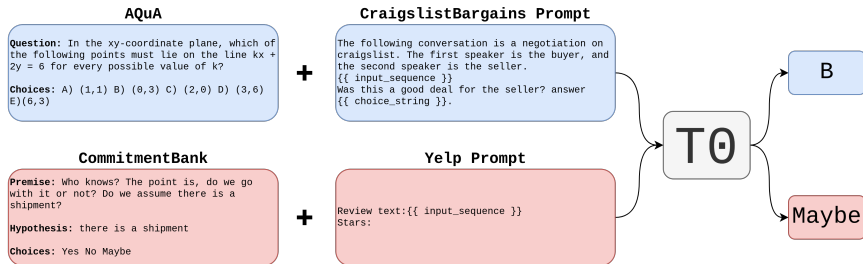


# Using Cross-Task Prompts in the Zero Shot Setting

[REDACTED]

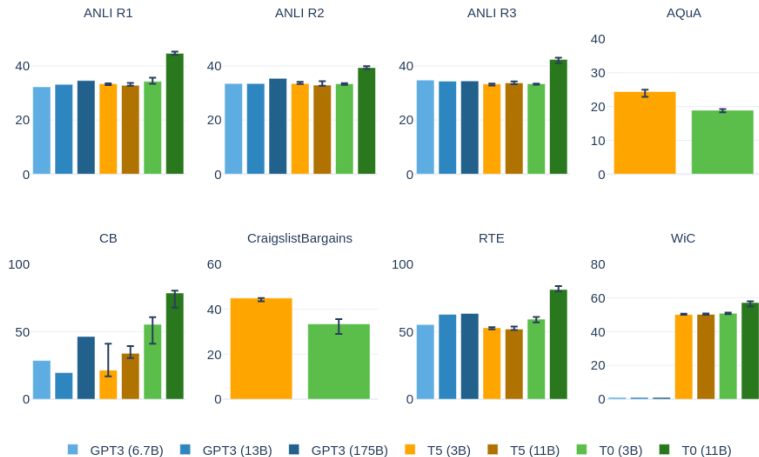
December 8, 2021

# Project Overview



**Figure:** High level 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. The bolded text in the example represents its fields and choices.

# Baselines



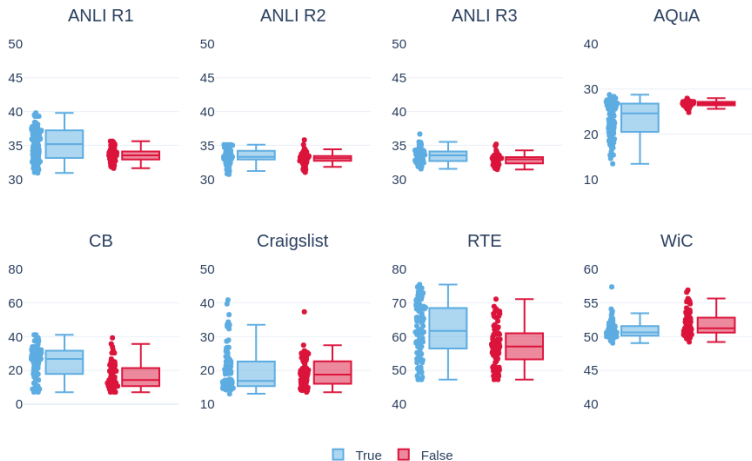
**Figure:** Median Accuracy with interquartile range for three models: **GPT-3**, **T5**, and **T0**. Darker indicates larger model.

# Cross Task Results - Accuracy

		ANLI R1	ANLI R2	ANLI R3	AQuA	CB[5]	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 [14]	<b>37.60</b>	<b>34.70</b>	<b>34.08</b>	25.98	<b>32.14</b>	21.44	64.62	<b>50.16</b>	24.50
	AQuA[11]	36.10	33.40	<b>35.42</b>	<b>17.32</b>	<b>33.93</b>	23.45	71.12	51.57	<b>18.25</b>
	COPA[16]	<b>39.30</b>	34.40	34.00	20.47	26.79	16.58	69.31	50.63	21.25
	Craigslist[8]	<b>31.40</b>	<b>31.30</b>	32.83	25.79	<b>8.04</b>	<b>26.72</b>	<b>49.82</b>	50.16	<b>71.25</b>
	MathQA[1]	37.30	33.50	34.25	19.29	26.79	16.25	<b>73.29</b>	51.10	24.50
	RTE[3, 6, 2, 4]	36.10	33.20	33.58	22.05	23.21	20.27	<b>61.37</b>	50.47	43.25
	SemEval2010[9]	33.10	32.00	32.58	<b>27.56</b>	14.29	25.63	55.23	50.47	66.50
	WiC [15]	32.75	33.45	<b>32.33</b>	26.57	13.39	18.01	55.05	<b>50.47</b>	64.25
Training Prompts	AppReviews[7]	34.20	33.10	33.62	27.17	19.64	<b>33.17</b>	61.55	50.31	33.50
	CommonGen[10]	33.75	33.35	32.50	25.39	13.39	23.62	51.81	51.18	58.75
	IMDB[12]	33.00	32.20	33.08	26.38	12.50	<b>14.57</b>	55.23	50.16	71.25
	XSum[13]	33.50	32.00	33.00	26.97	10.71	19.26	57.22	50.86	58.50
	Yelp[17]	33.25	32.35	33.04	26.77	12.50	24.29	62.27	<b>51.57</b>	41.75

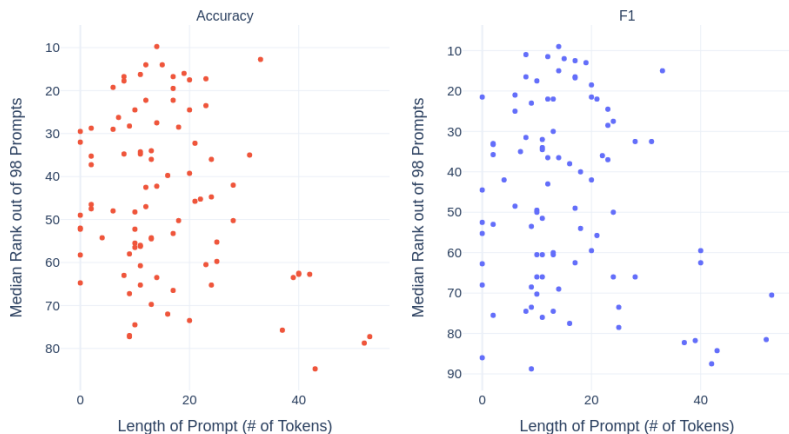
**Table:** Median Accuracy when using modified prompts for cross task. **Bolded** entries are prompts for the original task. **Green Cells** and **Red Cells** are the best and worst performing tasks for a column respectively. Rank is the median rank of prompts from this task out of 98 total prompts. Some tasks are not visualized for the sake of clarity.

# Choices in the Prompt



**Figure:** Comparison of the performance between prompts that have the choices in them compared with ones that do not. For AQuA, every prompt has the mathematical choices present, but when for prompts that do not present the choices, we hide the corresponding letter choices.

# Impacts of Prompt Length



**Figure:** Each dot represents one of the 98 prompts, with y-axis representing the median rank of the prompt across the 8 evaluation tasks. Length of the prompt is calculating excluding both choices and the choice phrase.

# References I

- [1] A. Amini, S. Gabriel, S. Lin, R. Koncel-Kedziorski, Y. Choi, and H. Hajishirzi. MathQA: Towards interpretable math word problem solving with operation-based formalisms. 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, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1245. URL <https://aclanthology.org/N19-1245>.
- [2] R. Bar-Haim, I. Dagan, B. Dolan, L. Ferro, D. Giampiccolo, B. Magnini, and I. Szpektor. 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, 2006.
- [3] L. Bentivogli, P. Clark, I. Dagan, and D. Giampiccolo. The fifth pascal recognizing textual entailment challenge. In *TAC*, 2009.
- [4] I. Dagan, O. Glickman, and B. Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

# References II

- [5] M.-C. De Marneff, M. Simons, and J. Tonhauser. The commitmentbank: Investigating projection in naturally occurring discourse. *proceedings of Sinn und Bedeutung* 23, 2019.
- [6] D. Giampiccolo, B. Magnini, I. Dagan, and B. Dolan. 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, 2007.
- [7] G. Grano, A. Di Sorbo, F. Mercaldo, C. A. Visaggio, G. Canfora, and S. Panichella. 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, 2017. Association for Computing Machinery. ISBN 9781450351584. doi: 10.1145/3121264.3121266. URL <https://doi.org/10.1145/3121264.3121266>.
- [8] H. He, D. Chen, A. Balakrishnan, and P. Liang. 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, Oct.–Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1256. URL <https://aclanthology.org/D18-1256>.



## References III

- [9] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. 'O S'eaghdha, S. Pad'ò, M. Pennacchiotti, L. Romano, and S. Szpakowicz. 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, July 2010. Association for Computational Linguistics. URL <https://www.aclweb.org/anthology/S10-1006>.
- [10] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, Nov. 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.165. URL <https://www.aclweb.org/anthology/2020.findings-emnlp.165>.
- [11] W. Ling, D. Yogatama, C. Dyer, and P. Blunsom. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *ACL*, 2017.
- [12] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. 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, June 2011. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P11-1015>.

# References IV

- [13] S. Narayan, S. B. Cohen, and M. Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *ArXiv*, abs/1808.08745, 2018.
- [14] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. 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, 2020.
- [15] M. T. Pilehvar and osé Camacho-Collados. Wic: 10, 000 example pairs for evaluating context-sensitive representations. *CoRR*, abs/1808.09121, 2018. URL <http://arxiv.org/abs/1808.09121>.
- [16] M. Roemmele, C. A. Bejan, and A. S. Gordon. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*, 2011.
- [17] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.