AFRL-RI-RS-TR-2023-201

FLASH: FAST LEARNING VIA AUXILIARY SIGNALS, STRUCTURED KNOWLEDGE, AND HUMAN EXPERTISE

UNIVERSITY OF PENNSYLVANIA

NOVEMBER 2023

FINAL TECHNICAL REPORT

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09. This report is available to the general public, including foreign nations. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TR-2023-201 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE CHIEF ENGINEER:

/ S /
PETER A. JEDRYSIK
Work Unit Manager

/ S /
MICHELLE R. GRIECO
signed for: JULIE BRICHACEK
Chief, Information Systems Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
NOVEMBER 2023	FINAL TECHNICAL REPORT	SEPTEMBER 2019	MAY 2023

4. TITLE AND SUBTITLE

FLASH: FAST LEARNING VIA AUXILIARY SIGNALS, STRUCTURED KNOWLEDGE, AND HUMAN EXPERTISE

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
FA8750-19-2-0201	N/A	61101E
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
		R2VV

6. AUTHOR(S)

Dan Roth

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

PRIME

University of Pennsylvania
3401 Walnut St Room 461C
Philadelphia PA 19104

SUB

Georgia Institute of Technology
801 Atlantic Drive
Atlanta GA 30332

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RISB
525 Brooks Road
Rome NY 13441-4505

DARPA/I20
675 N. Randolph St.
Arlington VA 22203-2114

10. SPONSOR/MONITOR'S ACRONYM(S)

AFRL/RI & DARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

AFRL-RI-RS-TR-2023-201

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The original goal of the FLASH project was to develop novel and efficient machine learning algorithms that leverage rich forms of structured knowledge. Specifically, we build on the hypothesis that appropriate use of structured knowledge can substantially reduce the amount of hand-labeled data needed to achieve state-of-the-art performance on standard machine learning tasks, and address two key challenges:

• Leveraging structure: Develop general algorithms for leveraging structure to learn new concepts from few or no hand-labeled examples.

• Inferring structure: Develop general algorithms for inferring structure, either by actively learning from scratch or by transferring it from other domains

The FLASH program executed this plan and produced a number of theoretical and practical contributions in all the areas mentioned above. In addition to developing theory, algorithms and representations we used those to develop applications in natural language and in computer vision.Moreover, in the course of the DARPA LwLL project, the focus of the field changed as a result of the success of large pre-trained generative AI models, including large language models (LLMs) like ChatGPT. While the goal of the work has not changed, our own research agenda adapted to these changes in the field while remaining close to the broad goals of our original proposal.

15. SUBJECT TERMS

Machine Learning, Structured prediction, Program synthesis, Knowledge integration, Vector embeddings, Constrained deep learning, Compositionality, Natural Language Processing, Computer Vision, Language Models.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
U	U	U	SAR	38

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
PETER A. JEDRYSIK	N/A

PREVIOUS EDITION IS OBSOLETE.

TABLE OF CONTENTS

1.0 SUMMARY

The original goal of the FLASH project was to develop novel and efficient machine learning algorithms that leverage rich forms of structured knowledge. Specifically, we build on the hypothesis that appropriate use of structured knowledge can substantially reduce the amount of hand-labeled data needed to achieve state-of-the-art performance on standard machine learning tasks, and address two key challenges:

- Leveraging structure: Develop general algorithms for leveraging structure to learn new concepts from few or no hand-labeled examples.
- Inferring structure: Develop general algorithms for inferring structure, either by actively learning from scratch or by transferring it from other domains.

The original formulation of our project focused on (1) Program Synthesis/Structured Prediction: the design of new structured prediction and program synthesis algorithms and their use to infer and leverage structure, (2) The study of neural representations (vector embeddings) and their use to develop new algorithms for embedding structure into vector representations, and use it to support transfer and leverage structure, and (3) inferring and using domain knowledge and auxiliary signals as a way to induce structure, better address transfer learning and develop new constrained deep learning algorithms to incorporate auxiliary signals.

The FLASH program executed this plan and produced a number of theoretical and practical contributions in all the areas mentioned above. In addition to developing theory, algorithms and representations we used those to develop applications in natural language and in computer vision.

Moreover, in the course of the DARPA LwLL project, the focus of the field changed as a result of the success of large pre-trained generative AI models, including large language models (LLMs) such as ChatGPT. While the goal of the work has not changed, our own research agenda adapted to these changes in the field while remaining close to the broad goals of our original proposal. For instance, much of our proposed work on vector embeddings broadened to studying LLMs, and our work on program synthesis incorporated neuromyotonic elements, as described in detail below.

The FLASH program produced a large number of publications in top conferences and the description below will not address all these contributions. Instead, we will focus on some of the key contributions in each technical area and refer the reader to the rich set of references below for complete information.

2.0 INTRODUCTION

2.1 General

Current state-of-the-art machine learning algorithms typically require millions of labeled examples, making them prohibitively expensive to apply in domains for which labels are expensive to acquire, e.g., for machine translation to languages with few speakers or for structured prediction tasks such as semantic image segmentation. The objective of the project is to build innovative frameworks for inferring and leveraging rich structure to enable learning with a few labels. The original formulation of our project focused on the following topics:

2.1.1 Program Synthesis/Structured Prediction

The goal in this part was to design new structured prediction and program synthesis algorithms and use them to infer and leverage structure and it consisted of these key directions:
(i) design novel structured prediction and synthesis algorithms by integrating search into deep reinforcement learning, (ii) adapt meta-reinforcement-learning to this setting, (iii) develop composition and decomposition approaches that account for learned and injected knowledge, and (iv) perform active learning for structured prediction.

The goal of the **structured prediction** work was to significantly improve the state-of-the-art in structured prediction, a key component to our proposed approaches for inferring structure. In the **program synthesis** component, we planned to make significant improvements to existing state-of-the-art program synthesis algorithms. In particular, we planned to (i) use machine learning, including deep reinforcement learning, to guide the synthesis algorithm, which we call learning-to-synthesize, (ii) incorporate neural components into synthesized programs, and (iii) use active learning to further reduce the amount of data needed to synthesize new concepts. This part also included a **knowledge integration** part where we planned to develop new algorithms for integrating structure across different sources. In particular, we planned to (i) develop new inference algorithms that combine probabilistic and logical reasoning, so we can integrate structure in a way that both accounts for uncertainty and respects logical constraints on the structure, and (ii) develop new algorithms for actively discovering new relationships to grow knowledge bases over time.

2.1.2 Vector Embeddings

The goal of this part was to develop new algorithms for embedding structure into vector representations, which can be used to support transfer and leverage structure. The plan here consisted of these key directions: (i) learn vector embeddings that are shared across different structures and domains, (ii) develop techniques for quantifying uncertainty in representations, and (iii) use these algorithms to transfer structure. Specifically, we planned to make significant

improvements to state-of-the-art unsupervised deep learning algorithms for learning vector embeddings, including devising general-purpose techniques for embedding structured representations into vector spaces, new meta-learning algorithms, and improving the quality of shared embeddings that can be used to transfer structure across domains and modalities while keeping track of uncertainty in the transfer.

2.1.3 Auxiliary Signals

The goal in this part was to develop algorithms for inferring and leveraging auxiliary signals. The plan here consisted of these directions: (i) using program synthesis and "semantic parsing" to infer auxiliary signals, (ii) using shared embeddings to transfer auxiliary signals to new domains, and (iii) developing new constrained deep learning algorithms to incorporate auxiliary signals. Specifically, we planned to develop new algorithms for constrained deep learning, which we would use to integrate auxiliary signals into state-of-the-art deep learning models. This includes developing new constrained optimization algorithms that enforce hard constraints on the model, and developing new algorithms that automatically encode soft constraints into the loss function to guide learning.

Overall, in addition to developing theory, algorithms, representations, and applications on top of these, both in natural language and in computer vision (CV), we planned to show performance improvements in a range of important machine learning tasks and to substantially reduce the amount of hand-labeled data needed to achieve state-of-the-art performance on standard machine learning tasks. Specifically, we planned to develop (i) new learning and inference technologies for structured prediction, program synthesis, and knowledge integration, (ii) new unsupervised deep learning algorithms for learning vector embeddings, (iii) new algorithms for constrained deep learning focused on integrating auxiliary signals, with applications to a range of Natural Language Processing (NLP) tasks, (iv) an object detection system that learns from few labeled examples, (v) a video activity recognition system that learns from few labeled examples, and (vi) a machine translation system that learns from few aligned sentences. The last three systems would learn from < 100 labeled examples or adapt to new domains from < 10 labeled examples. Altogether, this would dramatically reduce the amount of labeled data required to learn in NLP and CV, thus our work will make it substantially more cost-effective to leverage state-of-the-art machine learning algorithms in these settings.

Over the course of the DARPA LwLL project, the focus of the field changed as a result of the success of large pre-trained generative AI models, including large language models (LLMs) like ChatGPT. Our own research agenda adapted to these changes in the field while remaining close to the broad goals of our original proposal. For instance, much of our proposed work on vector embeddings broadened to studying LLMs, and our work on program synthesis incorporated neuromyotonic elements. In the rest of this Introduction we describe some of the key tasks and accomplishments of the FLASH team.

2.2 Task: Neurosymbolic Program Synthesis, Programming, and Learning Theory

Neurosymbolic programs, which are programmatic compositions of neural network components, have the promise to significantly improve data efficiency by leveraging structure in the data. In the course of this project we have developed a novel algorithm for learning neurosymbolic programs from a handful of examples, a neurosymbolic language and toolchain to combine the complementary benefits of deep learning and logical reasoning, and a learning theory account for some of the fundamental questions in the Neurosymbolic domain, facilitating better understanding of what is possible and how, and opening new venues for additional algorithmic improvements.

We have designed and implemented a novel framework for synthesizing neurosymbolic programs from a handful of examples; our experimental results demonstrate that these approaches can significantly outperform existing approaches at tasks such as classifying object trajectories in videos in the few-shot setting. Our algorithm additionally uses active learning to gather additional examples from the user to significantly improve performance.

In addition, we have developed a novel neurosymbolic learning algorithm for few-shot learning of novel concepts in semantic parsing. Given a single example of a novel concept, such as "thrice", our algorithm isolates the portion that is novel, and uses program synthesis to automatically infer how to generalize this concept into a subroutine. Then, this subroutine can be used in future occurrences of that concept.

In the Neurosymbolic Programming space we developed Scallop, a neurosymbolic language and toolchain to combine the complementary benefits of deep learning and logical reasoning. Scallop advanced the principles and practice of neurosymbolic AI by 1) developing new theories and algorithms for scalable differentiable reasoning, 2) implementing a toolchain for supporting the development and training of a wide range of challenging AI applications, and 3) performing extensive empirical studies demonstrating the feasibility and competitiveness of Scallop's solutions relative to both neural and neurosymbolic baselines.

We then developed a new system called Vieira which extends our neurosymbolic programming system Scallop with foundation models. The powerful yet incomplete nature of these models has spurred a wide range of mechanisms to augment them with capabilities such as in-context learning, information retrieval, and code interpreting. Vieira builds atop Scallop to unify these mechanisms in providing a general solution for programming with foundation models.
Like Scallop, Vieira follows a probabilistic relational paradigm and treats foundation models as stateless functions with relational inputs and outputs. It supports neurosymbolic applications by enabling the seamless combination of such models with logic programs, as well as complex, multi-modal applications by streamlining the composition of diverse sub-models. We implemented Vieira with 12 built-in foundation models and evaluated it on a suite of 9 challenging tasks that span language, vision, and structured and vector databases. Our evaluation

demonstrates that programs in Vieira are concise, can incorporate state-of-the-art foundation models such as GPT-4 and CLIP, and have comparable or better accuracy than competitive baselines.

On the Learning Theory side, we developed a framework to study and understand key scalability issues in neurosymbolic learning. In [61] we study the problem of multi-instance partial label learning (PLL), a problem that captures neurosymbolic learning and latent structural learning in NLP. This setting has several advantages over end-to-end neural architectures. One obvious advantage is the ability to extract and reuse the latent model. Another advantage is improved end-task accuracy. For instance, learning via multi-instance PLL can lead to architectures with higher accuracy than that of end-to-end architectures in NLP and visual QA (Question Answering) tasks. Moreover, several recently proposed neurosymbolic works are based on multi-instance PLL so this fundamental work will find more applications in the future.

2.3 Task: Structured Prediction and Domain Knowledge: Training, Parameter Efficiency, Auxiliary Signals, and Learning Theory

Most interesting and realistic machine learning applications in NLP and CV are structured prediction problems where the model needs to assign values to multiple interdependent variables. We have done significant theoretical and application work on the use of domain knowledge to improve structured prediction, both in the space of logic programs (that are structured too) and in the general case of structured prediction problems in NLP. Some of the key contributions in this space include:

- Investigated probabilistic reasoning frameworks to leverage *logic programs* and *knowledge graphs* within deductive databases to systematically train neural models for end-to-end tasks.
    - Due to the logical component, fewer parameters are required for machine reasoning and fewer training iterations required for convergence.
    - Tested against dense vision-text models for visual question answering (VQA), where our method outperforms baseline methods, even in low data regimes.
- Leveraged *logical programs* as *intermediate text embeddings* within Transformers to track the state of a program and execute the corresponding program for end tasks.
    - Generalized to VQA to improve performance over dense cross-attention image-text models.
        - Since logic programs are embedded using text, our methods perform well when programs are out of distribution or are noisy.
    - Generalized to a policy learning tasks to dynamically select actions given the world state to improve performance over baseline program-guided agents.
        - We show that our method is more sample efficient as well.
- We developed new ways to incorporate “free” supervision signals in the context of answering compositional questions requiring symbolic reasoning against natural language

paragraphs. Compositional, structured models are appealing because they explicitly decompose problems and provide interpretable intermediate outputs that give confidence that the model is not simply latching onto data artifacts. Learning these models is challenging, however, because end-task supervision only provides a weak indirect signal on what values the latent decisions should take. This often results in the model failing to learn to perform the intermediate tasks correctly. We introduced a way to leverage paired examples that provide stronger cues for learning latent decisions.

- We introduced Target-Aware Weighted Training (TAWT), a weighted training algorithm for cross-task learning based on minimizing a representation-based task distance between the source and target tasks. We showed that TAWT is easy to implement, is computationally efficient, requires little hyperparameter tuning, and enjoys non-asymptotic learning-theoretic guarantees. As a byproduct, the proposed representation-based task distance allows one to reason in a theoretically principled way about several critical aspects of cross-task learning, such as the choice of the source data and the impact of fine-tuning.
- We developed the foundations for learnability with incidental, indirect supervision signals by studying conditions for learnability from indirect supervision signals. Real-world applications often require improved models by leveraging a range of cheap incidental supervision signals. These could include partial labels, noisy labels, knowledge-based constraints, and cross-domain or cross-task annotations -- all having statistical associations with gold annotations but not exactly the same.
- Prior knowledge and symbolic rules in machine learning are often expressed in the form of label constraints, especially in structured prediction problems. We studied two common strategies for encoding label constraints in a machine learning pipeline, regularization with constraints and constrained inference, by quantifying their impact on model performance.
- Real-world applications often require improved models by leveraging a range of cheap incidental supervision signals. These could include partial labels, noisy labels, knowledge-based constraints, and cross-domain or cross-task annotations -- all having statistical associations with gold annotations but not exactly the same. However, we currently lack a principled way to measure the benefits of these signals to a given target task, and the common practice of evaluating these benefits is through exhaustive experiments with various models and hyperparameters. We developed a theory that addresses the question of whether we can, in a single framework, quantify the benefits of various types of incidental signals for a given target task without going through combinatorial experiments.

2.4 Task: Vector Embeddings: Semi-Supervised Object Detection

- Made a number of key contributions to *efficiently* training computer vision models, where efficiency includes both label efficiency and parameter efficiency.
- Significantly increased our understanding of semi-supervised learning, and developed Unbiased Teacher for Closed and Open-Set Object Detection, an influential line of work introducing successful semi-supervised object detection methods across a range of settings including various types of detectors and closed/open-set.
- Developed methods for parameter-efficient and robust finetuning of foundation Vision Transformer models.

Our team has made a number of key contributions to *efficiently* training computer vision models, where efficiency includes both label efficiency and parameter efficiency. In terms of label efficiency, we have developed some of the early and more popular semi- and self-supervised methods for object detection that have been strong-cited and built-upon, included code releases with significant impact and use, and resulted in strong evaluation results (including by other teams that leveraged our method). Specifically, we developed Unbiased Teacher for Closed and Open-Set Object Detection, an influential line of work introducing successful semi-supervised object detection methods. Our methods proposed a number of key principles, including addressing severe imbalance problems (e.g. background bounding boxes dominate in the data), teacher-student based pseudo-labeling where predictions from a teacher network are intelligently used to train the student network, as well as a novel relative uncertainty-based mechanism to filter out noisy predictions from the teacher network. We subsequently expanded this to the open-set case, where the unlabeled data can contain different categories than the labeled data, allowing us to achieve better performance even than fully labeled datasets that are noisy, such as Google's Open Images. This work formed the foundation of a large number of subsequent works, including real-world applications to domains such as healthcare as well as winning submissions to the evaluations (both by our team as well as other teams that used our open-source code). Our code has over 450 github stars, and has been integrated into the official Detectron2 codebase.

Finally, we have also developed methods for parameter-efficient and robust finetuning of foundation Vision Transformer models. Specifically, we have looked at parameter-efficient multi-task learning for dense vision tasks, proposing a method called PolyHistor. We achieved state-of-art performance using only 10% of the trainable parameters, and proposed a general method that can be leveraged for other tasks. We also showed that, since updating a smaller number of parameters significantly prevents overfitting, our method was highly applicable to low-labeled datasets as well.

Overall, our team has significantly pushed the state-of-art in low-labeled learning, leading to methods that both advance our scientific understanding as well as work under practical situations, including during the DARPA evaluations.

2.5 Task: Understanding Pre-Trained Models: Large Language Models for NLP

We made several key contributions to the understanding of large language models and how they can be used for NLP applications.

- We showed that a simple methodological improvement of deduplicating LLM training data results in improved model quality.
- We showed that LLMs could be used to create high quality synthetic data in conjunction with human editing, and that the resulting training data set could be used to mitigate undesirable biases.
- We explored different architectures for LLMs, contrasting fill-in-the-blank models versus continuation models, and showed that bidirectional LLMs have the same few-shot learning properties as unidirectional models like GPT-3.
- We applied LLMs to a variety of tasks including few-shot learning for text style transfer, fine-tuning for machine translation of low resource languages, and commonsense reasoning about wikihow-style tasks that involve goals, steps, and temporal ordering.
- We explored chain-of-thought prompting methods and introduced the notion of 'faithfulness' in explanation via the generation of executable code.
- We developed new way to decompose prompts so that we can identify the "propositions" required to support reasoning tasks.

The goal of the LwLL program was to dramatically reduce the amount of training data that is required to adapt a machine learning system to a new domain. Over the course of the program, it became clear that the best paradigm for doing this was via pre-training large models on general domain data and then fine-tuning them on a limited amount of in-domain data. Compared with training systems from scratch on in-domain data, this results in dramatically improved task performance quality across a huge range of tasks. On the NLP side, large language models (LLMs) became the dominant approach across the field during the LwLL program. Our team contributed significantly to the body of knowledge on LLMs and their applications in NLP.

Deduplicating Training Data. A first major contribution was demonstrating that methodological refinement—specifically deduplication of LLM training data—can measurably improve the resultant model's quality, and reduce undesirable model behavior like memorization.

Creating Synthetic Training Data. A further contribution arose from illustrating that LLMs could be used, in tandem with human editing, to generate high-quality synthetic data. Crucially, we found that the curated training data set could be employed to diminish undesirable model biases—for instance when training a system to write biographies using Wikipedia data, the model learns out-of-date biases like that most famous mathematicians are men.

LLM Architectures. In the course of our work, we also explored different architectures for LLMs. Through contrasting 'fill-in-the-blank' models with 'continuation' models, we showed that the less commonplace bidirectional LLMs shared the same few-shot learning capabilities as well-established unidirectional models, like GPT-3. This expanded the potential applicability and versatility of these models for diverse NLP tasks.

LLMs for NLP Tasks. We applied LLMs for a variety of NLP tasks—ranging from few-shot learning for text style transfer to fine-tuning for machine translation of under-resourced languages. This application was also extended to commonsense reasoning assignments involving wikihow-style tasks, which necessitated comprehension of goals, steps, and their temporal ordering.

Faithful Chain of Thought Models. Most recently, we have explored chain-of-thought prompting methods, and critiqued their 'faithfulness'. We showed that although generating chain-of-thought explanations tends to improve models' performance for a wide range of tasks, the explanations themselves are not guaranteed to be faithful to the model's predictions. Faithfulness is an important, foundational part of explainable models. We explored introducing faithful explanations through the generation of executable code. Instead of talking through a math word problem, like in standard chain-of-thought models, our faithful chain-of-thought model generates executable python code, which ensures that its output solution is faithful to the explanation. This methodology ensures that AI solutions not only perform tasks accurately but also account for the underlying logic and coherence of their solutions—a key advancement in AI transparency and interpretability.

Learning to Decompose via Intermediate Pre-Training: Explicit decomposition modeling, which involves breaking down complex tasks into simpler, and often more interpretable, sub-tasks has long been a central theme in developing robust and interpretable natural language understanding (NLU) systems. Despite the many datasets and resources that have been built as part of this effort, the majority of them have small-scale annotations and limited scope, which is not enough to solve general decomposition tasks. Pre-training of LLM suffers from similar problems due to the simplistic nature of the self-supervised training. We proposed large-scale intermediate pre-training of decomposition-based transformers using distant supervision from comparable texts as a way to support better decomposition and use it as the basis for building an end-to-end decomposition-based question answering system that is shown to produce better results on challenging reasoning datasets.

3.0 NEUROSYMBOLIC PROGRAM SYNTHESIS, PROGRAMMING, AND LEARNING THEORY

3.1 Methods, Assumptions, and Procedures

The key challenge for learning neurosymbolic programs is the need to optimize over both a discrete space of program structures as well as continuous parameters of programs. Our algorithm uses A* search to search over the space of possible programs to perform this search. As the heuristic for A* search, it uses logical reasoning to compute optimistic bounds on the objective value of the program, which is a valid search heuristic. By using this strategy, our algorithm guarantees optimality of the synthesized program.

In addition, our algorithm uses active learning to automatically identify promising examples to query the user to improve performance. It uses a standard greedy learning strategy to do so. At each step of active learning, it computes all neurosymbolic programs consistent with the examples so far. Then, for each possible user query, it computes the expected fraction of the search space that will be pruned by this query. Finally, it queries the user on the example with the highest pruning power, and uses the response to continue search.

We have implemented our algorithm in the context of data science queries over videos. In particular, we have designed a language that enables users to perform queries over objects in videos to identify videos with interesting behaviors. For instance, a data scientist might want to identify certain driving patterns from traffic videos, or search for animal herding patterns in wildlife videos.

We enabled general-purpose neurosymbolic programming using Scallop through three key features: 1) a flexible symbolic representation that is based on the relational data model; 2) a declarative logic programming language that is based on Datalog and supports recursion, aggregation, and negation; and 3) a framework for automatic and efficient differentiable reasoning that is based on the theory of provenance semirings.

We implemented these features in a comprehensive and open-source toolchain comprising 65K lines of code in the programming language Rust. It includes a compiler, an interpreter, and PyTorch bindings to integrate Scallop programs with existing machine learning pipelines. We also conducted tutorials on Scallop at numerous venues including SSFT 2022 (11th summer school on formal techniques), LoG 2022 (learning on graphs), and PLDI 2023 (programming language design and implementation).

On the Learning Theory side, we developed a framework to study and understand key scalability issues in neurosymbolic learning. In [61] we study the problem of multi-instance partial label learning (PLL), a problem that captures neurosymbolic learning and latent structural learning in NLP. This setting has several advantages over end-to-end neural architectures. One obvious

advantage is the ability to extract and reuse the latent model. Another advantage is improved end-task accuracy. For instance, learning via multi-instance PLL can lead to architectures with higher accuracy than that of end-to-end architectures in NLP and visual QA tasks. Moreover, several recently proposed neurosymbolic works are based on multi-instance PLL so this fundamental work will find more applications in the future.

3.2 Results and Discussion

We have demonstrated that our algorithm can synthesize highly accurate programs starting from just 2 positive examples; in particular, it significantly outperforms deep learning based approaches including state-of-the-art transformer models. In particular, our approach achieves an F1 score of at least 0.9 on 13 out of 17 benchmarks, whereas the transformer never achieves this performance. Similarly, we have implemented and evaluated our approach for few-shot novel concept learning. In our experiments, our approach achieves near-perfect accuracy at all tasks, whereas a deep learning based baseline only achieves accuracy of a few percentage points.

We evaluated Scallop using a suite of 8 neurosymbolic applications that span the domains of image and video processing, natural language processing, planning, and knowledge graph querying, in a variety of learning settings such as supervised learning, reinforcement learning, rule learning, and contrastive learning. Our evaluation demonstrated that Scallop is expressive and yields solutions of comparable, and often times superior, accuracy than state-of-the-art models. We showed additional benefits of Scallop's solutions in terms of runtime and data efficiency, interpretability, and generalizability.

On the Learning Theory side we considered a weakly supervised learning scenario where the supervision signal is generated by a transition function of labels associated with multiple input instances. We formulate this problem as multi-instance Partial Label Learning (multi-instance PLL), which is an extension to the standard PLL problem. Our problem is met in different fields, including latent structural learning and neurosymbolic integration. Despite the existence of many learning techniques, limited theoretical analysis has been dedicated to this problem. In [61], we provide the first theoretical study of multi-instance PLL with possibly an unknown transition. Our main contributions are as follows: First, we proposed a necessary and sufficient condition for the learnability of the problem. This condition nontrivially generalizes and relaxes the existing small ambiguity degree in PLL literature since we allow the transition to be deterministic. Second, we derived Rademacher-style error bounds based on the top- surrogate loss that is widely used in the neurosymbolic literature. Furthermore, we conclude with empirical experiments for learning with an unknown transition. The empirical results align with our theoretical findings; however, they also expose the issue of scalability in the weak supervision literature.

4.0 STRUCTURED PREDICTION AND DOMAIN KNOWLEDGE: TRAINING, PARAMETER EFFICIENCY, AUXILIARY SIGNALS, AND LEARNING THEORY

4.1 Methods, Assumptions, and Procedures

Most interesting and realistic machine learning applications in NLP and CV are structured prediction problems where the model needs to assign values to multiple interdependent variables. We have done significant theoretical and application work on the use of domain knowledge to improve structured prediction, both in the space of logic programs (that are structured too) and in the general case of structured prediction problems in NLP.

Logic Programs Space:

- Leverage Datalog and define a provenance semiring to assign probabilities for neural outputs and combinations in a symbolic fashion.
    - Use top-k proofs to approximate probabilities, making larger tasks feasible with our system.
    - Incorporate predefined logical rules represented as Datalog programs, as well as knowledge graphs as explicit facts to use for reasoning.
    - Train neural models on reasoning tasks end-to-end efficiently.
- Encode program structure as intermediate vector representations in a Transformer for instruction following.
    - Further encode the program structure dependencies to enable loops and branching in the program structure.
    - Attend dependency embeddings with the targeted modality for reasoning tasks.
- In the Datalog system, we assume logic programs are present and are specified correctly. Furthermore, we do not account for noise provided by our knowledge graph.
- For Transformer based methods, the explicit structure of the program has to be correct to perform well, although it is robust to the syntax of program language.
- Datalog based models improved training efficiency in simpler MNIST addition tasks as well as complex visual question answering and reasoning tasks (VQAR).
- Transformer based models improved performance on program guided tasks such as GQA as well as an agent policy learning task on Minecraft.

4.2 Results and Discussion

Datalog based methods reach close to max performance in MNIST tasks with only a few minutes of wall clock training time, while baselines take much longer to converge. For VQR tasks it outperforms custom neural module network approaches as well as large visual-language cross attention based methods. It also converges faster, even though all baseline methods have access to the same logic programs used.

The Transformer based approach improves over baselines on GQA, including specialized neural module networks. Furthermore we test embedding the program with different encodings and

having structural guidance to ablate our design choices. We are able to use the same architecture to train a reinforcement learning agent in a Minecraft environment, where we obtain better rewards versus other program guided agents across short, long, and complex tasks. Systematically combining logic and neural networks has yielded a range of solutions. Ones that are more logic/symbolic to encode program specifications, and ones that are more neural to encode the variances in high dimensional data. We present two works on either side of this spectrum.

We are able to leverage Datalog, which has already been an established deductive database to represent our logical programs for evaluation. Using deduction is useful for machine learning reasoning tasks, since evidence can be smaller components of raw data, which has to be reasoned about to provide a final answer. For example, asking "How many people are standing to the left of the tree?" decomposes into identifying people and a tree in an image, then analyzing the relation of the people to the tree, and finally counting the number of people to give the final answer. Here the objects detected are the evidence, where the final answer is deduced from using Datalog. This system provides a powerful abstraction of how to combine deep models, which determine probabilistic facts, with reasoning systems that can be built upon to train those models. Transformer based logic methods allow flexibility of the tasks used, since images, texts, and even game boards can be encoded through cross attention embeddings. This makes this approach a powerful model when combined with logic driven tasks. The key challenge is posing tasks that use program guidance and as well as have methods to provide program apriori. This enables using logic and state tracking of a task within a deep network to capture the variance of real world tasks.

Structure Prediction:

- Compositional, structured models are appealing because they explicitly decompose problems and provide interpretable intermediate outputs that give confidence that the model is not simply latching onto data artifacts. Learning these models is challenging, however, because end-task supervision only provides a weak indirect signal on what values the latent decisions should take. This often results in the model failing to learn to perform the intermediate tasks correctly. In this work, we introduce a way to leverage paired examples that provide stronger cues for learning latent decisions. When two related training examples share internal substructure, we add an additional training objective to encourage consistency between their latent decisions. Such an objective does not require external supervision for the values of the latent output, or even the end task, yet provides an additional training signal to that provided by individual training examples themselves. We apply our method to improve compositional question answering using neural module networks on the DROP dataset. We explore three ways to acquire paired questions in DROP: (a) discovering naturally occurring paired examples within the dataset, (b) constructing paired examples using templates, and (c) generating paired

examples using a question generation model. We empirically demonstrate that our proposed approach improves both in- and out-of-distribution generalization and leads to correct latent decision predictions.

- A key feature of human intelligence is the ability to generalize beyond the training distribution, for instance, parsing longer sentences than seen in the past. Currently, deep neural networks struggle to generalize robustly to such shifts in the data distribution. We study robust generalization in the context of using recurrent neural networks (RNNs) to learn regular languages. We hypothesize that standard end-to-end modeling strategies cannot generalize well to systematic distribution shifts and propose a compositional strategy to address this. We compare an end-to-end strategy that maps strings to labels with a compositional strategy that predicts the structure of the deterministic finite-state automaton (DFA) that accepts the regular language. We theoretically prove that the compositional strategy generalizes significantly better than the end-to-end strategy. In our experiments, we implement the compositional strategy via an auxiliary task where the goal is to predict the intermediate states visited by the DFA when parsing a string. Our empirical results support our hypothesis, showing that auxiliary tasks can enable robust generalization. Interestingly, the end-to-end RNN generalizes significantly better than the theoretical lower bound, suggesting that it is able to achieve at least some degree of robust generalization.

- We introduced Target-Aware Weighted Training (TAWT), a weighted training algorithm for cross-task learning based on minimizing a representation-based task distance between the source and target tasks. We showed that TAWT is easy to implement, is computationally efficient, requires little hyperparameter tuning, and enjoys non-asymptotic learning-theoretic guarantees. As a byproduct, the proposed representation-based task distance allows one to reason in a theoretically principled way about several critical aspects of cross-task learning, such as the choice of the source data and the impact of fine-tuning.

- Real-world applications often require improved models by leveraging a range of cheap incidental supervision signals. These could include partial labels, noisy labels, knowledge-based constraints, and cross-domain or cross-task annotations -- all having statistical associations with gold annotations but not exactly the same. However, we currently lack a principled way to measure the benefits of these signals to a given target task, and the common practice of evaluating these benefits is through exhaustive experiments with various models and hyperparameters. In [28], we study whether we can, in a single framework, quantify the benefits of various types of incidental signals for a given target task without going through combinatorial experiments. We propose a unified PAC-Bayesian motivated informativeness measure, PABI, that characterizes the

uncertainty reduction provided by incidental supervision signals. We demonstrate PABI's effectiveness by quantifying the value added by various types of incidental signals to sequence tagging tasks. Experiments on named entity recognition (NER) and question answering (QA) show that PABI's predictions correlate well with learning performance, providing a promising way to determine, ahead of learning, which supervision signals would be beneficial.

- Prior knowledge and symbolic rules in machine learning are often expressed in the form of label constraints, especially in structured prediction problems. We studied two common strategies for encoding label constraints in a machine learning pipeline, regularization with constraints and constrained inference, by quantifying their impact on model performance. For regularization, we show that it narrows the generalization gap by precluding models that are inconsistent with the constraints. However, its preference for small violations introduces a bias toward a suboptimal model. For constrained inference, we show that it reduces the population risk by correcting a model's violation, and hence turns the violation into an advantage. Given these differences, we further explore the use of two approaches together and propose conditions for constrained inference to compensate for the bias introduced by regularization, aiming to improve both the model complexity and optimal risk.

- Learning from indirect supervision signals is important in real-world AI applications when, often, gold labels are missing or too costly. In [60], we develop a unified theoretical framework for multi-class classification when the supervision is provided by a variable that contains nonzero mutual information with the gold label. The nature of this problem is determined by (i) the transition probability from the gold labels to the indirect supervision variables and (ii) the learner's prior knowledge about the transition. Our framework relaxes assumptions made in the literature, and supports learning with unknown, non-invertible and instance-dependent transitions. Our theory introduces a novel concept called separation, which characterizes the learnability and generalization bounds. We also demonstrate the application of our framework via concrete novel results in a variety of learning scenarios such as learning with superset annotations and joint supervision signals.

5.0 VECTOR EMBEDDINGS: SEMI-SUPERVISED OBJECT DETECTION

5.1 Methods, Assumptions, and Procedures

Semi-Supervised Object Detection: Unbiased Teacher. The availability of large-scale datasets and computational resources has allowed deep neural networks to achieve strong performance on a wide variety of tasks. However, training these networks requires a large number of labeled examples that are expensive to annotate and acquire. As an alternative, Semi-Supervised Learning (SSL) methods have received growing attention. Yet, these advances have primarily focused on image classification, rather than object detection where bounding box annotations require more effort.

In [41], we revisit object detection under the SSL setting: an object detector is trained with a single dataset where only a small amount of labeled bounding boxes and a large amount of unlabeled data are provided, or an object detector is jointly trained with a large labeled dataset as well as a large external unlabeled dataset. A straightforward way to address Semi-Supervised Object Detection (SS-OD) is to adapt from existing advanced semi-supervised image classification methods. Unfortunately, object detection has some unique characteristics that interact poorly with such methods. For example, the nature of class-imbalance in object detection tasks impedes the usage of pseudo-labeling. In object detection, there exists foreground-background imbalance and foreground classes imbalance. These imbalances make models trained in SSL settings prone to generate biased predictions. Pseudo-labeling methods, one of the most successful SSL methods in image classification, may thus be biased towards dominant and overly confident classes (background) while ignoring minor and less confident classes (foreground). As a result, adding biased pseudo-labels into the semi-supervised training aggravates the class-imbalance issue and introduces severe overfitting. Taking a two-stage object detector as an example, there exists heavy overfitting on the foreground/background classification in the RPN and multi-class classification in the ROIhead (but not on bounding box regression).

To overcome these issues, we propose a general framework – Unbiased Teacher: an approach that jointly trains a Student and a slowly progressing Teacher in a mutually-beneficial manner, in which the Teacher generates pseudo-labels to train the Student, and the Student gradually updates the Teacher via Exponential Moving Average (EMA), while the Teacher and Student are given different augmented input images. Inside this framework, (i) we utilize the pseudo-labels as explicit supervision for both RPN and ROIhead and thus alleviate the overfitting issues in both RPN and ROIhead. (ii) We also prevent detrimental effects due to noisy pseudo-labels by exploiting the Teacher-Student dual models. (iii) With the use of EMA training and the Focal loss, we can address the pseudo-labeling bias problem caused by class-imbalance and thus improve the quality of pseudo-labels. As the result, our object detector achieves significant performance improvements.

Unbiased Teacher v2. As mentioned above, we have developed state-of-the-art Semi-Supervised Object Detection (SS-OD) methods to leverage only limited labeled data but more abundant unlabeled data to improve performance. However, our method and other state-of-the-art SS-OD methods apply self-training techniques, which generate pseudo-labels and enforce the consistency between unlabeled data with different augmentations. Despite the significant improvement, there are still two remaining issues that are left untackled: (1) there is no prior SS-OD work on anchor-free detectors and (2) prior works are ineffective in pseudo-labeling on the bounding box regression.

First, anchor-free detectors have been recently getting more attention in the community of object detection, with the promise of achieving competitive accuracy, computational efficiency, and potential generalization to new datasets or environments. In spite of these advances, existing SS-OD works mainly focus on anchor-based detectors but do not empirically verify their effectiveness on anchor-free detectors. In fact, when we adapt recent state-of-the-art SS-OD methods to anchor-free detectors, we observe that, compared with its improvement on anchor-based models, the improvement is much smaller on anchor-free models. With extensive analysis, we find that some advanced techniques performing favorably in the fully-supervised setting do not work in the semi-supervised setting with limited supervision. For example, the centerness score becomes unreliable for box selection under the semi-supervised setting, and the localization-based labeling method is not robust to the localization noise in pseudo-labels.

Second, following the Teacher-Student framework, the existing SS-OD works apply an unsupervised regression loss with the pseudo-boxes generated from confidence thresholding (i.e., a threshold on the box score). However, we find that this approach inherits some potential issues that can be further addressed. For instance, (1) instead of using one single metric (e.g., box score or box IoU) to jointly represent the quality of four boundaries, the confidence/uncertainty of each boundary should be predicted individually; (2) confidence in the classification branch might not be able to reflect the quality of boundary prediction on the regression branch. Instead, we propose to predict uncertainties on the regression branch to select pseudo-labels for boundary prediction; (3) Lastly, simply relying on Teacher's confidence/uncertainty prediction to select pseudo labels for regression cannot prevent misleading instances for the regression task. Instead, we propose to exploit the relative uncertainties between the Teacher and Student to select the boundary-level pseudo-labels, in which the Teacher has lower uncertainty than the Student. Integrating the three components, we propose Listen2Student to improve the unsupervised regression loss for the SS-OD tasks.

Open-Set Semi-Supervised Object Detection. One limitation of the above work, and works that have built on it, is that they often consider a scenario where the labeled set is randomly sampled from a dataset and use the remaining images as the unlabeled set. This implicitly assumes the label spaces of labeled and unlabeled data are identical. However, this closed-set assumption is unlikely to happen in real-world situations, where unlabeled images collected in the wild might contain out-of-distribution (OOD) objects, which are unseen, undefined, and unknown in the available labeled set.

We are thus interested in a more practical yet challenging problem, Open-Set Semi-Supervised Object Detection (OSSOD), which aims to leverage the unconstrained unlabeled images (i.e., images containing unseen OOD objects), to improve an object detector trained with the available labeled data. When adding unlabeled data containing open-set categories, we observe that the existing successful SSOD method leads to a lower performance gain or even degraded results. This is different from the common belief that SSL methods can benefit from using more unlabeled data. We attribute the above phenomena to the semantic expansion issue, where OOD objects are mispredicted as in-distribution objects with high confidence and misused as pseudo-labels with confidence thresholding.

To eliminate the detrimental effect of OOD samples, we propose to add an additional OOD filtering process into the existing SSOD training pipeline. More concretely, we first consider online OOD detectors to perform OOD filtering. An online OOD detector is a prediction head/branch we straightforwardly add on the object detector using existing OOD methods. However, we find that such methods cannot produce satisfactory results due to interference with other tasks, e.g., bounding box localization and box classification. In order to address this, we propose a simple but effective strategy that uses an offline OOD detection module, which is disentangled from the architecture of the object detector. This OOD detector is based on a self-supervised DINO [2] model, and it provides several advantages. First, the pre-training of DINO does not require label annotations, so it is suitable for the low-label setting and alleviates the concern of limited amounts of labels for OOD detection tasks. Secondly, it is more effective in detecting OOD objects in the pseudo-labels compared with other (online) OOD methods, and this eliminates the detrimental effect of OOD samples for OSSOD tasks. Lastly, since the architecture of OOD detector and object detector are independent, the training of the two models can be done separately and prevent interference as we observed in online OOD detectors.

5.2 Results and Discussion

Our papers [41, 40, 39] detail our results, but a summary of our results is as follows:

- **Semi-Supervised Object Detection:** We benchmark Unbiased Teacher with SSL setting using the MS-COCO and PASCAL VOC datasets, namely COCO-standard, COCO-additional, and VOC. When using only 1% labeled data from MS-COCO (COCO-standard), Unbiased Teacher achieves 6.8 absolute mAP improvement against the state-of-the-art method, STAC. Unbiased Teacher consistently achieves around 10 absolute mAP improvements when using only 0.5, 1, 2, 5% of labeled data compared to supervised baseline. Unbiased Teacher also demonstrates 3.9 mAP improvement when trained with additional unlabeled data from MS-COCO (COCO-additional). Our code has been released publicly.
- **Unbiased Teacher v2:** Our method achieves significant improvements compared to the state-of-the-art SSOD methods when using both anchor-free and anchor-based detectors on several SS-OD benchmarks, including COCO-standard, COCO-additional, and VOC. We also provide ablation studies to examine the effectiveness of our Listen2Student.
- **Open-Set Semi-Supervised Detection:** In our experiments, we first provide a systematic analysis between several OOD detection methods, and our results suggest that offline methods show consistent improvements over the online detection methods. We also show that using the offline OOD detector can filter the OOD objects in pseudo-labels and consistently improve against the existing SSOD methods under different open-set scenarios, including under different combination of unlabeled sets, varying number of in-distribution (ID) classes, and different number of images. We also find that using the pseudo-labels generated from our framework is even more effective than using the ground-truth labels provided in OpenImages.

Overall, our Unbiased Teacher line of work has been significantly influential, receiving significant uptake and citations (340+ citations since 2021), and drove a large number of follow-on works. Our methods proposed a number of key principles, including addressing severe imbalance problems (e.g. background bounding boxes dominate in the data), teacher-student based pseudo-labeling where predictions from a teacher network are intelligently used to train the student network, as well as a novel relative uncertainty-based mechanism to filter out noisy predictions from the teacher network. We subsequently expanded this to the open-set case, where the unlabeled data can contain different categories than the labeled data, allowing us to achieve better performance even than fully labeled datasets that are noisy, such as Google's Open Images. This work formed the foundation of a large number of subsequent works, including real-world applications to domains such as healthcare as well as winning submissions to the DARPA LwLL evaluations (both by our team as well as other teams that used our open-source code, which has over 450 GitHub stars).

Moving forward, the field has significantly changed with the popularity of large-scale self-supervised models (DINO, MAE, etc.) as well as multi-modal vision-and-language models (e.g. CLIP). Our unbiased teacher method has inspired a number of follow-ons that utilize these models, including the large number of open-vocabulary detectors. The combination of self-supervised pre-trained models, combined with multi-modal models, provides a promising path to models that are extremely generalizable under limited labeled data for the downstream task.

6.0 UNDERSTANDING PRE-TRAINED MODELS: LARGE LANGUAGE MODELS FOR NLP

6.1 Methods, Assumptions, and Procedures

Deduplicating Training Data

In [34], we find that existing language modeling datasets contain many near-duplicate examples and long repetitive substrings. As a result, over 1% of the unprompted output of language models trained on these datasets is copied verbatim from the training data. We develop two tools that allow us to deduplicate training datasets. Deduplication allows us to train models that emit memorized text ten times less frequently and require fewer training steps to achieve the same or better accuracy. We can also reduce train-test overlap, which affects over 4% of the validation set of standard datasets, thus allowing for more accurate evaluation. Code for deduplication is released at https://github.com/google-research/ deduplicate-text-datasets.

Methods:

- Exact substring matching using suffix arrays to identify and remove duplicated sequences of 50+ tokens
- Approximate matching between documents using MinHash to estimate overlap and remove near-duplicate documents

Assumptions:

- Duplication in training data is problematic and causes issues like overfitting, memorization, and train-test contamination
- Removing duplication will mitigate these issues and improve model quality

Procedures:

- Analyzed duplication in four common NLP datasets: C4, Wiki-40B, LM1B, RealNews
- Found 1-14% of examples are near-duplicates, and over 1% of model outputs are memorized
- Removed duplicates from datasets using the methods above
- Trained Transformer language models on original and deduplicated versions of C4
- Evaluated on perplexity and likelihood of emitting memorized text

Creating Synthetic Training Data

In [68], we explored using large language models to generate synthetic training data, which is then refined by human editors. This allows for efficiently creating high-quality labeled datasets. This indirectly addresses the goals of LwLL. Rather than focusing on decreasing the amount of labeled training required for machine learning systems, we instead focus on creating training datasets inexpensively by generating synthetic training using LLMs.

For this particular study, we created a synthetic replacement for the WikiBio dataset. The original WikiBio dataset consists of 728,321 infoboxes describing notable individuals, mapped to biographies written in natural English about those individuals. The dataset contains many biases such as the fact that mathematicians are more often men or that most biographies on English Wikipedia are of Americans and Europeans. Machine learning methods trained on biased data may perpetuate these biases in undesirable ways. We used synthetic data generation as a way of controlling inherent biases in the training dataset.

Methods:

- We used two 137B parameter transformer language models developed by Google. One was a completion-based LLM and one was a chat-based LLM.
- The chat-based LLM was used to generate attribute lists for fictional individuals, where the frequency of certain features like gender, nationality, etc were controlled in order to remove biases from the original dataset.
- Human editors revise attribute lists for factual plausibility, appropriateness and formatting.
- LLM generates biography text conditioned on edited attributes.
- Human editors revise biographies for faithfulness, fluency and formatting
- The process results in the synthetic WikiBio benchmark dataset SynthBio

Assumptions:

- Our work assumes that balancing demographic properties like gender or nationality is desirable and appropriate for removing biases correlated with other features like profession.
- Our method assumes synthetic data generation plus human refinement can efficiently produce high-quality training data.

Procedures:

- Dataset creation procedure involves iterative collaboration between models and humans
- Humans perform quality control and editing after each round of model generation

LLM Architectures

We highlight two papers related to LLM architectures: [31] and [47]. In [31], we argue that the task of inserting text into a specified position in a passage, known as fill in the blank (FITB), is more general than the more typical continuation-based LMs. We investigated the feasibility of using a single model to do both tasks, and showed that models pre-trained with a FITB-style objective are capable of both tasks, while models pre-trained for continuation are not. FITB models are bidirectional, meaning that they can use both left and right contexts, unlike unidirectional LMs which only take the left contexts into account when continuing. Unidirectional LLMs like GPT-3 are known to be able to perform in-context few-shot learning with only a few labeled examples. This emergent prompt-based learning capabilities had only been demonstrated for unidirectional language models. However, bidirectional language models like T5 have many desirable properties. We introduced SAP (Sequential Autoregressive Prompting), a technique that enables the few-shot prompting of bidirectional models.

Methods:

- Our approach uses pre-trained and finetuned T5 models for fill-in-the-blank and continuation
- We compare against baseline LLM model finetuned with prefix tuning
- We proposed a new a sequential autoregressive prompting (SAP) technique for bidirectional models like mT5

Assumptions:

- FITB models are more general than continuation models since they can support both insertion and continuation interactions
- Bidirectional models can perform few-shot learning through prompting without fine-tuning

Procedures:

- After training the different model architecture, we assess models on automatic fluency metrics and human judgments, and measured their performance on tasks like machine translation (using 14 languages)

Faithful Chain of Thought Models

In [43] we enhanced the chain-of-thought (CoT) prompting technique. CoT prompting allows LLMs to perform better reasoning on a wide range of tasks including things like math word problems. CoT brought striking performance gains, by prompting an LM to generate a reasoning chain along with the answer, given only a few in-context examples. In addition to performance improvement, CoT is also claimed to "provide an interpretable window into the behavior of the model" However, it lacks one fundamental property of explanation, faithfulness, meaning "an explanation (i.e., the reasoning chain) should accurately represent the reasoning process behind the model's prediction" (i.e., how the model arrives at the final answer). We introduced a new

technique called Faithful CoT. Faithful CoT is a faithful-by-construction framework where the answer is the result of deterministically executing the reasoning chain. Specifically, we break down a complex reasoning task into two sequential stages: Translation and Problem Solving. During Translation, an LM translates a Natural Language (NL) query into a reasoning chain, which interleaves NL and a task-dependent Symbolic Language (SL), such as Python, Datalog, or Planning Domain Definition Language (PDDL). In the Problem Solving stage, the reasoning chain is executed by a deterministic solver, e.g., a Python/Datalog interpreter, or a PDDL planner, to derive the answer.

Methods:

- We propose a 2-stage pipeline with Translation and Problem Solving stages
- In Translation, an LM converts a natural language query into a reasoning chain interleaving NL and symbolic language (e.g. Python, Datalog)
- In Problem Solving, an external deterministic solver executes the symbolic program to derive the answer

Assumptions:

- Decomposing reasoning into symbolic execution ensures the faithfulness of the reasoning chain explanation
- Interleaving natural language and symbolic language improves interpretability

Procedures:

- Evaluated on math word problems, multi-hop QA, planning tasks, and logical inference
- Compared performance against standard prompting and chain of thought prompting
- Analyzed the contribution of different prompt components through ablation studies

6.2 Results and Discussion

Deduplicating Training Data

Key results:

- Deduplication reduces train-test overlap and verbatim memorization by ~10x
- No loss in perplexity, sometimes gains up to 10% on validation sets
- Models reach higher accuracy with fewer training steps

Discussion:
Deduplicating training data provides substantial benefits for large language models. Removing near-duplicate examples and repetitive substrings reduces overfitting, memorization, and train-test contamination. [34] has been cited over 150 times, and the methodology has been adopted in LLM training by most groups.

In the experiments in the paper, deduplication reduced verbatim memorization by 10x and maintained or improved perplexity. This allows models to reach higher accuracy with fewer training steps.

The techniques like suffix array matching and MinHash offer scalable ways to deduplicate even very large (terabyte-scale) text datasets. Deduplication also improves training efficiency, as the processed datasets can be significantly smaller.

Limitations:
The analysis focused on intrinsic metrics like perplexity and memorization. Further analysis is needed on how deduplication impacts performance on downstream NLP tasks. While verbatim memorization is reduced, there may still be issues with "semantic memorization" where models generate content semantically similar to training examples. More advanced techniques could help address this.

Deduplication focuses on identical or near-identical text. However, training data can be "unfair" in less obvious ways related to representation, biases, and harmful content. Deduplication does not address these issues.

Deduplicating training data is a valuable technique for improving large language models. It directly addresses duplicate content issues and provides measurable gains. However, it is just one part of a broader effort to improve training data and model quality. Researchers should consider complementary techniques like data filtering and augmentation when building language models.

Creating Synthetic Training Data

Key Results:

- We demonstrated that LLMs could be used to create high quality synthetic data in conjunction with human editing by creating the SynthBio dataset, a synthetic replacement for the popular WikiBio dataset.
- SynthBio has less noise, more balance, and greater faithfulness than original WikiBio ("faithfulness" refers to the biographies accurately reflecting the information in the infoboxes, without embellishment or unsupported facts).
- Models trained on WikiBio perform worse on SynthBio, requiring more grounded text.

Discussion:
Creating high-quality synthetic training data using large language models in tandem with human editing indirectly meets the goals of LwLL by driving down the cost and reducing the difficulty of obtaining high quality training data. The iterative process of having models generate initial data which is then refined by human editors provides a collaborative curation workflow that combines the generative capabilities of LLMs with human judgment. This approach was shown to produce the synthetic WikiBio benchmark SynthBio, which has greater faithfulness, balance, and less noise compared to the original real-world WikiBio dataset.

However, there remain significant challenges in controlling subtle biases and ensuring full representativeness when generating synthetic datasets programmatically. For example, SynthBio

still exhibited some biases like inconsistencies in gender identity and an overrepresentation of Western attributes. More research is needed to enable finer-grained control over the data generation and curation process. Nevertheless, this human-AI collaborative approach shows promise for efficiently creating training data and could likely benefit other NLP tasks that require labeled datasets, like machine translation, summarization, and question answering.

Overall, the key contribution of this work is demonstrating how large language models can be effectively leveraged in conjunction with human editing and refinement to bootstrap the synthetic data generation process. This allows creating training sets that are high-quality, customizable, and efficient to produce. As language models continue to advance in their generative capabilities, tapping into their strengths while still maintaining human oversight over dataset creation will be an important direction for efficiently producing the data needed to train next-generation AI systems.

LLM Architectures

Key Results:

- Models pre-trained on fill-in-the-blank-style objectives are better for both the fill-in-the-blank task and the continuation task
- Bidirectional models using SAP can match or exceed unidirectional models on machine translation

Discussion:
Our work explored different LLM architectures, and showed that less common architectures and training methods –bidirectional models with fill-in-the-blank objectives like span corruption– sometimes surpass the more common GPT-style unidirectional LLM architecture. Both [31] and [48] contribute to the goals of the LwLL program in exploring how to maximize model capabilities with less training data. The fill-in-the-blank work shows how model architectures and pre-training can reduce the need for fine-tuning data. The bidirectional prompting work demonstrates how synthetic data generation and prompting can minimize the amount of human-labeled examples needed. The techniques could likely benefit other NLP tasks like summarization and question answering. The papers demonstrate promising directions for efficient and adaptable models that require less training data through careful model architecture and pre-training scheme design.

Faithful Chain of Thought Models

Key Results:

- Our Faithful CoT model outperformed standard CoT prompting on 9 of 10 reasoning datasets
- It achieved new state-of-the-art results on 7 datasets, showing accuracy boost from faithfulness

- An ablation study showed pivotal role of the symbolic solver in achieving high accuracy
- Our analysis revealed strengths like robustness to exemplars and limitations like brittleness of symbolic reasoning

Discussion:
This work makes an important contribution in improving the faithfulness of chain of thought explanations from large language models. By decomposing reasoning into natural language and deterministic symbolic execution, the model is forced to explicitly show its work and ensure the coherence between the explanation chain and the predicted answer. This stands in contrast to standard CoT methods, where the reasoning chain does not necessarily reflect the model's actual inference process.

The results demonstrate that enforcing faithfulness not only improves interpretability, but also synergistically boosts accuracy across a diverse set of reasoning tasks. This highlights the importance of explanation faithfulness as a foundational principle for building transparent and trustworthy AI systems. The design of interleaving natural language with symbolic formalisms also strikes a balance between human-understandability and rigorous logical formalization.

While the approach still has limitations like brittleness of syntactic reasoning, it provides a general framework for injecting faithfulness into reasoning tasks across multiple domains. As large language models continue to advance in reasoning capabilities, ensuring their explainability through principles like faithfulness will only grow in importance. This work offers a promising path forward, and future work should explore methods to improve the faithfulness of the initial translation step as well. Overall, the faithful CoT approach represents an important step toward reliable and interpretable reasoning with language models.

Learning to Decompose:

Explicit decomposition modeling, which involves breaking down complex tasks into simpler, and often more interpretable, sub-tasks has long been a central theme in developing robust and interpretable NLU systems. Despite the many datasets and resources that have been built as part of this effort, the majority of them have small-scale annotations and limited scope, which is not enough to solve general decomposition tasks. In [77], we look at large-scale intermediate pre-training of decomposition-based transformers using distant supervision from comparable texts, in particular, large-scale parallel news. We show that with such intermediate pre-training, developing robust decomposition-based models for a diverse range of tasks becomes more feasible. On semantic parsing, our model improves 20% to 30% on two datasets, Overnight and TORQUE, over the baseline language model. We further use such a pre-trained model as the basis for building an end-to-end decomposition-based question answering system and show that our QA system improves over state-of-the-art models including GPT-3 on both HotpotQA and StrategyQA by 8% and 4%, respectively.

7.0 CONCLUSIONS

Even in the LLM era, when models are pre-trained on a huge number of examples in a self-supervised way, a lot of training data is needed, and a lot of supervision is still needed beyond the pre-training stage. Our program aimed at dramatically reducing the amount of labeled data required to learn, and we have shown in a sequence of work that with an innovative theory, work on structured prediction and neurosymbolic paradigms, the use of domain knowledge incorporated via constraints, and ingenious improvements over standard pre-training approaches, we could improve the state of the art in many applications and develop understanding of new paradigms, in ways that would greatly benefit the research community.

BIBLIOGRAPHY OF OUR PUBLISHED WORKS

1. Shuxiao Chen and Koby Crammer and Hangfeng He and Dan Roth and Weijie Su, "Weighted Training for Cross-Task Learning" ICLR - 2022
2. Shuxiao Chen and Hangfeng He and Weijie Su, "Label-Aware Neural Tangent Kernel: Toward Better Generalization and Local Elasticity" NeurIPS - 2020
3. Jeffrey Young-Min Cho, Harry Li Zhang, Chris Callison-Burch. “Unsupervised Entity Linking with Guided Summarization and Multiple-Choice Selection.” EMNLP 2022.
4. Soham Dan and Osbert Bastani and Dan Roth, "Few-Shot Novel Concept Learning for Semantic Parsing" EMNLP-Findings - 2021
5. Soham Dan, Osbert Bastani, Dan Roth. Understanding Robust Generalization in Learning Regular Languages. ICML 2022.
6. Soham Dan and Xinran Han and Dan Roth, "Compositional Data and Task Augmentation for Instruction Following" EMNLP-Findings - 2021
7. Soham Dan and Hangfeng He and Dan Roth, "Understanding Spatial Relations through Multiple Modalities" LREC - 2020
8. Soham Dan and Dan Roth, "On the Effects of Transformer Size on In- and Out-of-Domain Calibration" EMNLP-Findings - 2021
9. Soham Dan and Michael Zhou and Dan Roth, "Generalization in Instruction Following Systems" NAACL - 2021
10. Zhun Deng and Hangfeng He and Weijie Su, "Toward Better Generalization Bounds with Locally Elastic Stability" ICML - 2021
11. Daniel Deutsch and Rotem Dror and Dan Roth, "A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods" TACL - 2021
12. Daniel Deutsch and Tania Bedrax-Weiss and Dan Roth, "Towards Question-Answering as an Automatic Metric for Evaluating the Content Quality of a Summary" TACL - 2021
13. Daniel Deutsch and Rotem Dror and Dan Roth, "On the Limitations of Reference-Free Evaluations of Generated Text" EMNLP - 2022
14. Daniel Deutsch and Rotem Dror and Dan Roth, "Re-Examining System-Level Correlations of Automatic Summarization Evaluation Metrics" NAACL - 2022
15. Daniel Deutsch and Dan Roth, "SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics" Workshop on Natural Language Processing Open Source Software (NLP-OSS), EMNLP - 2020
16. Daniel Deutsch and Dan Roth, "Benchmarking Answer Verification Methods for Question Answering-Based Summarization Evaluation Metrics" ACL-Findings - 2022
17. Daniel Deutsch and Dan Roth, "Incorporating Question Answering-Based Signals into Abstractive Summarization via Salient Span Selection" EACL - 2023
18. Daniel Deutsch and Dan Roth, "Understanding the Extent to which Content Quality Metrics Measure the Information Quality of Summaries" CoNLL - 2021

19. Hantian Ding and Jinrui Yang and Yuqian Deng and Hongming Zhang and Dan Roth, "Towards Open-Domain Topic Classification" NAACL - 2022
20. Liam Dugan, Daphne Ippolito, Arun Kirubarajan, Sherry Shi, Chris Callison-Burch. "Real or Fake Text?: Investigating Human Ability to Detect Boundaries Between Human-Written and Machine-Generated Text." AAAI 2023.
21. Xingyu Fu and Ben Zhou and Ishaan Preetam Chandratreya and Carl Vondrick and Dan Roth, "There's a Time and Place for Reasoning Beyond the Image" ACL - 2022
22. Hannah Gonzalez, Liam Dugan, Eleni Miltsakaki, Zhiqi Cui, Jiaxuan Ren, Bryan Li, Shriyash Upadhyay, Etan Ginsberg, Chris Callison-Burch "Enhancing Human Summaries for Question-Answer Generation in Education." Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023).
23. Hannah Gonzalez, Jiening Li, Helen Jin, Jiaxuan Ren, Hongyu Zhang, Ayotomiwa Akinyele, Adrian Wang, Eleni Miltsakaki, Ryan Baker, Chris Callison-Burch, "Automatically Generated Summaries of Video Lectures May Enhance Students' Learning Experience." Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023).
24. Ruohao Guo and Dan Roth, "Constrained Labeled Data Generation for Low-resource Named Entity Recognition" ACL-Findings - 2021
25. Nitish Gupta and Kevin Lin and Dan Roth and Sameer Singh and Matt Gardner, "Neural Module Networks for Reasoning over Text" ICLR - 2020
26. Nitish Gupta and Sameer Singh and Matt Gardner and Dan Roth, "Paired Examples as Indirect Supervision in Latent Decision Models" EMNLP - 2021
27. Hangfeng He and Qiang Ning and Dan Roth, "QUASE: Question-Answer Driven Sentence Encoding" ACL - 2020
28. Hangfeng He and Mingyuan Zhang and Qiang Ning and Dan Roth, "Foreseeing the Benefits of Incidental Supervision" EMNLP - 2021
29. J Huang, Z Li, B Chen, K Samel, M Naik, L Song, X Si, "Scallop: From probabilistic deductive databases to scalable differentiable reasoning" *Advances in Neural Information Processing Systems 34 (2021)*, 25134-25145
30. Philip Huebner and Elior Sulem and Cynthia Fisher and Dan Roth, "BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language" CoNLL - 2021
31. Daphne Ippolito, Liam Dugan, Emily Reif, Ann Yuan, Andy Coenen, Chris Callison-Burch. "The Case for a Single Model that can Both Generate Continuations and Fill-in-the-Blank." NAACL 2022.
32. Zsolt Kira, Co-Lead Organizer, "Learning from Limited and Imperfect Data (L2ID)" *2022 ECCV Workshop* https://www.youtube.com/@learningwithlimitedandimpe9143/playlists, accessed 8/16/23.
33. Jordan Kodner and Nitish Gupta, "Overestimation of Syntactic Representations in Neural Language Models" ACL - 2020

34. Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, Nicholas Carlini. “Deduplicating Training Data Makes Language Models Better.” ACL 2022.
35. Bryan Li, Mohammad Sadegh Rasooli, Ajay Patel, Chris Callison-Burch. “Multilingual Bidirectional Unsupervised Translation Through Multilingual Finetuning and Back-Translation.” 2023 LoResMT @ EACL 2023
36. Sha Li, Ruining Zhao, Manling Li, Heng Ji, Chris Callison-Burch, Jiawei Han. “Open-Domain Hierarchical Event Schema Induction by Incremental Prompting and Verification.” ACL 2023.
37. Ziyang Li, Jiani Huang, Mayur Naik, “Scallop: A Language for Neurosymbolic Programming” PLDI 2023.
38. YC Liu, CY Ma, J Tian, Z He, Z Kira, “Polyhistor: Parameter-Efficient Multi-Task Adaptation for Dense Vision Tasks” (*NeurIPS 2022*)
39. YC Liu, CY Ma, X Dai, J Tian, P Vajda, Z He, Z Kira, “Open-Set Semi-Supervised Object Detection” *(ECCV 2022)*
40. YC Liu, CY Ma, Z Kira, “Unbiased Teacher v2: Semi-Supervised Object Detection for Anchor-Free and Anchor-Based Detectors” *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022),* 9819-9828
41. YC Liu, CY Ma, Z He, CW Kuo, K Chen, P Zhang, B Wu, Z Kira, P Vajda, “Unbiased Teacher for Semi-Supervised Object Detection” *(ICLR 2021)*
42. Josh Magnus Ludan, Yixuan Meng, Tai Nguyen, Saurabh Shah, Qing Lyu, Marianna Apidianaki, Chris Callison-Burch. “Explanation-based Finetuning Makes Models More Robust to Spurious Cues.” ACL 2023.
43. Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, Chris Callison-Burch. “Faithful Chain-of-Thought Reasoning.” AACL-IJCNLP 2023.
44. Qing Lyu and Hongming Zhang and Elior Sulem and Dan Roth, "Zero-shot Event Extraction via Transfer Learning: Challenges and Insights" ACL - 2021
45. Qing Lyu*, Li Zhang*, Chris Callison-Burch. “Reasoning about Goals, Steps, and Temporal Ordering with WikiHow.” EMNLP 2020.
46. Qing Lyu, Hua Zheng, Daoxin Li, Li Zhang, Marianna Apidianaki, Chris Callison-Burch. “Is ‘My Favorite New Movie’ My Favorite Movie? Probing the Understanding of Recursive Noun Phrases.” NAACL 2022.
47. Bonan Min and Hayley Ross and Elior Sulem and Amir Pouran Ben Veyseh and Thien Huu Nguyen and Oscar Sainz and Eneko Agirre and Ilana Heinz and Dan Roth,"Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey" ACM Computing Surveys - 2023
48. Ajay Patel, Bryan Li, Mohammad Sadegh Rasooli, Noah Constant, Colin Raffel, Chris Callison-Burch. “Bidirectional Language Models Are Also Few-shot Learners.” ICLR 2023.

49. Zheng Qi and Elior Sulem and Haoyu Wang and Xiaodong Yu and Dan Roth, "Capturing the Content of a Document through Complex Event Identification" *SEM - 2022
50. Mukund Raghothaman, Jonathan Mendelson, David Zhao, Mayur Naik, Bernhard Scholz, “Provenance-Guided Synthesis of Datalog Programs.” POPL 2020.
51. Mohammad Sadegh Rasooli, Chris Callison-Burch, Derry Wijaya. “‘Wikily’ Supervised Neural Translation Tailored to Cross-Lingual Tasks.” EMNLP 2021.
52. Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, Jason Wei. “A Recipe For Arbitrary Text Style Transfer with Large Language Models.” ACL 2022.
53. Alla Rozovskaya and Dan Roth, "How Good (really) are Grammatical Error Correction Systems?" EACL - 2021
54. Krunal Shah and Nitish Gupta and Dan Roth, "What Do We Expect from Multiple-Choice QA Systems?" EMNLP-Findings - 2020
55. Erfan Sadeqi Azer and Daniel Khashabi and Ashish Sabharwal and Dan Roth, "Not All Claims are Created Equal: Choosing the Right Statistical Approach to Assess Hypotheses" ACL - 2020
56. Sanjay Subramanian and Ben Bogin and Nitish Gupta and Tomer Wolfson and Sameer Singh and Jonathan Berant and and Matt Gardner, "Obtaining Faithful Interpretations from Compositional Neural Networks" ACL - 2020
57. Elior Sulem and Jamaal Hay and Dan Roth, "Do We Know What We Don't Know? Studying Unanswerable Questions beyond SQuAD 2.0" EMNLP-Findings - 2021
58. Elior Sulem and Jamaal Hay and Dan Roth, "Yes, No or IDK: The Challenge of Unanswerable Yes/No Questions" NAACL - 2022
59. Aalok Thakkar, Nathaniel Sands, George Petrou, Rajeev Alur, Mayur Naik, Mukund Raghothaman, “Mobius: Synthesizing Relational Queries with Recursive and Invented Predicates” OOPSLA 2023.
60. Kaifu Wang and Hangfeng He and Tin D. Nguyen and Piyush Kumar and Dan Roth, "On Regularization and Inference with Label Constraints" ICML - 2023
61. Kaifu Wang and Qiang Ning and Dan Roth, "Learnability with Indirect Supervision Signals" NeurIPS - 2020
62. Kaifu Wang and Efthymia Tsamoura and Dan Roth, "On Learning Latent Models with Multi-Instance Weak Supervision" NeurIPS - 2023
63. Yinjun Wu, Adam Stein, Jacob Gardner, Mayur Naik, “Learning to Select Pivotal Samples for Meta Re-weighting” AAAI 2023.
64. Yahan Yang and Elior Sulem and Insup Lee and Dan Roth, "Bootstrapping Small & High Performance Language Models with Unmasking-Removal Training Policy" EMNLP - 2023
65. Yahan Yang and Elior Sulem and Insup Lee and Dan Roth, "Penn & BGU BabyBERTa+ for Strict-Small BabyLM Challenge" BabyLM Challenge (technical report) - 2023

66. Yue Yang, Artemis Panagopoulou, Marianna Apidianaki, Mark Yatskar and Chris Callison-Burch. “Visualizing the Obvious: A Concreteness-based Ensemble Model for Noun Property Prediction.” Findings of EMNLP 2022.
67. Yue Yang, Artemis Panagopoulou, Qing Lyu, Li Zhang, Mark Yatskar, Chris Callison-Burch. “Visual Goal-Step Inference using wikiHow.” EMNLP 2021.
68. Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, Mark Yatskar. “Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification.” CVPR 2023.
69. Ann Yuan, Daphne Ippolito, Vitaly Nikolaev, Chris Callison-Burch, Andy Coenen, and Sebastian Gehrmann. “SynthBio: A Case Study in Human-AI Collaborative Curation of Text Datasets.” NeurIPS 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. 2021.
70. Hanlin Zhang, Jiani Huang, Ziyang Li, Mayur Naik, Eric Xing, “Improved Logical Reasoning of Language Models via Differentiable Symbolic Programming” Findings of ACL 2023.
71. Li Zhang, Liam Dugan, Hainiu Xu, Chris Callison-Burch, “Exploring the Curious Case of Code Prompts.” Proceedings of the 1st Workshop on Natural Language Reasoning and Structured Explanations (NLRSE).
72. Li Zhang, Qing Lyu, Chris Callison-Burch. “Intent Detection with WikiHow.” AACL-IJCNLP 2020.
73. Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, Chris Callison-Burch. “Causal Reasoning About Entities and Events in Procedural Texts.” Findings of EACL 2023.
74. Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Leon Zhou, Hainiu Xu, Li Zhang, Lara Martin, Rotem Dror, Sha Li, Heng Ji, Martha Palmer, Susan Windisch Brown, Reece Suchocki, Chris Callison-Burch, “Human-in-the-loop Schema Induction.” ACL 2023 System Demonstrations.
75. Z Zhao, K Samel, B Chen, L Song, “Proto: Program-guided transformer for program-guided tasks” *Advances in neural information processing systems 34 (2021)*, 17021-17036
76. Ben Zhou and Kyle Richardson and Qiang Ning and Tushar Khot and Ashish Sabharwal and Dan Roth, "Temporal Reasoning on Implicit Events from Distant Supervision" NAACL – 2021
77. Ben Zhou and Kyle Richardson and Xiaodong Yu and Dan Roth, “Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts” EMNLP 2022
78. Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, Prithviraj Ammanabrolu, “I Cast Detect Thoughts: Learning to Converse and Guide with Intents and Theory-of-Mind in Dungeons and Dragons.” ACL 2023.

79. Shuyan Zhou, Li Zhang, Yue Yang, Qing Lyu, Pengcheng Yin, Chris Callison-Burch, Graham Neubig. “Show Me More Details: Discovering Hierarchies of Procedures from Semi-structured Web Data.” ACL 2022.
80. Andrew Zhu, Karmanya Aggarwal, Alexander Feng, Lara Martin, Chris Callison-Burch. “FIREBALL: A Dataset of Dungeons and Dragons Actual-Play with Structured Game State Information.” ACL 2023.

Publications in Submission

81. Xingyu Fu, Ben Zhou, Sihao Chen, Mark Yatskar, Dan Roth, “Dynamic Clue Bottlenecks: Inherently Interpretable VQA with Abductive Reasoning.” In submission.
82. Jiani Huang, Ziyang Li, David Jacobs, Mayur Naik, Ser-Nam Lim, “LASER: Neuro-Symbolic Learning of Semantic Video Representations.” In submission.
83. Stephen Mell, Steve Zdancewic, and Osbert Bastani, “Optimal Synthesis of Neurosymbolic Programs via Abstract Interpretation.” In submission.

LIST OF SYMBOLS, ABBREVIATIONS, AND ACRONYMS

CoT: Chain of Thought
CV: Computer Vision
DARPA: Defense Advanced Research Projects Agency
DFA: Deterministic Finite-State Automaton
EMA: Exponential Moving Average
FITB: Fill in the Blank
FLASH: Fast Learning via Auxiliary signals, Structured knowledge, and Human expertise
GPT: Generative Pre-trained Transformer (language model)
GQA: Graph Question Answering dataset
ID: In-Distribution
ILP: Inductive Logic Programming
LLM: Large Language Model
LwLL: Learning with Less Labeling
MNIST: Modified National Institute of Standards and Technology database
NER: Named Entity Recognition
NL: Natural Language
NLP: Natural Language Processing
NLU: Natural Language Understanding
OOD: Out-of-Distribution
OSS: Open Source Software
OSSOD: Open-Set Semi-Supervised Object Detection
PABI: PAC-Bayesian motivated informativeness measure
PDDL: Planning Domain Definition Language
PLL: Partial Label Learning
QA: Question Answering
PoS: Part of Speech
RPN: Region Proposal Network
RNN: recurrent neural network
ROI: Region of Interest
SAP: Sequential Autoregressive Prompting
SL: Symbolic Language
SOTA: State-of-the-art
SSL: Semi-Supervised Learning
SS-OD: Semi-Supervised Object Detection
TAWT Target-Aware Weighted Training
VQA: Visual Question Answering
VQAR: Visual Question Answering with Reasoning