Dissecting Compositional Generalization: Correlation Analysis on Generalization Performance of Sequence-to-Sequence Models on Different Compositional Problems

Yoobin Cheong, Yeong Koh and **Yoon Tae Park**
Center for Data Science
New York University
60 5th Avenue, New York, NY
{yc5206, yk2678, yp2201}@nyu.edu

Abstract

Compositional generalization is the ability to generalize systematically to a new data distribution by combining known components (Montague, 1975). In detail, this ability can be defined as identifying smaller parts that the data can be decomposed into, and generalizing to novel data by combining these known, smaller parts. In this project, we provide an evaluation of 9 different compositional generalization datasets by using a Sequence-to-Sequence model, T5 (Raffel et al., 2019), with 20 randomly initialized weights. Also, we provide 36 correlation analysis results by selecting different pairs of tasks and calculating their correlation. Our result shows that some strong positive correlation tasks exist, indicating that those tasks recruit similar capacities. On the other hand, there are negative correlations or uncorrelated results which indicate that based on the model we tested, performance on one task does not entail similar performance on another task. This suggests that these negatively correlated or uncorrelated tasks require distinct solutions that go beyond a simple fine-tuning approach. We believe that extending the analyses to other models and tasks will help identify groupings of tasks that require distinct solutions, and lead to a better understanding of the nature of the problems labeled "compositional generalization".

1 Introduction

Compositional generalization is the ability to generalize systematically to a new data distribution by combining known components. This ability is often presented as a solution to generalize outside of the training data, but it is also considered that neural networks struggle to achieve. As a reflection to these challenges, there exist many benchmark tests that claim to evaluate the abstract capacity of compositional generalization. Some recent studies show significant improvements in instruction learning in SCAN using two representations (Li et al., 2019) or establish a new state of the art on the CFQ compositional generalization benchmark using MLM pre-training (Furrer et al., 2020). On the other hand, another recent study shows that the seq2seq models achieve near-zero accuracy on structural generalization tasks in COGS although they perform well on lexical tasks (Weißenhorn et al., 2022). These studies also get extended to other datasets, such as PCFGset (Hupkes et al., 2019), or a specific subset in compositional generalization as in the study of evaluating the capacity in Question and Answering tasks (Liu et al., 2021). While these studies enrich "compositional generalization", whether these tests that fall under this abstraction target the same underlying capacity is an open question.

In this project, we experiment on multiple tasks to evaluate performance given different weight initializations and study their correlations to look for any evidence of compositional generalization capacity. As a means to investigate this idea, we apply different instantiations of the T5 model across multiple tasks. Through this analysis, we examine whether good performance on one task is a predictor of good performance on a different compositional generalization task, which would suggest that these tasks recruit similar problem-solving capacities. Grouping of tasks that are identified as being highly correlated would be able to provide insights towards a better characterization of the compositional generalization, and furthermore be useful for task selection in multitask learning or transfer learning.

2 Experimental Design

2.1 Experiment steps

The initial experiment starts with testing a Sequence-to-Sequence model on a collection of

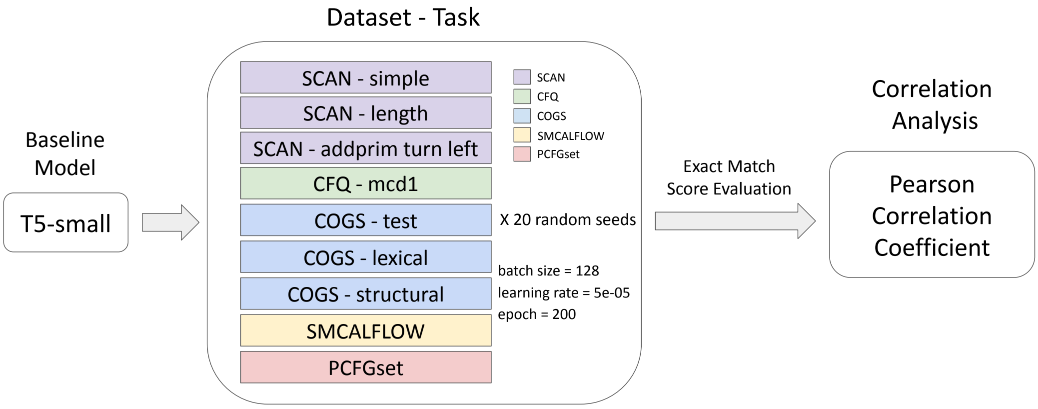

Figure 1: Experimental Design

publicly available benchmark datasets that propose to measure "compositional generalization". Our experiment can be summarized in 4 steps (Figure 1).

Step 1 Select task X and pick one initialization seed of the non-pretrained T5-small model. Finetune on task X and evaluate its performance based on the exact match score. Experiment with different hyperparameter settings to make sure that the expected exact match score (Furrer et al., 2020) is achieved. Then, repeat this process for 20 different initialization seeds using the selected hyperparameters.

Step 2 Perform the same experiment on another task, Y. Apply the same strategies as Step 1, but change the hyperparameters if the result does not meet the expected performance.

Step 3 Calculate Pearson Correlation Coefficient for all different combination of task pairs (X: selected task 1, Y: selected task 2, X != Y).

Step 4 Repeat Steps 1-3 for different task pairs, X and Y.

2.2 Experiment structure

We have defined our experimental structure as below:

Dataset We have used various datasets constructed to test compositional generalization tasks: SCAN (Lake and Baroni, 2018), CFQ (Keysers et al., 2019), COGS (Kim and Linzen, 2020), PCFG SET(Hupkes et al., 2019), and SMCalFlow (Andreas et al., 2020). As a baseline experiment, we select the SCAN dataset that consists of a set of commands and corresponding action sequences that an agent should perform where the two components are defined compositionally based on primitives and modifiers. Among several subsets of tasks included in the SCAN, we select "simple", "length", and "addprim turn left" dataset/tasks to compute correlations.

In addition, the experiments get expanded to other datasets: COGS (for subsets of test, lexical, and structural) and CFQ that are based on semantic parsing, SMCalFlow featuring natural conversations, and PCFG SET testing different aspects of compositional generalization (Table 1).

Dataset	Input (string)	Output (string)
SCAN simple	run opposite left after walk right	I_TURN_RIGHT I_WALK I_TURN_LEFT I_TURN_LEFT I_RUN
COGS lexical	A ball was blessed.	ball (x_1) AND bless . theme (x_3 , x_1) active_to_passive
CFQ mcd1	Who wrote M1 and wrote a film	SELECT DISTINCT ?x0 WHERE { M0 ns:film.producer.films executive$_{producedM1}$}
SMCcalFlow	what date is tomorrow?	(Yield (Tomorrow))
PCFGset	swap_first_last repeat copy J4 A9 N7 V8	V8 A9 N7 V8 J4 A9 N7 J4

Table 1: Examples of datasets

Models/architectures The model we use is T5-small, which is an encoder-decoder transformer architecture. In detail, we finetune the non-pretrained T5-small architecture with randomly initialized weights to compare the exact match scores of different dataset/tasks.

Experimental setup We use a batch size of 128, a learning rate of 5e-05, and 200 epochs as our default hyperparameter setting. For tasks where our default setting leads to an unexpectedly low performance, we perform further hyperparameter tuning by adjusting the batch size, learning rate,

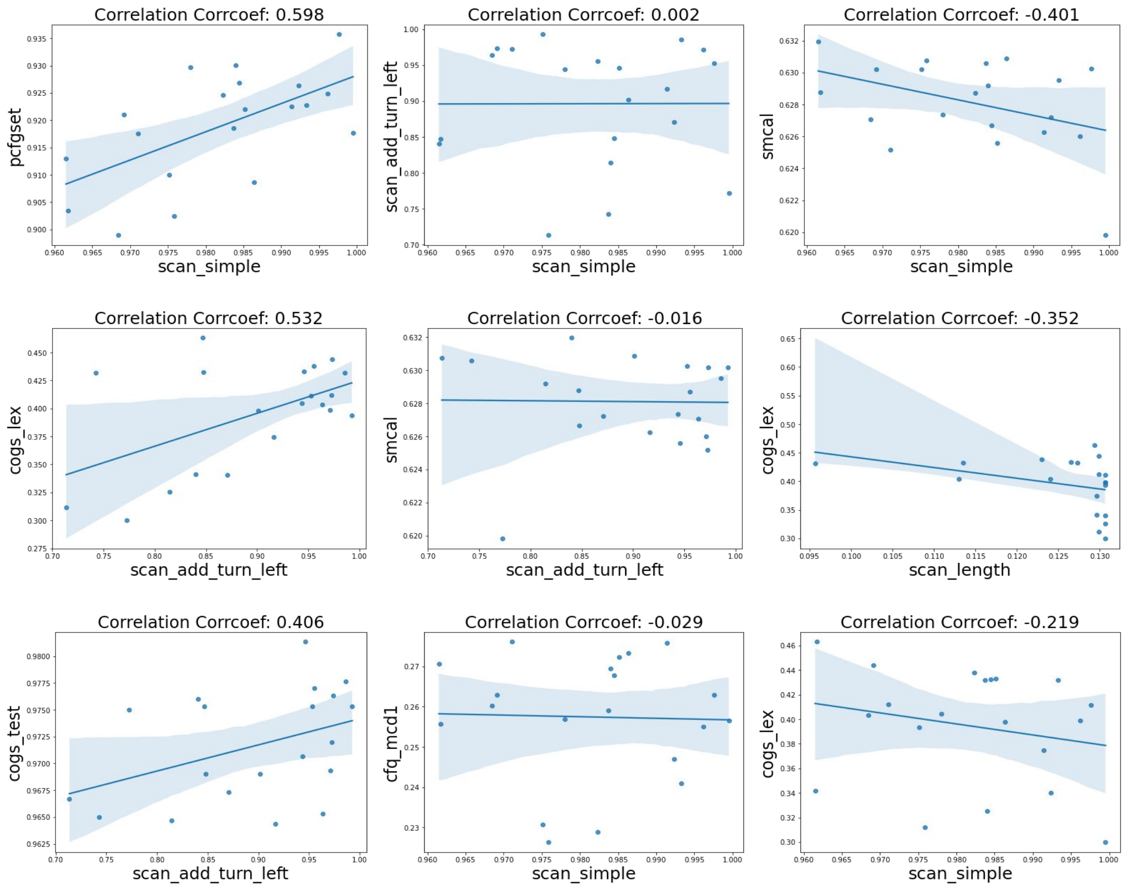

Figure 2: Top 3 Positive, Neutral, and Negative Correlations

and the total number of epochs and tracking the validation accuracy using the Tensorboard to look for convergence. To enhance consistency of each experiment, we test different weight instantiations for each model using 20 fixed random seeds from 0 to 19.

2.3 Evaluation

For each task, we evaluate the performance of the model using the exact match score metric. This metric compares the entire prediction output against the label, and therefore works as a strict evaluation for all tasks.

After we calculate the exact match scores for all tasks given different random seeds, we compare the performances of the models through Pearson Correlation Coefficient (Equation 1) to measure the degree of compositional generalization capacities between any pairs of tasks.

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

3 Results and Analysis

In total, we have conducted 36 experiments which consist of 9 different tasks in various datasets. In comparison to the mean correlation of 0.186, which is calculated across all pairs of tasks, our results show that there are some tasks that are highly correlated, negatively correlated and uncorrelated across the datasets (Figure 2).

The most highly correlated tasks were SCAN-simple and PCFGset with a correlation coefficient of 0.598. This result indicates that these two tasks recruit similar capacities, meaning that the skills that are required to solve these two datasets may be related. The correlations, however, can be low as 0.002 for other pairs of tasks such as the SCAN-simple and SCAN-add-turn-left pair, showing evidence that they may recruit different capacities even though these tasks are framed under compositional generalizations. A similar conclusion can be drawn for the most negatively correlated tasks such as the SCAN-simple and SMCalFlow pair. In other words, the performance on SCAN-simple does not provide any indication on performance of

SCAN-add-turn-left or SMCalFlow. Therefore, the problem-solving capacities needed to solve these tasks are not related and the two tasks require distinct solutions that go beyond a simple finetuning approach.

In addition to the correlation comparison, we observed that different initializations may affect the model's performance. Especially in the case of SCAN-add turn left, our experiments reveal that the accuracy of the model, which is measured by the exact match score, can fluctuate significantly, from 0.71 to 0.99, just by changing the random states. Other tasks also show some variations in exact match scores; the accuracy of the model varies by 0.16 in COGS-gen-lexical and 0.04 in PCFGset depending on the initialization setting.

Below table shows overall mean and standard deviation of each dataset, given 20 random seeds (Table 2).

Dataset	Mean (Exact Match)	Standard Deviation (Exact Match)
SCAN-simple	0.9819	0.0112
SCAN-length	0.1258	0.0086
SCAN-add turn left	0.8962	0.0833
CFQ-mcd1	0.2574	0.0151
COGS-test	0.9716	0.0050
COGS-gen-lexical	0.3945	0.0460
COGS-gen-structural	0.0001	0.0002
SMCalFlow	0.6281	0.0027
PCFGset	0.9188	0.0097

Table 2: Mean and Standard Deviation of Dataset

4 Conclusion and Future Work

Our analysis found that there were some highly correlated tasks, suggesting that those pairs of tasks recruit similar problem solving capacities. Our experiment could be valuable for further research on compositional generalization, as in depth analysis on highly correlated tasks would be able to reveal some components of compositional generalization capacity of the model. In the cases of negative or null correlations, the performance of one task does not give an indication of performance of another task, and it means that a simple finetuning approach would not help in finding a solution to both tasks. Therefore, our study shows that this correlation analysis will help identify groups of tasks that may require distinct solutions.

However, the limitation here is that we might need more data to conclude that this analysis is statistically significant, as even 20 random seeds may not be enough to get robust estimates of the correlation coefficients, making our study underpowered. Also, we need to note that the exact match score may not capture the model's capacity fully, as this metric only reflects the percentage of examples in which the output sequence exactly matches the gold output. For example, a model that gets only one token wrong but everything else in the sequence correct would still get zero accuracy although the loss for this model could further go down over the course of training.

This research can be further developed by applying the same experiments to other relevant datasets and models and analyzing the results. For a highly correlated pair of tasks, it can be further examined to see whether this trend is robustly found in other model families, in which case it is likely that these tasks require distinct problem solving capacities despite being framed under compositional generalization.

5 Collaboration Statement

All team members contributed roughly equally to the project. Yoobin Cheong identified a set of compositional generalization datasets and models to test, and organized the different datasets into a single unified format as a part of the experimental setup. Yoon Tae Park built a codebase for the training and evaluation pipeline. Yeong Koh performed correlation analysis with the evaluation results and documented any interesting findings using appropriate visualization method.

Acknowledgments

This research was supervised and mentored by Najoung Kim, a faculty fellow at CDS, as a part of her ongoing research project.

References

Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H. Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, and Alexander

Zotov. 2020. Task-oriented dialogue as dataflow synthesis. *Transactions of the Association for Computational Linguistics*, 8:556–571.

Daniel Furrer, Marc van Zee, Nathan Scales, and Nathanael Schärli. 2020. Compositional generalization in semantic parsing: Pre-training vs. specialized architectures.

Dieuwke Hupkes, Verna Dankers, Mathijs Mul, and Elia Bruni. 2019. Compositionality decomposed: how do neural networks generalise?

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2019. Measuring compositional generalization: A comprehensive method on realistic data.

Najoung Kim and Tal Linzen. 2020. Cogs: A compositional generalization challenge based on semantic interpretation.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2873–2882. PMLR.

Yuanpeng Li, Liang Zhao, Jianyu Wang, and Joel Hestness. 2019. Compositional generalization for primitive substitutions.

Linqing Liu, Patrick Lewis, Sebastian Riedel, and Pontus Stenetorp. 2021. Challenges in generalization in open domain question answering.

Richard Montague. 1975. Formal philosophy. *Canadian Journal of Philosophy*, 4(3):573–578.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Pia Weißenhorn, Yuekun Yao, Lucia Donatelli, and Alexander Koller. 2022. Compositional generalization requires compositional parsers.

A Tables & Figures

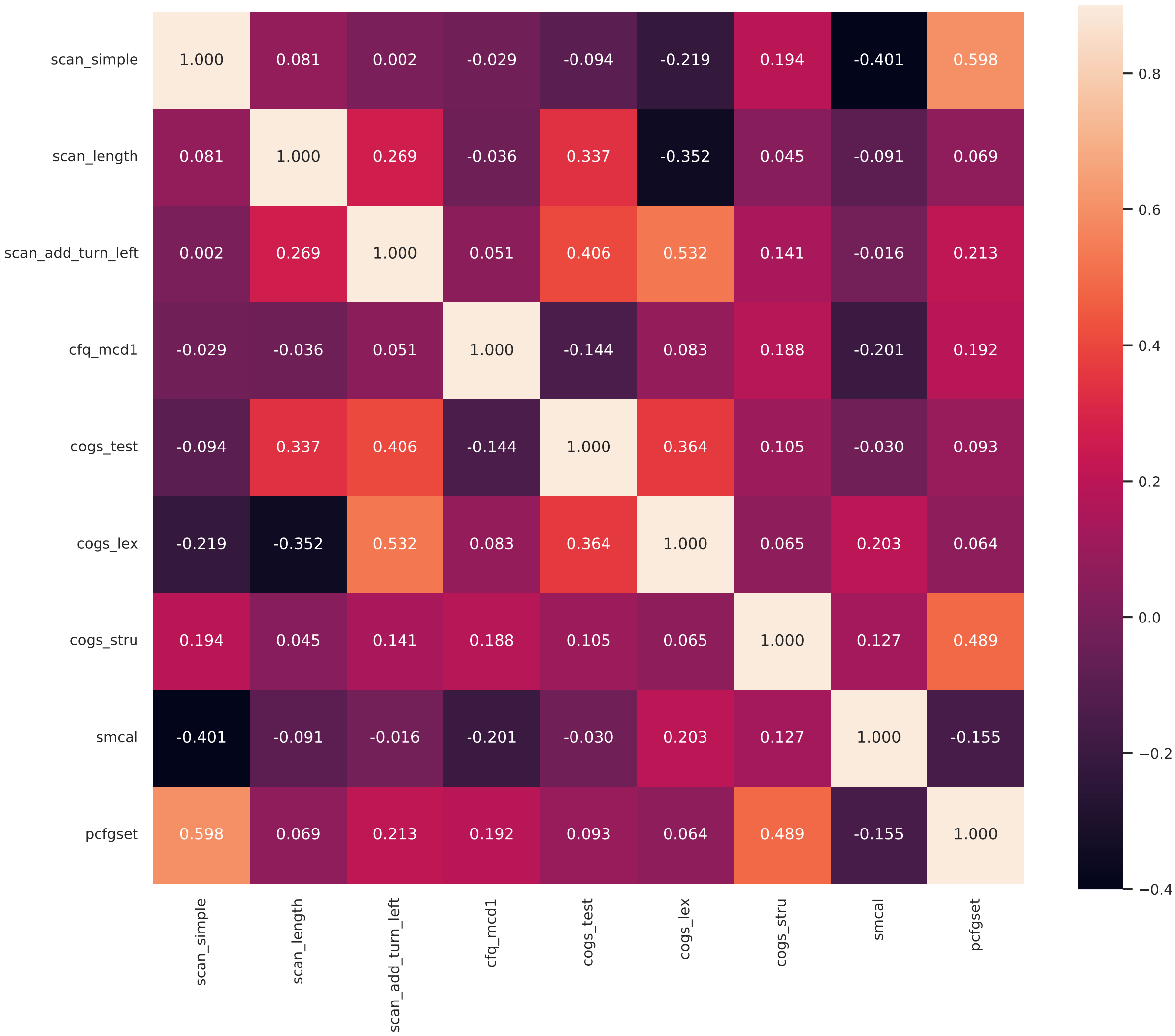

Figure 3: Correlation between each datasets

Dataset	Parameters				Evaluation		
	Batch size	Epoch	Learning rate	Random_state	Train_loss	Eval_loss	Exact_match
SCAN-simple (t5-small)	128	200	5e-05	0	0.007334859276	2.87E-05	0.9976087996
				1	0.007168962152	1.48E-05	0.9995217599
				2	0.00745033166	0.0001524734689	0.9863701578
				3	0.007383443061	0.0001326526399	0.9851745576
				4	0.007203197166	4.16E-05	0.9961740794
				5	0.007397583679	0.000206119963	0.9758488761
				6	0.007208527084	0.0003531243128	0.9684361549
				7	0.007478091261	0.0001606552687	0.9837398374
				8	0.007662578965	0.0001460619969	0.9839789574
				9	0.0071786152	0.000166904123	0.9823051172
				10	0.007481207946	5.75E-05	0.9933046389
				11	0.007730158044	3.87E-04	0.9617407939
				12	0.007281659856	0.0003610258282	0.9710664754
				13	0.007317202424	6.58E-05	0.9923481588
				14	0.007546341226	8.62E-05	0.9913916786
				15	0.007398119799	0.0001566017017	0.9844571975
				16	0.007428055377	0.0003860375145	0.9691535151
				17	0.007086516726	0.0004421735357	0.9615016738
				18	0.007623361932	0.0002568508789	0.975131516
				19	0.007242355264	0.0002268649987	0.9780009565
SCAN-length	128	200	5e-05	0	0.01627728453	0.3362577558	0.1306122449
				1	0.01661655124	0.4148291647	0.1306122449
				2	0.01687743472	0.4099161625	0.1306122449
				3	0.01731976546	0.4386719465	0.1265306122
				4	0.01687757496	0.4218914509	0.1306122449
				5	0.01763484394	0.6451332569	0.1298469388
				6	0.01656827554	0.4104941785	0.1130102041
				7	0.01699380749	0.8328958154	0.09566326531
				8	0.01731525771	0.435950011	0.1306122449
				9	0.01634657263	0.70394665	0.1229591837
				10	0.01752527649	0.4093101025	0.1272959184
				11	0.01839856074	0.3833969235	0.1293367347
				12	0.01684564049	0.3909729719	0.1298469388
				13	0.01621996049	0.3790671229	0.1306122449
				14	0.01802135299	0.4425477087	0.1295918367
				15	0.01628392334	0.3716238737	0.1135204082
				16	0.0170185344	0.3878252208	0.1298469388
				17	0.01617544814	0.7303277254	0.1295918367
				18	0.01731261221	0.440002948	0.1306122449
				19	0.01668305402	0.3349840939	0.1239795918
SCAN-add-turn-left	128	200	5e-05	0	0.005631486087	0.0008338540792	0.9528145695
				1	0.00554699323	0.009626055136	0.7723509934
				2	0.005672240059	0.003712043399	0.9014900662
				3	0.005781763382	0.001139399246	0.946192053
				4	0.005552975937	0.0006659567007	0.9710264901
				5	0.005907344594	0.008401786909	0.7135761589
				6	0.005466845818	0.0005905316793	0.9635761589
				7	0.005750493922	0.009326213039	0.7425496689
				8	0.005932550937	0.005914927926	0.8145695364
				9	0.005573085849	0.00107624114	0.9552980132
				10	0.005761650562	0.0002954888332	0.9859271523
				11	0.005962120522	0.003605121048	0.8468543046
				12	0.005518810169	0.0005021891557	0.9726821192
				13	0.005617676925	0.003065675963	0.8708609272
				14	0.005773737664	0.002543240087	0.9163907285
				15	0.005705563335	0.006835408509	0.8476821192
				16	0.005719707635	0.0004298555723	0.9735099338
				17	0.00544044316	0.006603810005	0.8402317881
				18	0.005825395844	0.0001134381964	0.9925496689
				19	0.005523590214	0.0008582745213	0.9437086093

Table 3: SCAN-simple, length, addprim turn left

Dataset	Parameters				Evaluation		
	Batch size	Epoch	Learning rate	Random_state	Train_loss	Eval_loss	Exact_match
CFQ-mcd1	128	100	5e-05	0	0.004951808645	0.1878000498	0.2628676471
				1	0.004980482707	0.1880038083	0.2565173797
				2	0.004901150268	0.1812835932	0.2733121658
				3	0.004975926309	0.166784361	0.2722259358
				4	0.005021963932	0.1743126959	0.2550969251
				5	0.00496463786	0.1911886632	0.2263536096
				6	0.00504349443	0.1662412286	0.2602774064
				7	0.004857401847	0.1746738702	0.2591076203
				8	0.005031509503	0.1809579432	0.2693850267
				9	0.005036753918	0.1776161343	0.2288602941
				10	0.004866998356	0.1873328835	0.2408923797
				11	0.004950200409	0.1841841936	0.2556818182
				12	0.004837122428	0.1778272092	0.2760695187
				13	0.004927948455	0.1835150868	0.2469084225
				14	0.004936061721	0.1856077909	0.2757352941
				15	0.005026184425	0.1784915775	0.2677139037
				16	0.005020736903	0.1863505989	0.2629512032
				17	0.004937560061	0.1815197319	0.2705548128
				18	0.005094192096	0.1766964942	0.2307820856
				19	0.004976910875	0.1746816039	0.2569351604
COGS-test	128	100	5e-5	0	0.02152565842	0.001290984452	0.9753333333
				1	0.0220092939	0.001323060016	0.975
				2	0.001475576079	0.02204668308	0.969
				3	0.02136987701	0.0008196550771	0.9813333333
				4	0.02153392957	0.001518002711	0.9693333333
				5	0.02102481274	0.001736165374	0.9666666667
				6	0.02214540267	0.001705732546	0.9653333333
				7	0.02205265573	0.00186376716	0.965
				8	0.02293341639	0.001544175437	0.9646666667
				9	0.001099817222	0.02169316459	0.977
				10	0.02183453109	0.001265278668	0.9776666667
				11	0.021336415	0.00110143458	0.9753333333
				12	0.02161602858	0.001374388579	0.972
				13	0.02108788021	0.001556925476	0.9673333333
				14	0.02237227758	0.001625619712	0.9643333333
				15	0.02235104603	0.001475795638	0.969
				16	0.02175871141	0.001353516476	0.9763333333
				17	0.02069682311	0.001043781405	0.976
				18	0.02182525016	0.001097325468	0.9753333333
				19	0.02292406146	0.001269277418	0.9706666667
COGS-lexical	128	100	5e-5	0	0.02152565842	0.001290984452	0.4114444444
				1	0.0220092939	0.001323060016	0.2999444444
				2	0.02204668308	0.001475576079	0.3978888889
				3	0.02136987701	0.0008196550771	0.4331111111
				4	0.02153392957	0.001518002711	0.3987222222
				5	0.02102481274	0.001736165374	0.3118888889
				6	0.02214540267	0.001705732546	0.4035
				7	0.02205265573	0.00186376716	0.4317222222
				8	0.02293341639	0.001544175437	0.3253888889
				9	0.02169316459	0.001099817222	0.4381666667
				10	0.02183453109	0.001265278668	0.432
				11	0.021336415	0.00110143458	0.463
				12	0.02161602858	0.001374388579	0.4119444444
				13	0.02108788021	0.001556925476	0.3403333333
				14	0.02237227758	0.001625619712	0.3746111111
				15	0.02235104603	0.001475795638	0.4326666667
				16	0.02175871141	0.001353516476	0.4440555556
				17	0.02069682311	0.001043781405	0.3415
				18	0.02182525016	0.001097325468	0.3936666667
				19	0.02292406146	0.001269277418	0.4046666667

Table 4: CFQ, COGS-test, COGS-lexical

Dataset	Parameters				Evaluation		
	Batch size	Epoch	Learning rate	Random_state	Train_loss	Eval_loss	Exact_match
COGS-structural	128	100	5e-5	0	0.02152565842	0.001290984452	0.0006666666667
				1	0.0220092939	0.001323060016	0.0
				2	0.0220092939	0.001323060016	0.0
				3	0.02136987701	0.0008196550771	0.0
				4	0.02153392957	0.001518002711	0.0003333333333
				5	0.02102481274	0.001736165374	0.0
				6	0.02214540267	0.001705732546	0.0
				7	0.02205265573	0.00186376716	0.0
				8	0.02293341639	0.001544175437	0.0
				9	0.02169316459	0.001099817222	0.0
				10	0.02183453109	0.001265278668	0.0
				11	0.021336415	0.00110143458	0.0
				12	0.02161602858	0.001374388579	0.0
				13	0.02108788021	0.001556925476	0.0
				14	0.02237227758	0.001625619712	0.0
				15	0.02235104603	0.001475795638	0.0003333333333
				16	0.02175871141	0.001353516476	0.0
				17	0.02069682311	0.001043781405	0.0003333333333
				18	0.02182525016	0.001097325468	0.0
				19	0.02292406146	0.001269277418	0.0003333333333
smcalflow	128	100	5e-5	0	0.01077946567	0.01097799093	0.6302623551
				1	0.01093530331	0.01168796793	0.6198223849
				2	0.01089000497	0.01108726487	0.6308724832
				3	0.01085175046	0.01151832659	0.6255847061
				4	0.01069616254	0.01140160207	0.6259914582
				5	0.01099479732	0.0107978005	0.6307368992
				6	0.01085265057	0.0110064866	0.6270761304
				7	0.01070500909	0.01134838164	0.6306013152
				8	0.01100033832	0.01115417667	0.6291776829
				9	0.01079977547	0.0108904466	0.6287031388
				10	0.01073021164	0.01101701334	0.6295166429
				11	0.01085639492	0.0112897912	0.6287709308
				12	0.01075443088	0.01129097864	0.625177954
				13	0.01090981366	0.01134538278	0.6272117145
				14	0.01078900167	0.01134158671	0.6262626263
				15	0.01073956388	0.01130808704	0.6266693783
				16	0.01075372163	0.01114666741	0.6301945631
				17	0.01084914325	0.01066517085	0.6319571554
				18	0.01083294807	0.01107729413	0.6301945631
				19	0.01074614954	0.01119221188	0.6273472985
pcfgset	128	100	5e-5	0	0.06196381574	0.002386444714	0.9357062031
				1	0.07333237646	0.002891392913	0.9177039399
				2	0.07397078275	0.002985373139	0.9086513733
				3	0.06745049371	0.002669894602	0.9220244831
				4	0.0692433616	0.002966131316	0.9249048452
				5	0.07775206461	0.003159118816	0.9024791688
				6	0.07561014314	0.003680542344	0.8989815863
				7	0.06902646326	0.002739246236	0.9185269005
				8	0.0663543432	0.002354255412	0.9300483489
				9	0.0691221766	0.002808668884	0.924596235
				10	0.06893443989	0.002638096688	0.9227445736
				11	0.07146570918	0.003128418699	0.9034049995
				12	0.0711666968	0.002470331267	0.9176010698
				13	0.06896213721	0.002661463805	0.9263450262
				14	0.06786934178	0.002591838827	0.9225388335
				15	0.06740149009	0.002695811214	0.9268593766
				16	0.06498633327	0.002793221036	0.9209957823
				17	0.07074090619	0.002925436012	0.9129719165
				18	0.07092044324	0.00337511301	0.9099886843
				19	0.0697736337	0.00259421696	0.9296368686

Table 5: COGS-structural, smcalflow, pcfgset