh et al., 2021) on generalized prompts from a wide range of tasks with eight datasets to provide a quantitative analysis of how the qualitative aspects of a prompt effect its performance.

We collected 95 prompts across 20 tasks and evaluated each on eight datasets. We find that using a prompt performs better for every evaluation task than not using a prompt. Further, we find that the set of prompts with the highest performance is not the ones designed for the specific task for seven of the eight datasets. In our ablations, we find that adding a small amount of task-specific NL to the generalized prompt increases performance by a median of 4.65%. Finally, we find that prompts with the choices present outperform those that do not and that presenting the options as multiple distinct choices further improves the results. Additionally, the longer a prompt is, the worse it performs.

2 Methodology

Our overall approach is detailed in Figure 1. Given a downstream multiple-choice task $\mathcal{T}_o$ =

$\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, evaluate how well T0 (Sanh et al., 2021) performs when using a prompt $\mathcal{P}_\text{d}$ designed for a different task $\mathcal{T}_\text{d}$.

2.1 Fixed Choice Tasks

For the purposes of this paper, we limit the scope of the tasks we look at to be only multiple choice tasks whose choices are constant across all examples. Thus, for $\mathcal{T}_\text{o} = \{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$, every $\mathbf{y}_i \in \mathcal{C}_\text{o}$ where $\mathcal{C}_\text{o} = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$ with lengths $\ell = \{\ell_1, \ldots, \ell_c\}$. To make a prediction for the example point $\mathbf{x}_i$, we follow Holtzman et al. (2021) and use rank scoring: taking the choice with the highest probability as defined by

$$\underset{\mathcal{C}_j \in \mathcal{C}_\text{o}}{\text{argmax}} \prod_{k=1}^{\ell_j} P(\mathcal{C}_j^k | \mathbf{x}_i, \mathcal{C}_j^1 \ldots \mathcal{C}_j^{k-1}) \quad (1)$$

However, the lengths in ℓ are not guaranteed to be the same and thus Equation 1 will unintentionally penalize longer choices (Brown et al., 2020; Holtzman et al., 2021). We thus follow prior work and take the choice with the highest Average Log-Likelihood

$$\underset{\mathcal{C}_j \in \mathcal{C}_\text{o}}{\text{argmax}} \frac{\sum_{k=1}^{\ell_j} \log[P(\mathcal{C}_j^k | \mathbf{x}_i, \mathcal{C}_j^1 \ldots \mathcal{C}_j^{k-1})]}{\ell_j} \quad (2)$$

2.2 Generalized Prompts

We use the PromptSource[1] framework proposed by Sanh et al. (2021) for templates as it provides a standardized format for managing prompts. We standardize the input fields and answer formats across all tasks such that they fell into three general categories: CLASSIFICATION, ENTAILMENT, and QUESTION ANSWERING (QA). Table 1 displays an example of what the generalization would look like for an ENTAILMENT prompt. To use a prompt with a task that does not have the same number of inputs, we add additional task specific NL to better align the inputs. To use the example prompt from Table 1 with a sentiment classification task, we would map the input text from the task to the `premise` field and pass "`what is the sentiment`" as the hypothesis. In prompts where answers choices are present, we replace them with an additional input field for the choice string(i.e. "yes, no, or maybe") to hold how the choices are presented constant across all prompts. A detailed breakdown of the tasks used for the generalized prompts can be found in Table 3.

[1] https://github.com/bigscience-workshop/promptsource

3 Experimental Setup

3.1 Datasets

For all datasets, we use the most recent version on HuggingFace (Wolf et al., 2020). Every evaluation is done using the validation split as per Sanh et al. (2021). For the evaluation datasets, we again follow Sanh et al. (2021) and use: Adversarial NLI (**ANLI**) (Nie et al., 2020), CommitmentBank (**CB**) (De Marneff et al., 2019), Recognizing Textural Entailment (**RTE**) (Dagan et al., 2005; Bar-Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009), and Words In Context (**WiC**) (Pilehvar and osé Camacho-Collados, 2018).

Two additional tasks are used to evaluate T0's performance on complex tasks in unseen domains: **Algebra Question Answering (AQuA)** (Ling et al., 2017) and **CraigslistBargains** (He et al., 2018). Descriptions for these two tasks can be found in Appendix B.

For the generalized prompts, we collect 86 prompts across 19 distinct tasks. In addition to these, we also include 12 prompts with no additional NL for each of the three categories for 4 different ablations. More details can be found in Appendix A.

3.2 Model and Metrics

We evaluate the performance of the 3B parameter T0, and T5 (Raffel et al., 2020) models with the HuggingFace implementation (Wolf et al., 2020) as we were limited to a single RTX Titan 24GB card. Following prior works (Sanh et al., 2021; Brown et al., 2020; Holtzman et al., 2021), we report the accuracy and F1 scores on each of the eight datasets.

As each task will have different mean metrics, we cannot compare the raw accuracy and F1 scores across tasks. For a given prompt, we calculate the median accuracy and F1 ranks compared to the 95 other prompts for all evaluation tasks. As the rank is ascending, **lower values** for median accuracy rank (MAR) and median F1 rank (MFR) indicate a better performing prompt.

4 Results

4.1 Baselines On New Tasks

As we evaluate T0-3B on two new tasks, we first want to gauge how the model performs as shown in both Figure 2 and Table 4. We follow Sanh et al. (2021) in using rank scoring without length nor-

Task	Prompt
Original	`Sentence A: {{sentence1}} Sentence B: {{sentence2}} "{{word}}" has a similar meaning in sentences A and B. True or False?`
Generalized	`Sentence A: {{premise}} Sentence B: {{hypothesis}} "{{domain}}" has a similar meaning in sentences A and B. {{ choice_string }}?`
Example	`Sentence A: What is 2+2? Sentence B: Choices are: ∞, -10, fish, 4, √2 "math problem" has a similar meaning in sentences A and B. "A", "B", "C", "D" or "E"?`

Table 1: Sample prompt from WordsInContext Task (Pilehvar and osé Camacho-Collados, 2018) and its generalized form. Each {{ }} represents an input from the dataset. The colors are the alignment of inputs.

malization as defined by Equation 1 and find that for both AQuA and CraigslistBargains, the base T5 model performs better than T0. However, T0's unweighted multi-class F1 is better than that of T5 on both tasks, indicating that T5 achieves a higher score due to only predicting a subset of the choices. Figure 3 provides further evidence for this hypothesis. In both AQuA and CraigslistBargains, T0 more evenly distributes its predictions across the possible choices where as T5 heavily favors a subset of the choices. Thus, the disparity in the accuracy is likely a result of class imbalance in the evaluation datasets in which T5 is 'lucky' in heavily predicting the class that was more populous. We consider this to be 'luck' as T5 outperforming T0 only occurs in only two of the datasets we examined.

4.2 Generalized Prompts

We report the results of the cross-task evaluation in Table 2. We find that the only task in which the original[2] prompt performs best is ANLI R2 with an accuracy of 34.70. Conversely, the worst performing prompts for the AQuA task were its original prompts with an accuracy of 17.32. Furthermore, we find that there is no task out of the eight used for evaluation where not using a prompt has the best performance.

We report how the added NL discussed in subsection 2.2 effects the zero-shot performance in Table 5. Across the eight evaluation tasks, adding some task specific text leads to an average increase of 4.65% in the accuracy and a 2.34% increase to the unweighted multiclass F1. However, there was also a decrease of 4.21% and 13.90% to minimum accuracy and unweighted multiclass F1 respectively. This implies that the added extra text helps to better amplify the negative and positive elements of a prompt.

4.3 Qualitative Analysis of Prompts

Table 5 displays the rank statistics across multiple ablations and Table 7 displays the correlations of a prompt's qualities with their rank. In prompts which have choices, the MAR is 33.12 compared to 52.25 when the choices are left out. However, the range as indicated by the Q1 and Q3 MARs is significantly larger when the choices are included, implying that adding the choice string causes high variance.

We also find that when the choices are presented as multiple distinct choices[3] the MAR is further improves to 22.25. In comparison, the prompts with choices that are not in this format have a MAR of 36.00. These results provide further evidence to the findings from Wei et al. (2021a) that clearly distinguishing the options in a prompt improves performance.

Next, we find that the median rank of prompts used in training is 50.25 compared to 42.00 of the unseen prompts. The F1 scores display a similar pattern with training prompts having a MFR of 55.75 while unseen prompts have a median of 36.50. Although this implies that prompts that share tokens with those used for training will perform worse, we find that there is no significant correlation between the number of tokens a prompt $\mathcal{P}$ shares with those used in training and its rank.

Finally, we find that longer prompts have a slightly negative impact on the performance of a prompt. Figure 4 shows that, with the 95 prompts used across eight tasks, the best performing prompts are those whose length is in the range

[2]By original we are referring to the prompts specifically designed for a task.

[3]Presented with a clear delimiters/separation. For example A) yes B) no C) maybe

		ANLI R1	ANLI R2	ANLI R3	AQuA	CB	Craigslist	RTE	WiC	Rank
	No Prompt	34.15	33.35	33.42	26.77	24.11	16.83	59.57	50.24	46.25
Unseen Prompts	ANLI	**37.60**	**34.70**	**34.08**	25.95	**32.14**	21.44	64.62	50.16	24.50
	AQuA	36.10	33.40	35.42	**17.32**	33.93	23.45	71.12	51.57	18.25
	COPA	39.30	34.40	34.00	20.47	26.79	16.58	69.31	50.63	21.25
	Craigslist	31.40	31.30	32.83	25.79	8.04	**26.72**	49.82	50.16	71.25
	MathQA	37.30	33.50	34.25	19.29	26.79	16.25	73.29	51.10	24.50
	RTE	36.10	33.20	33.58	22.05	23.21	20.27	**61.37**	50.47	43.25
	SemEval2010	33.10	32.00	32.58	27.56	14.29	25.63	55.23	50.47	66.50
	WiC	31.75	33.45	32.33	26.57	13.39	18.01	55.05	**50.47**	64.25
Training Prompts	AppReviews	34.20	33.10	33.62	27.17	19.64	33.17	61.55	50.31	33.50
	IMDB	33.00	32.20	33.08	26.38	12.50	14.57	55.23	50.16	71.25
	Yelp	33.25	32.35	33.04	26.77	12.50	24.29	62.27	51.57	41.75

Table 2: Median Accuracy when using modified prompts for cross task zero-shot evaluation. **Bolded** entries are prompts for the original task. Green Cells and Red Cells are the best and worst performing tasks for a column respectively. Rank is the median rank of prompts from this task out of 95 total prompts. ANLI and CB both use the same prompts for their original task prompts per PromptSource. Some tasks are left out for clarity. The full table can be found in Table 6.

$[14, 21)$ as their MAR is 28.50 and MFR is 36.50. The Q1 values are 18.00 and 15.38, respectively. In comparison, we find that prompts with lengths with lengths ≥ 25 have a median MAR of 50.25 and MFR of 72.00, indicating a negative impact on performance. Surprisingly, we find that prompts whose lengths are < 14 have a median MAR of 47.75 and MFR of 49.00. While this is a negative impact on performance, it is not as large as that in longer prompts.

5 Related Works

Pre-trained Language Models In the past few years, large pre-trained models have rose to prominence due to their strong performance on a wide range of NLP tasks (Radford et al., 2019; Brown et al., 2020; Lewis et al., 2020). In particular, T5 (Raffel et al., 2020) explored transferred learning for large LMs by transforming all NLP problems to a text-to-text format. One aspect of large LMs is that they perform well in zero-shot settings (Radford et al., 2019; Brown et al., 2020; Vu et al., 2020).

NL Prompting A drawback of these large LMs is that their size makes it costly to fine-tune them. This lead to the rise of the "*pre-train, prompt, and predict*" paradigm in which a downstream tasks are modified to resemble those used in training through the use of NL prompts (Liu et al., 2021b). These prompts have improved few-shot and zero-shot performance across a vast number of models and tasks (Brown et al., 2020; Schick and Schütze, 2021b,a; Mishra et al., 2021; Scao and Rush, 2021; Shin et al., 2020). Recent models such as FLAN (Wei et al., 2021b) and T0 (Sanh et al., 2021) have shown that even better zero-shot performance can be achieved through using a multi-task pre-training objectives with a diverse set of prompts.

6 Conclusion

In this paper, we examined T0's performance on a range of fixed multiple-choice tasks. We find that T0 does worse than T5 on two unseen complex tasks. Next, we evaluated how the performance of a prompt transfers between tasks. Our results show that using a prompt performs consistently better than not using any prompt. We conclude with a quantitative analysis of what aspects of a prompt affect its performance. We find that prompts with the choices in them are 66.82% better than those that leave the choices out. Next, we find that prompts not used in pre-training are 19.64% better than those that were. Finally, we find that prompts whose length are between 14 and 24 perform better than both longer and shorter prompts. Further work should examine prompt transfer with larger models while also expanding the number of prompts and tasks used.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based for-

malisms. 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 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the second PASCAL challenges workshop on recognising textual entailment*, volume 6, pages 6–4. Venice.

Max Bartolo, Alastair Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. In *TAC*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer.

Marie-Catherine De Marneff, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *proceedings of Sinn und Bedeutung 23*.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL workshop on textual entailment and paraphrasing*, pages 1–9. Association for Computational Linguistics.

Giovanni Grano, Andrea Di Sorbo, Francesco Mercaldo, Corrado A. Visaggio, Gerardo Canfora, and Sebastiano Panichella. 2017. Android apps and user feedback: A dataset for software evolution and quality improvement. In *Proceedings of the 2nd ACM SIGSOFT International Workshop on App Market Analytics*, WAMA 2017, page 8–11, New York, NY, USA. Association for Computing Machinery.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343, Brussels, Belgium. Association for Computational Linguistics.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid 'O S'eaghdha, Sebastian Pad'o, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38, Uppsala, Sweden. Association for Computational Linguistics.

Ari Holtzman, Peter West, Vered Schwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface form competition: Why the highest probability answer isn't always right. In *EMNLP*.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know?

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020a. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of EMNLP*. To appear.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020b. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2021a. What makes good in-context examples for gpt-3?

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ArXiv*, abs/2107.13586.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hanna Hajishirzi. 2021. Cross-task generalization via natural language crowdsourcing instructions.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745.

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*. Association for Computational Linguistics.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Mohammad Taher Pilehvar and osé Camacho-Collados. 2018. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, 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, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Teven Le Scao and Alexander M. Rush. 2021. How many data points is a prompt worth? In *NAACL*.

Timo Schick and Hinrich Schütze. 2021a. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Timo Schick and Hinrich Schütze. 2021b. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2339–2352, Online. Association for Computational Linguistics.

Richard Shin, Christopher H. Lin, Sam Thomson, Charles Chen, Subhro Roy, Emmanouil Antonios Platanios, Adam Pauls, Dan Klein, Jason Eisner, and Benjamin Van Durme. 2021. Constrained language models yield few-shot semantic parsers.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Tu Vu, Tong Wang, Tsendsuren Munkhdalai, Alessandro Sordoni, Adam Trischler, Andrew Mattarella-Micke, Subhransu Maji, and Mohit Iyyer. 2020. Exploring and predicting transferability across NLP tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7882–7926, Online. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021a. Finetuned

language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021b. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Orion Weller, Nicholas Lourie, Matt Gardner, and Matthew E. Peters. 2020. Learning from task descriptions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1361–1375, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall.

A Generalized Prompts

The tasks we took prompts from that were not used to train T0 are:

- COPA (Roemmele et al., 2011)
- FinancialNews (Malo et al., 2014)
- LAMBADA (Paperno et al., 2016)
- MathQA (Amini et al., 2019)
- MultiXSci (Lu et al., 2020)
- NumerSense (Lin et al., 2020a)
- SemEval2010 (Hendrickx et al., 2010)
- ZEST (Weller et al., 2020)

The tasks we took prompts from that *were* used to train T0 are:

- AppReviews (Grano et al., 2017)
- Adversarial QA (Bartolo et al., 2020)
- CommonGen (Lin et al., 2020b)
- IMDB (Maas et al., 2011)
- XSum (Narayan et al., 2018)

Task Name	#	Type	MCQ
Text	6	All Three	✓
Text+Choices	6	All Three	✓
ANLI	9	ENTAILMENT	
AQuA	5	CLASSIFICATION	✓
COPA	3	ENTAILMENT	
Craigslist	4	CLASSIFICATION	✓
FinancialNews	4	CLASSIFICATION	
LAMBADA	3	CLASSIFICATION	
MathQA	5	CLASSIFICATION	✓
MultiXSci	4	CLASSIFICATION	
NumerSense	5	CLASSIFICATION	
RTE	5	ENTAILMENT	
SemEval2010	3	CLASSIFICATION	
WiC	4	ENTAILMENT	
ZEST	4	QA	
AppReviews	2	CLASSIFICATION	
AdversarialQA	4	QA	
CommonGen	2	CLASSIFICATION	
IMDB	5	CLASSIFICATION	
XSum	4	CLASSIFICATION	
Yelp	4	CLASSIFICATION	

Table 3: Number of prompts used by task for the generalized prompts. # is the number of prompts used from this task. MCQ indicates if the task had prompts that are formatted as an multiple choice question with choice letters.

B Complex Task Datasets

The two complex task used to evaluate T0's performance are:

Algebra Question Answering (AQuA) Dataset of multiple choice algebraic word problems. The choices for this task are {A,B,C,D,E} and each letter maps to a potential mathematical answer (Ling et al., 2017).

CraigslistBargains A collection of dialogues involving two-parties negotiating the price of an item for sale on Craigslist. For the scope of this paper, we use the task of classifying who won the negotiation (He et al., 2018).

C Additional Results

This section is for results that could not be included in the main body of the paper due to the page limits.

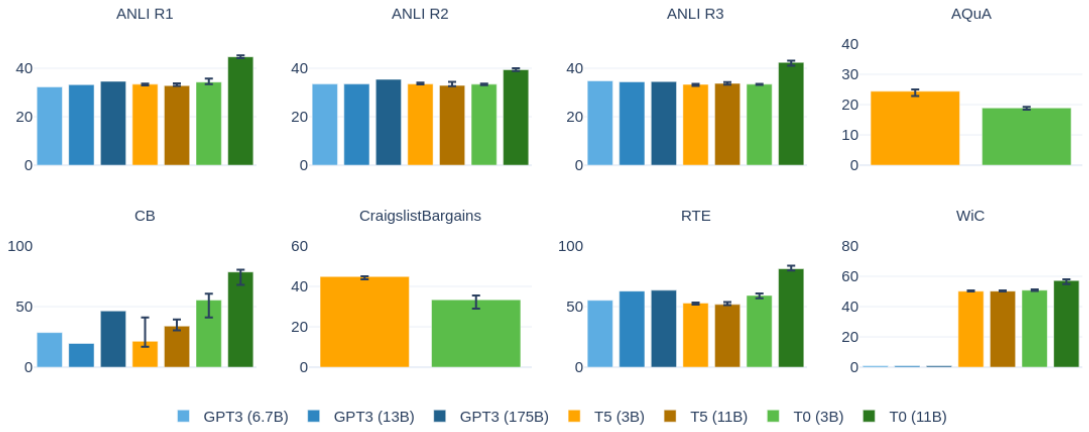

Figure 2: Median Accuracy with interquartile range for three models: **GPT-3**, **T5**, and **T0**. Darker indicates larger model. Results for GPT-3 Model are from Brown et al. (2020). Results for the 11B T0 and T5 models are taken from Sanh et al. (2021)

Task	Model	Accuracy	F1
ANLI R1	T0	34.30	26.17
	T5	33.40	20.28
ANLI R2	T0	33.40	23.70
	T5	33.50	21.25
ANLI R3	T0	33.42	21.82
	T5	33.33	24.84
AQuA	T0	18.90	15.04
	T5	24.41	12.74
CB	T0	55.36	38.62
	T5	21.43	19.41
CraigslistBargains	T0	33.42	18.45
	T5	44.89	16.47
RTE	T0	59.21	69.56
	T5	52.89	36.08
WiC	T0	50.86	8.81
	T5	50.16	5.44

Table 4: Median accuracy and F1 for the corresponding Figure 2.

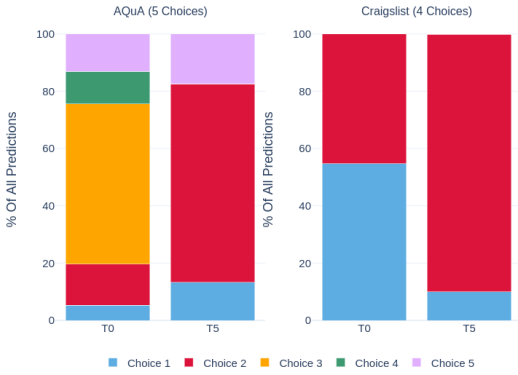

Figure 3: Distribution of choices for T0 and T5 on the AQuA and CraigslistBargains.

	Accuracy	F1
Has Choices	-0.14	-0.27
Is MCQ	-0.16	-0.25
Training Prompt	0.07	0.13
Length	0.14	0.18

Table 7: Correlations with metric rank for a given prompt quality. Per the definition of rank, a lower score is better and therefore a negative correlation indicates a quality improves performance. Length is measured as the raw number of tokens in a prompt.

Ablation	Accuracy				F1			
	Mean	Median	Q1	Q3	Mean	Median	Q1	Q3
Training Prompts	51.46	50.25	45.25	63.00	54.18	55.75	38.00	66.00
Unseen Prompts	42.72	42.00	23.50	60.75	42.46	36.50	22.00	62.50
With Choices	39.44	33.12	20.19	58.62	39.37	31.00	19.12	61.75
No Choices	51.73	52.25	44.75	60.50	55.93	53.50	43.00	66.00
Is MCQ	25.80	22.25	16.50	26.00	23.14	16.50	13.75	25.25
Not MCQ	43.28	36.00	26.38	62.62	43.95	36.50	22.00	65.50
Extra Text	44.99	46.75	28.81	60.31	46.44	46.50	27.75	66.00
No Extra Text	44.41	48.00	32.75	57.62	45.43	44.50	26.62	62.25

Table 5: Accuracy and F1 ranks for different ablations. It is calculated by taking the median rank of a given prompt across all 8 tasks then taking the Mean, Median, Q1, and Q3 of that. Lower is better. Q1 and Q3 are the first and third quartile.

		ANLI R1	ANLI R2	ANLI R3	AQuA	CB	Craigslist	RTE	WiC	Rank
	No Prompt	34.15	33.35	33.42	26.77	24.11	16.83	59.57	50.24	46.25
Unseen Prompts	ANLI	**37.60**	**34.70**	**34.08**	25.95	**32.14**	21.44	64.62	50.16	24.50
	AQuA	36.10	33.40	35.42	**17.32**	33.93	23.45	71.12	51.57	18.25
	COPA	39.30	34.40	34.00	20.47	26.79	16.58	69.31	50.63	21.25
	Craigslist	31.40	31.30	32.83	25.79	8.04	**26.72**	49.82	50.16	71.25
	FinNews	33.05	31.65	32.83	25.95	18.75	19.68	55.78	50.31	64.00
	LAMBADA	34.00	32.40	32.50	26.77	19.64	16.08	57.76	50.78	58.50
	MathQA	37.30	33.50	34.25	19.29	26.79	16.25	73.29	51.10	24.50
	Multi-XSci	34.20	32.70	32.75	27.17	19.64	19.43	58.84	50.31	54.75
	NumerSense	37.70	33.30	33.17	25.20	25.00	15.75	65.70	50.63	40.50
	RTE	36.10	33.20	33.58	22.05	23.21	20.27	**61.37**	50.47	43.25
	SemEval2010	33.10	32.00	32.58	27.56	14.29	25.63	55.23	50.47	66.50
	WiC	31.75	33.45	32.33	26.57	13.39	18.01	55.05	**50.47**	64.25
	ZEST	35.20	32.65	33.38	26.77	23.21	17.76	66.79	50.71	38.25
Training Prompts	AppReviews	34.20	33.10	33.62	27.17	19.64	33.17	61.55	50.31	33.50
	CommonGen	33.75	33.35	32.50	25.39	13.39	23.62	51.81	51.18	58.75
	IMDB	33.00	32.20	33.08	26.38	12.50	14.57	55.23	50.16	71.25
	XSum	33.50	32.00	33.00	26.97	10.71	19.26	57.22	50.86	58.50
	Yelp	33.25	32.35	33.04	26.77	12.50	24.29	62.27	51.57	41.75

Table 6: Median Accuracy when using modified prompts for cross task zero-shot evaluation. **Bolded** entries are prompts for the original task. Green Cells and Red Cells are the best and worst performing tasks for a column respectively. Rank is the median rank of prompts from this task out of 95 total prompts. ANLI and CB both use the same prompts for their original task prompts per PromptSource.

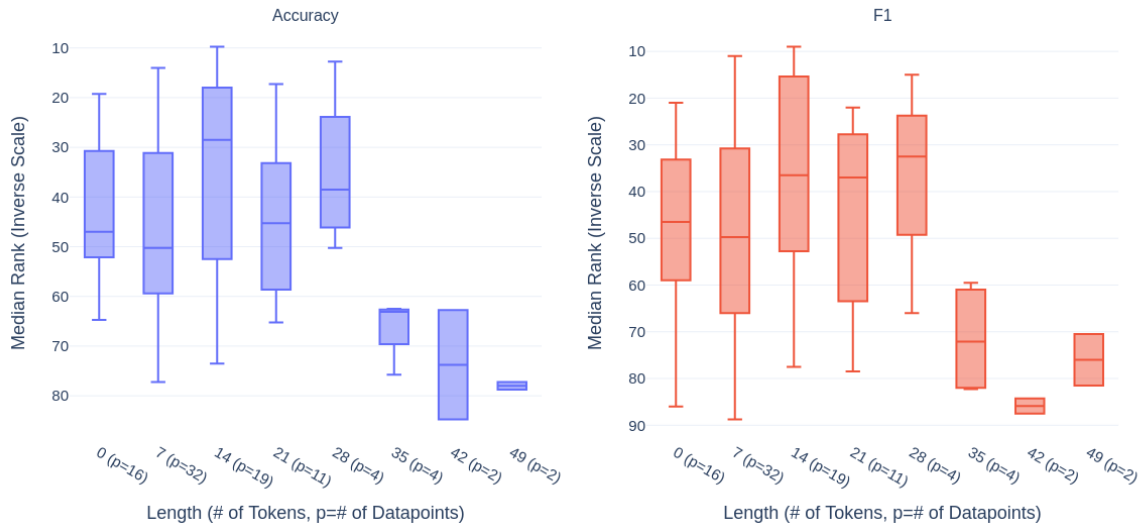

Figure 4: Accuracy and F1 rank compared with the number of tokens in the prompt. The tick value is the lower bound of the range. p=The number of prompts that fall into that respective range.