On the Transferability of Causal Knowledge for Language Models

Gourab Dey
Stony Brook University
gdey@cs.stonybrook.edu

Yash Kumar Lal
Stony Brook University
ylal@cs.stonybrook.edu

Abstract

Language understanding includes identifying causal connections between events in a discourse, such as news and instructional text. We study the transferability of causal knowledge across these two domains by analyzing the extent to which understanding preconditions in narratives such as news articles can help models reason about plans such as cooking recipes, and vice-versa. Our experiments show that using instructions to pretrain small models on one domain before similarly finetuning it on the other shows a slight improvement over just finetuning it. We also find that finetuning the models on a mix of both types of data is better (∼3-7% absolute) for understanding causal relations in instructional text. While we find that the improvements do not translate to larger or already instruction tuned models, our analysis highlights the aspects of a plan that are better captured through the interoperability of causal knowledge.

1 Introduction

Understanding underlying causal relationships is an important component of understanding narratives such as news articles. These causal relationships often show up as implicit preconditions and effects of the described events or actions. Preconditions provide a form of logical connection between events that explains why they occur together. They include background information and provide the glue to reason about chains of events common in narratives.

Preconditions also form the base for reasoning about other forms of text. Instructional texts such as how-to procedures often contain prerequisites and details about world states. Recognizing the causal elements in a story aids in identifying prerequisites in instructional text, while grasping procedural preconditions can enhance one's ability to track news events. Humans use a shared framework to comprehend preconditions and other causal relations regardless of the type of text they are reading. In this paper, we aim to study whether understanding aspects of causal knowledge about narratives can help models better understand instructional text and vice versa.

We use PeKo (Kwon et al., 2020), a dataset of annotated preconditions between event pairs in news articles, and CAT-Bench (Lal et al., 2024a), a benchmark testing step order prediction in cooking recipes. First, we establish the performance of different T5 (Raffel et al., 2020) and FlanT5 (Wei et al., 2021) models by finetuning them on each dataset individually. Next, we study how much understanding causal relations within one domain helps understand those in the other. This is done through causal pretraining, i.e., pretraining models on the first domain, finetuning on the second as well as evaluating on it. Finally, we study whether models are able to capture different types of causal knowledge when trained on a data mix from both domains.

Our experiments show that learning various types of causal relations impacts models differently. Base models benefit from training over such knowledge in different domains while larger models already contain it through their pretraining. Our analysis finds that causal pretraining and multi-task finetuning help understand long range relations in plans and cases where two steps in the plan are not dependendent on each other, and highlights areas to better use different types of causal knowledge together.

2 Related Work

There is a vast body of research on extracting different types of relations between events including temporal (Pustejovsky et al., 2003), causal (Girju, 2003), paraphrasal (Lin and Pantel, 2001), and precondition relationships (Kwon et al., 2020, 2021).

Proceedings of the The 7th Workshop on Narrative Understanding, pages 8–14
May 4, 2025 ©2025 Association for Computational Linguistics

ATOMIC (Sap et al., 2019) is a crowd-sourced dataset of event-event relations, where given a simple target event (verb phrase and its arguments), crowd workers provided various types of commonsense knowledge. The Rich Event Description (RED) dataset (O'Gorman et al., 2016) was created to model a broad set of event-event relations in news. CaTeRS (Mostafazadeh et al., 2016) contains data similar to preconditions captured through just one causal relation but focuses on 5 sentence short stories and only contains ∼400 data points. EventStoryLine (Caselli and Vossen, 2017) is also small in size and further does not explicitly capture preconditions. The Precondition Knowledge (PeKo) dataset (Kwon et al., 2020) contains large-scale crowdsourced annotations about precondition relations between event pairs in news stories.

Understanding instructional text involves multiple aspects such as tracking entity states (Bosselut et al., 2018; Henaff et al., 2017), linking actions (Pareti et al., 2014; Lin et al., 2020; Donatelli et al., 2021), next event prediction (Nguyen et al., 2017; Zellers et al., 2019; Zhang et al., 2020a) and more. Zhang et al. (2020b) formalize several multiple-choice tasks related to step- and goal- relations in procedures. Kiddon et al. (2015) explore predicting dependencies in cooking recipes and related tasks. Similar work has been done on identifying dependencies in multimodal instructions with images and text (Pan et al., 2020; Wu et al., 2024). CaT-Bench (Lal et al., 2024b) clearly studies the prediction and explanation of temporal ordering constraints on the steps of an instructional plan.

Humans have the ability to utilize knowledge from previous experiences when learning a new task. Prior work has explored techniques of transfer learning and domain adaptation to learn skills in various contexts. Zoph et al. (2016); Kocmi and Bojar (2018) explored using parallel data from high resource languages to improve translation in low resource languages. Gururangan et al. (2020); Han and Eisenstein (2019) use domain adaptation techniques for models to learn new tasks. Similar to these, we investigate whether understanding causal knowledge in one domain helps with another.

3 Data

To study the transferability of causal knowledge, we use CaT-Bench and PeKo, which contain information about dependencies between a plan's steps and preconditions about events respectively.

CaT-Bench (Lal et al., 2024b) is a dataset of causal dependency questions defined on cooking recipes to evaluate the causal and temporal reasoning abilities of models over instructional plans. Specifically, it focuses on the ability to recognize temporal dependencies between steps i.e., deciding if one step must happen before or after another. For a recipe in the dataset, containing an ordered number of steps, the dataset contains either of two binary questions: (1) Must $step_i$ happen before $step_j$? and (2) Must $step_j$ happen after $step_i$? We pool questions from dependent pairs of steps into Dep, and the rest into NonDep[1].

PeKo is a dataset consisting of crowdsourced annotations of preconditions between event pairs in news articles. Kwon et al. (2020) first subsample events and their temporal relations using CAEVO (Chambers et al., 2014) from the New York Times Annotated Corpus (Sandhaus, 2008). The resultant set was then filtered to retain only pairs of events that have a "before" or "after" temporal relation between them. These were further sampled and given to annotators who evaluated whether or not the candidate precondition event was an actual precondition for the target event resulting in 30k annotations.

4 Experiment Details

We provide critical information about the models and training regimes we use for our experiments.

4.1 Models

We conduct our experiments with the base and large models of the T5 and FlanT5 model family.

T5 reframes all text-based language problems into a text-to-text format. It is based on the encoder-decoder transformer architecture and is fine-tuned across a wide range of tasks by converting inputs and outputs into text strings. This unified approach allows T5 to effectively transfer learned knowledge from one task to another, achieving then state-of-the-art results across a wide range of benchmarks.

FlanT5 involves fine-tuning a T5 model with a diverse set of task-specific instructions before applying it to downstream tasks. Different from previous standard pretraining and finetuning methods, this approach enhances the model's ability to generalize across different tasks by explicitly teaching it to follow instructions during the finetuning

[1] Note that the answers to all the questions in the Dep set are 'yes', and the answers to NonDep questions are 'no'.

	T5-B	T5-L	FLAN-B	FLAN-L
PEKO / PEKO (FT)	0.76	0.80	0.78	0.80
CAT-BENCH → PEKO / PEKO (CP)	0.78	0.80	0.78	0.80
BOTH / PEKO (MFT)	0.79	0.80	0.79	0.81
CAT-BENCH / CAT-BENCH (FT)	0.8	0.92	0.91	0.95
PEKO → CAT-BENCH / CAT-BENCH (CP)	0.82	0.89	0.91	0.92
BOTH / CAT-BENCH (MFT)	0.87	0.91	0.90	0.93

Table 1: Macro F1 of different T5 and FLANT5 trained models on PEKO and CAT-BENCH. The dataset listed in red denote the data the model was trained on, and the dataset listed in green denotes the benchmark on which the F1 score is calculated. B represents base sized models and L represents the large sized models. → denotes that the model has been sequentially trained on the dataset before → first and then on the dataset listed after it.

phase.

4.2 Experiments

We first manually craft an instruction for the task corresponding to each dataset and prepend[2] it to all data points. We then follow three distinct training regimes as described below.

Finetuning (FT) In this regime, we finetune a model on the corresponding dataset to establish its performance on the task.

Causal Pretraining (CP) We first pretrain a model on one dataset before finetuning on the other. To do so, for instance, we first pretrain a model on PEKO and then finetune on CAT-BENCH to study whether learning preconditions about real world events in news helps better understand aspects of causal knowledge within plans. Theoretically, the model learns to detect causal dependencies from the first stage before adapting to the target dataset. This is aimed to test the transferability of causal knowledge between narratives such as news and instruction following content such as recipes.

Multi-Task Finetuning (MTF) We combine the corresponding splits of both instruction prepended datasets, shuffle them and finetune a model on it. This setting studies whether a model can learn different aspects of causal knowledge when given data from varying domains.

5 Results

Table 1 shows the performance of different models trained using the various training regimes described above. First, we find that all the FT models for PEKO achieve better performance than the best finetuned models reported in Kwon et al. (2020). Particularly, comparing models of the same size, T5 and FLANT5 are better on this binary classification task than the reported BERT model even though they are generative models. FT models for CAT-BENCH show improved performance over any of the reported baselines, which is expected as the baselines are only zero- and few-shot settings.

We observe mixed results when using the causal pretraining regime (CP). It is clear that first learning about preconditions from news events helps T5-BASE understand cause and effect relations implicitly encoded within the steps of a plan. We hypothesize that larger models already encode such knowledge in their parameters and such pretraining does not affect downstream performance. These findings also hold when first learning about plans followed by news events. Clearly, transferring causal knowledge between generic news events and highly specific actions in a plan lead to improved reasoning across both.

Multi-task training (MFT) over both datasets together improves T5-BASE performance over finetuning (FT) regardless of the target task. In fact, there is a large improvement (∼7%) on CAT-BENCH in this regime, and a small improvement in understanding news events in PEKO. However, while the opposite is true for T5-LARGE, the drop is negligible. This training paradigm does not impact FLANT5 performance on PEKO but mixing causal information in news with that in plans leads to slight decrease in understanding the latter.

Overall, we find that the training regime heavily impacts a model's performance on a causal understanding task. Simply following one regime will not lead to improvements across all tasks, and it is

[2] We also experiment with no prefix and an alternate prefix.

important to explore the different options.

6 Analysis

Having established the differences in training regimes across different settings, we investigate the abilities T5-BASE on CAT-BENCH to better understand our results.

6.1 Reasoning as a function of Step Distance

We study how the distance between the steps in question impacts model performance across training regimes. A question is said to be about *close* steps $(step_i, step_j)$ if $(j - i) < 3$, and *distant* otherwise. For CP and MFT, we calculate the number of cases where the corresponding regime corrects an error found in FT.

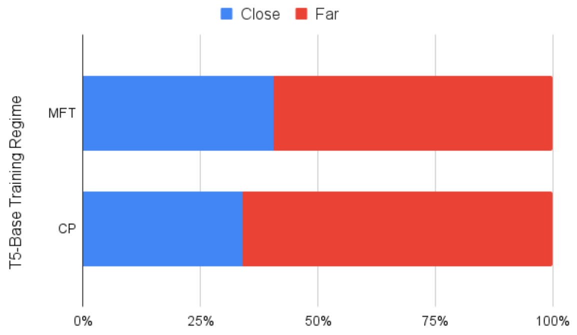

Figure 1: Distribution of improvements produced due to different T5-BASE training regimes for CAT-BENCH as a function of distance between the steps being asked about.

Figure 1 shows the distribution of these corrections as divided by the distance between the pair of steps in question. We hypothesize that models are likely to predict a dependencies between steps that are distant from each other, since it is likely that steps towards the end depend on ones near the start. We find that both CP and MFT improve reasoning more for distant steps rather than closer ones, indicating that the extra data helps understand indirect connections, or lack thereof, between steps.

6.2 Reasoning over Directional Dependencies

We study how models handle questions about different aspects of the same pair of steps. Typically, questions about why a step must happen *before* another require reasoning about preconditions and causes, while answering why a step must happen *after* another requires understanding the effects of any performed actions.

Table 2 shows the performance of T5-BASE on questions testing the 'before' and 'after' order between steps. We find that causal pretraining (CP) helps the model for questions about both dependent and non-dependent pairs of steps. In fact, CP helps the most on the non-dependent subset which is harder to detect.

		Before	After
DEP	FT	0.82	0.82
	CP	0.84	0.83
NONDEP	FT	0.77	0.76
	CP	0.80	0.79

Table 2: Performance (macro F1) of T5-BASE on CAT-BENCH when just finetuned (FT) on the target dataset as compared to using causal pretraining (CP) split by the type of dependence relations between the plan steps.

We also use the dependency related annotations in CAT-BENCH to understand the types of improvements the different training setups brings over finetuning. To do so, we extract the cases where FT fails but CP or MFT fix that error.

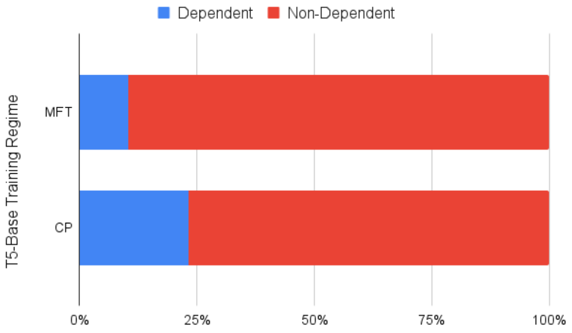

Figure 2: Distribution of improvements produced due to different T5-BASE training regimes for CAT-BENCH as a function of whether there is a dependency between within the pair of steps being asked about.

Figure 2 shows that the overwhelming majority of improvements are found for step pairs without a dependency. Detecting that two steps do not depend on each other is harder than the inverse since it involves eliminating all possibilities of there being a dependency between the steps.

7 Conclusion

With the ubiquity of causal relations, we study the transferability of such knowledge between critical, real-world domains. We investigate how learning about preconditions in news events impacts models' abilities to reason about causes and effects in plans and vice versa. Comparing different training setups reveals that, while different domains require varying finetuning strategies, transferring causal knowledge is helpful for smaller models. Larger models often already encode such information. Our error analysis reveals aspects of a plan that such regimes help with, highlighting areas of improvement for future research.

Limitations

We limit our investigation to two encoder-decoder pretrained models which are much smaller (in terms of number of parameters) than decoder-only large language models such as GPT-3 and others. Nonetheless, these small models are pretrained on large swathes of text and capture a model causal knowledge related to the world in their parameters. While we study such models as an artifact possibly reflecting a view of the world, we acknowledge that they don't capture all aspects of it. Even with our findings, they must be deployed only after extensive testing to study how they impact people. Finally, our work only investigates English-language documents and this limits the generalizability of our findings to other languages.

References

Antoine Bosselut, Omer Levy, Ari Holtzman, Corin Ennis, Dieter Fox, and Yejin Choi. 2018. Simulating action dynamics with neural process networks. *ICLR*.

Tommaso Caselli and Piek Vossen. 2017. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86, Vancouver, Canada. Association for Computational Linguistics.

Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics*, 2:273–284.

Lucia Donatelli, Theresa Schmidt, Debanjali Biswas, Arne Köhn, Fangzhou Zhai, and Alexander Koller. 2021. Aligning actions across recipe graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6930–6942, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Roxana Girju. 2003. Automatic detection of causal relations for question answering. In *Proceedings of the ACL 2003 Workshop on Multilingual Summarization and Question Answering*, pages 76–83, Sapporo, Japan. Association for Computational Linguistics.

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

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2017. Tracking the world state with recurrent entity networks. *ICLR*.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, Lisbon, Portugal. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 244–252, Brussels, Belgium. Association for Computational Linguistics.

Heeyoung Kwon, Nathanael Chambers, and Niranjan Balasubramanian. 2021. Toward diverse precondition generation. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 160–172, Online. Association for Computational Linguistics.

Heeyoung Kwon, Mahnaz Koupaee, Pratyush Singh, Gargi Sawhney, Anmol Shukla, Keerthi Kumar Kallur, Nathanael Chambers, and Niranjan Balasubramanian. 2020. Modeling preconditions in text with a crowd-sourced dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3818–3828, Online. Association for Computational Linguistics.

Yash Kumar Lal, Vanya Cohen, Nathanael Chambers, Niranjan Balasubramanian, and Ray Mooney. 2024a. CaT-bench: Benchmarking language model understanding of causal and temporal dependencies in plans. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19336–19354, Miami, Florida, USA. Association for Computational Linguistics.

Yash Kumar Lal, Vanya Cohen, Nathanael Chambers, Niranjan Balasubramanian, and Raymond Mooney. 2024b. Cat-bench: Benchmarking language model understanding of causal and temporal dependencies in plans. *Preprint*, arXiv:2406.15823.

Angela Lin, Sudha Rao, Asli Celikyilmaz, Elnaz Nouri, Chris Brockett, Debadeepta Dey, and Bill Dolan. 2020. A recipe for creating multimodal aligned datasets for sequential tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4871–4884, Online. Association for Computational Linguistics.

Dekang Lin and Patrick Pantel. 2001. Dirt @sbt@discovery of inference rules from text. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '01, page 323–328, New York, NY, USA. Association for Computing Machinery.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. CaTeRS: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61, San Diego, California. Association for Computational Linguistics.

Dai Quoc Nguyen, Dat Quoc Nguyen, Cuong Xuan Chu, Stefan Thater, and Manfred Pinkal. 2017. Sequence to sequence learning for event prediction. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 37–42, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Tim O'Gorman, Kristin Wright-Bettner, and Martha Palmer. 2016. Richer event description: Integrating event coreference with temporal, causal and bridging annotation. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 47–56, Austin, Texas. Association for Computational Linguistics.

Liang-Ming Pan, Jingjing Chen, Jianlong Wu, Shaoteng Liu, Chong-Wah Ngo, Min-Yen Kan, Yugang Jiang, and Tat-Seng Chua. 2020. Multi-modal cooking workflow construction for food recipes. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1132–1141, New York, NY, USA. Association for Computing Machinery.

Paolo Pareti, Benoit Testu, Ryutaro Ichise, Ewan Klein, and Adam Barker. 2014. Integrating know-how into the linked data cloud. In *International Conference on Knowledge Engineering and Knowledge Management*, pages 385–396. Springer.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

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.

Evan Sandhaus. 2008. The New York Times Annotated Corpus. Abacus Data Network.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3027–3035.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Te-Lin Wu, Alex Spangher, Pegah Alipoormolabashi, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2024. Understanding multimodal procedural knowledge by sequencing multimodal instructional manuals. *Preprint*, arXiv:2110.08486.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Hongming Zhang, Muhao Chen, Haoyu Wang, Yangqiu Song, and Dan Roth. 2020a. Analogous process structure induction for sub-event sequence prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1541–1550, Online. Association for Computational Linguistics.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020b. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

A Experiment Details

A.1 Hyperparameters

Here, we describe the hyperparameters we use to train our models. For both T5-BASE and FLANT5-BASE, we use a learning rate of 3e-4 for FT and MTF. For transfer during the CP stage, we use a lower learning rate of 1e-4, specifically we find that using a higher learning rate leads to a degradation in performance here for the FLANT5-BASE models. For T5-LARGE and FLANT5-LARGE, we use a learning rate of 5e-5 for CAT-BENCH and 1e-4 for PEKO during FT, and a learning rate of 5e-5 for MTF. For the transfer stage, we use a learning rate of 1e-4, and surprisingly find that a lower learning rate here leads to poor performance in contrast to the base models. All models were trained with a batch size of 64 and for a maximum of 7 epochs with early stopping.

A.2 Dataset Details

	Train	Validation	Test
CAT-BENCH	13,868	1,616	2,840
PEKO	23,158	2,895	2,895

Table 3: Number of examples in different splits of each dataset

Table 3 presents statistics of the datasets - PEKO and CAT-BENCH - used for our experiments.