UNIVERSITY OF MICHIGAN

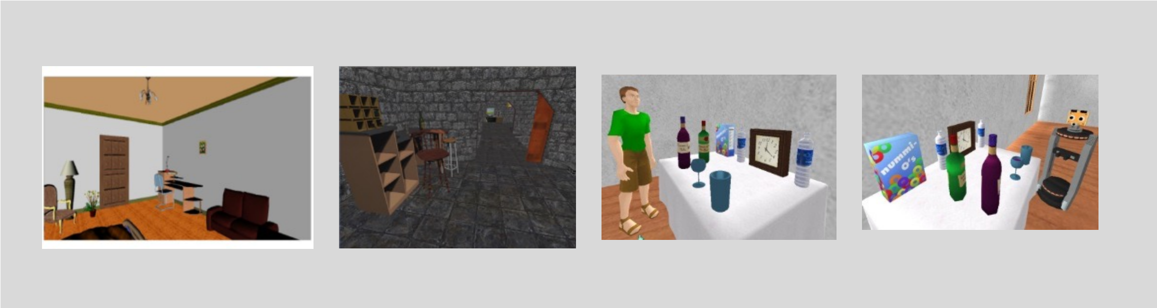
Large Pre-Trained Language Models for Physical Action Understanding and Planning

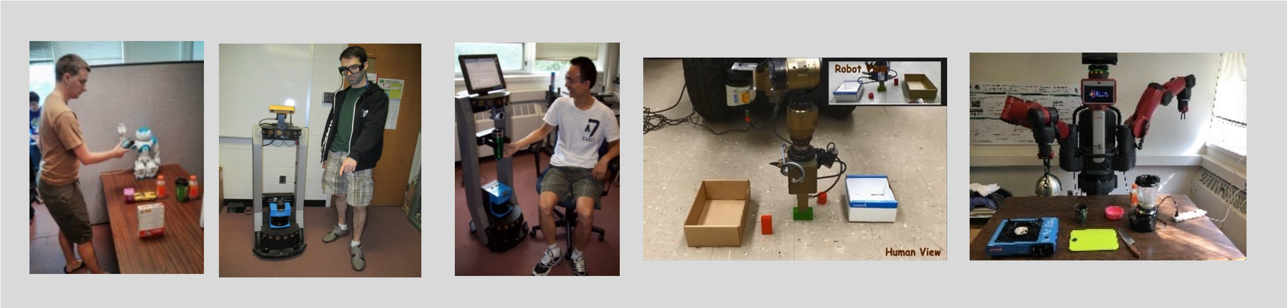
Shane Storks & Jianing “Jed” Yang

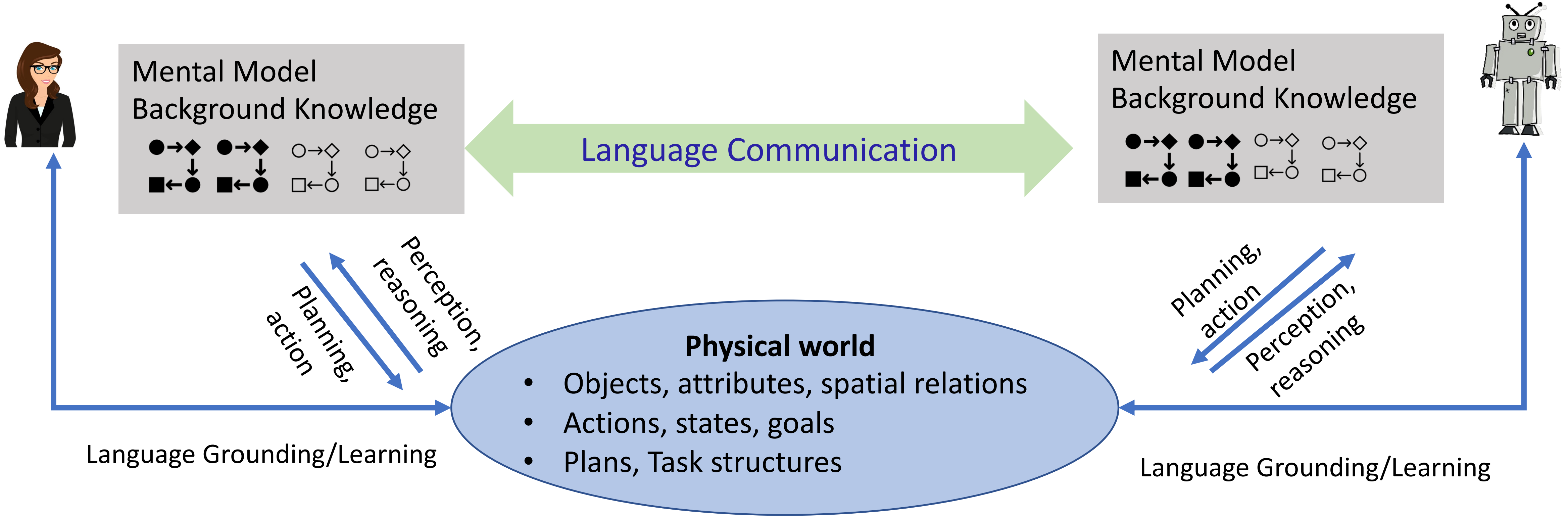
SLED Research Group @ University of Michigan

Oct. 21st, 2022

Situated Language and **Embodied** Dialog







(Slide from Joyce Chai)

Understanding Physical Causality

6 -8 months

Notice relationships between events. Perform basic actions to make things happen

18 months

Combine simple actions to make things happen. Change the way how they interact with the world to see how it changes the outcome

36 months

Make prediction about what may happen and reflect upon what caused something to happen

(Slide from Joyce Chai)

Alan M Leslie and Stephanie Keeble. 1987. Do six-month-old infants perceive causality? Cognition,25(3):265–288

Lisa M Oakes and Leslie B Cohen. 1990. Infant perception of a causal event. Cognitive Development,5(2):193–207

Elizabeth S Spelke. 1994. Initial knowledge: six suggestions. Cognition, 50(3):431-45.

Outline

1. Understanding the ability of large language models (LMs) to learn **verifiable physical commonsense reasoning**
2. Applying large LMs as a tool to inform **planning of physical actions**

Motivation

- NLP tasks commonly boil natural language understanding (NLU) down to simple text classification tasks
  - Data bias and lack of transparency make it unclear whether underlying problems are truly solved
  - We want to examine system's underlying reasoning capability
- Tiered Reasoning for Intuitive Physics (`TRIP`) provides traces of a multi-tiered, human-annotated reasoning process:
  - Low-level, concrete physical states
  - High-level end task of plausibility classification

Storks, S., Gao, Q., & Chai, J. (2021). Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. Findings of EMNLP 2021.

Tiered Reasoning for Intuitive Physics (TRIP)

Story A

1. Ann sat in the chair.
2. Ann turned off the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. Ann wrote in the book.

Story B

1. Ann sat in the chair.
2. **2.** Ann turned off the telephone.
3. Ann picked up a pencil.
4. Ann opened the book.
5. **5.** Ann heard the telephone ring.

!

Which story is more plausible? A

Why not B***?***

Conflicting sentences: 2 → 5

Physical states:

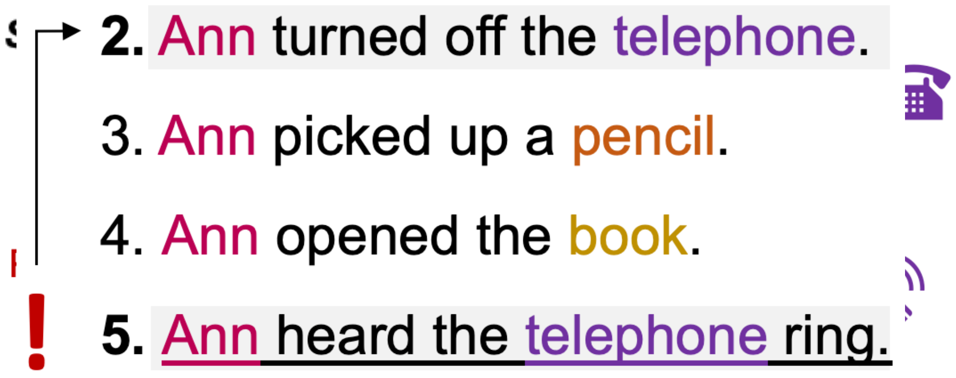
Powered(telephone) → ¬Powered(telephone)

X

Powered(telephone) → Powered(telephone) Running(telephone)

Storks, S., Gao, Q., & Chai, J. (2021). Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding. Findings of EMNLP 2021.

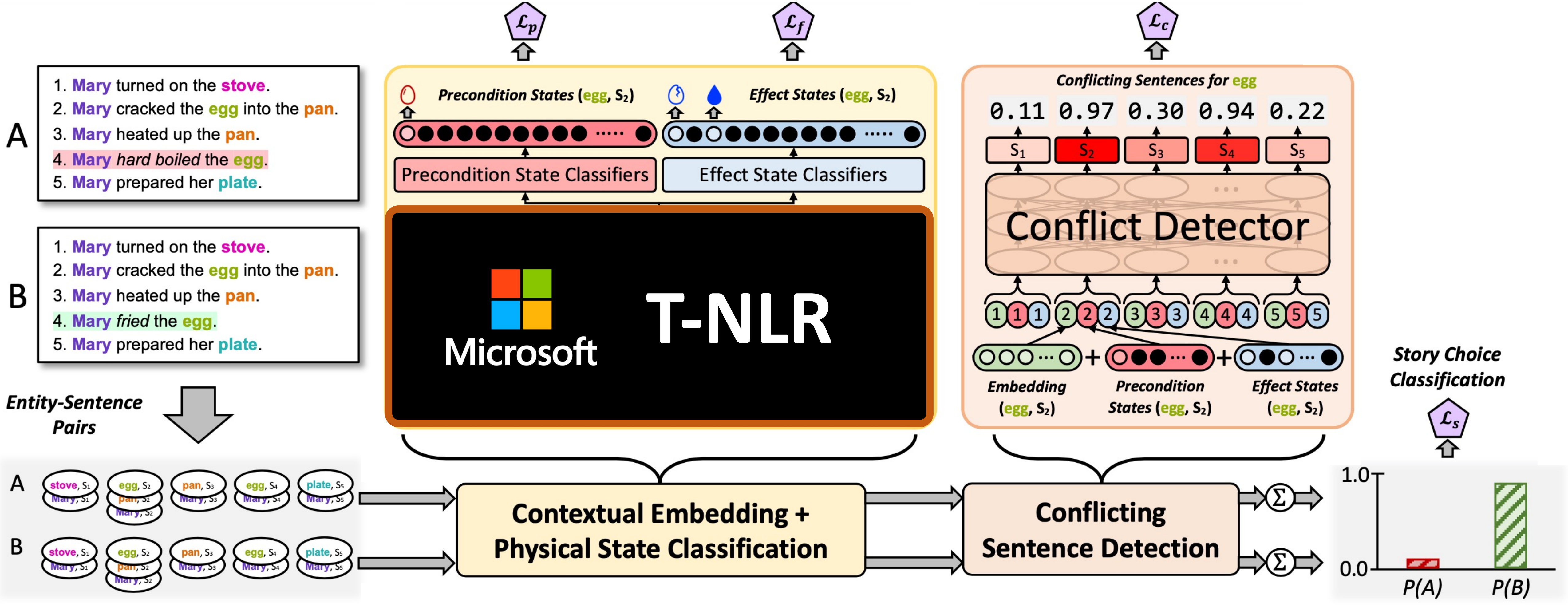
Evaluation Metrics



Metric	Story Choice	Conflicting Sentences	Physical States
Accuracy	✓		
Consistency	✓	✓	
Verifiability	✓	✓	✓

Goal: Accuracy ≈ Consistency ≈ Verifiability

Tiered Baseline



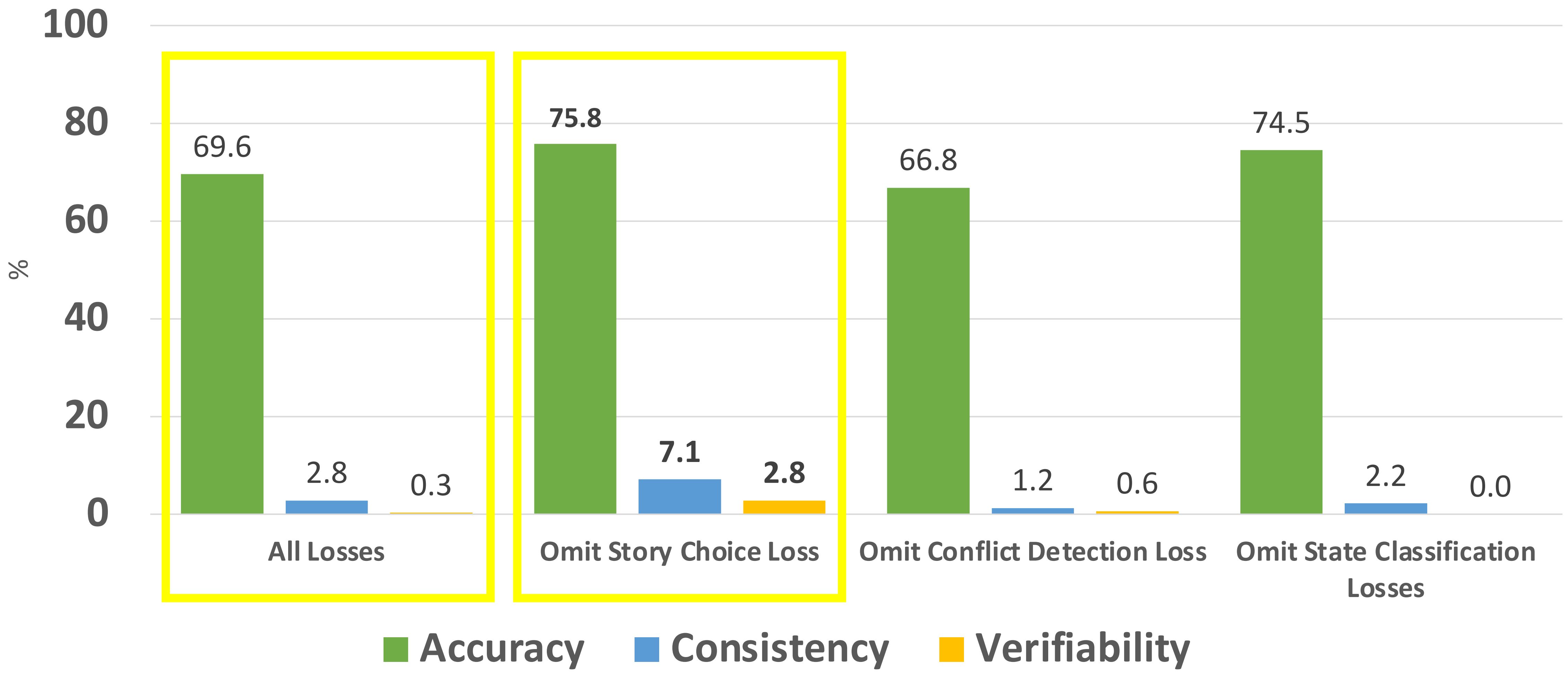
$$\mathcal{L} = \lambda_p \mathcal{L}_p + \lambda_f \mathcal{L}_f + \lambda_c \mathcal{L}_c + \lambda_s \mathcal{L}_s$$

Simply fine-tuning a pre-trained LM on the end task (plausibility prediction) **can achieve up to 97% accuracy**.

However…

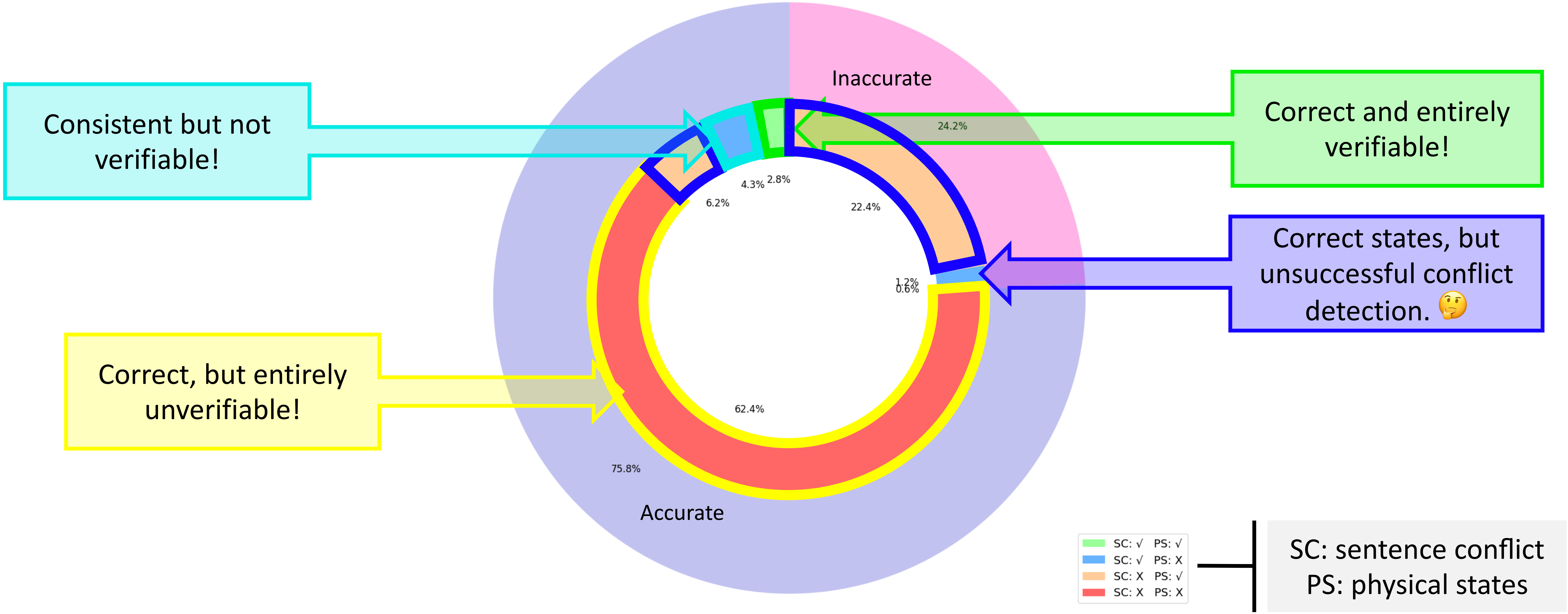
(Slide from Joyce Chai)

Results of T-NLR (Large) on TRIP

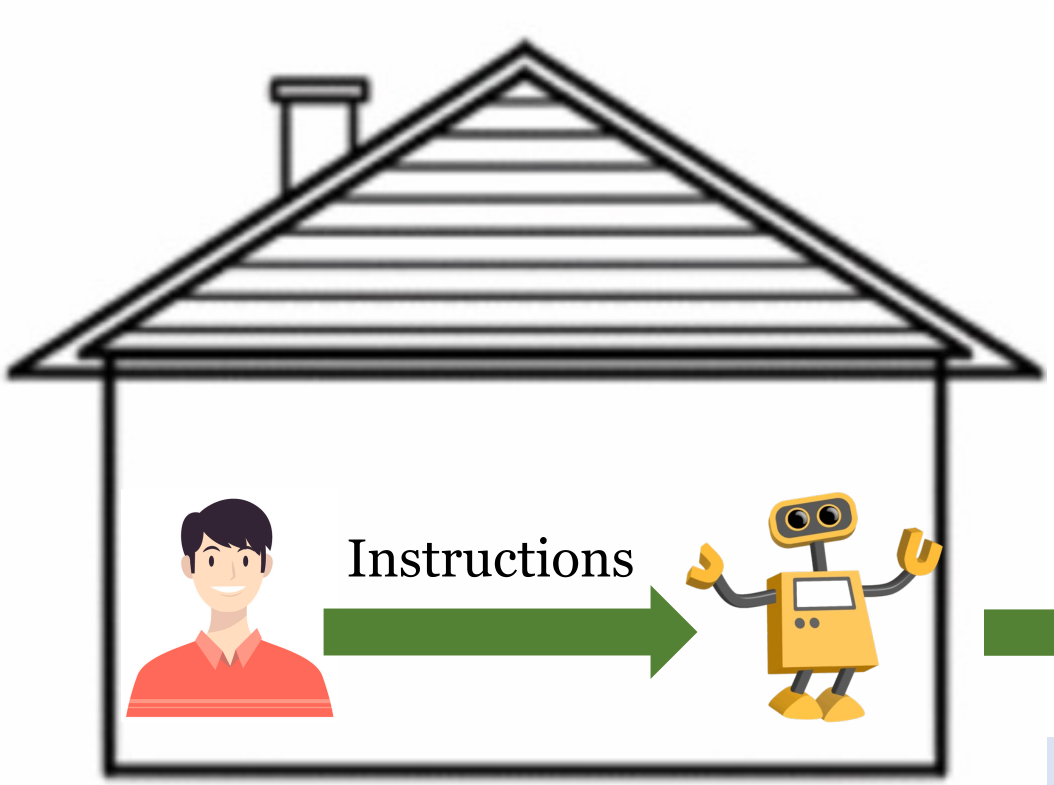


Gao, J & Tiwary, S. (2021). Efficiently and effectively scaling up language model pretraining for best language representation model on GLUE and SuperGLUE.

Error Distribution of T-NLR



Embodied Task Reasoning



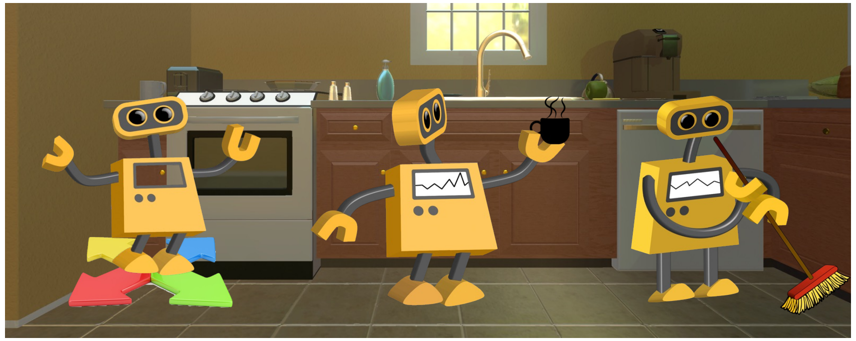
Household Task Domain

Task Learning

Navigation

Manipulation

Compositional

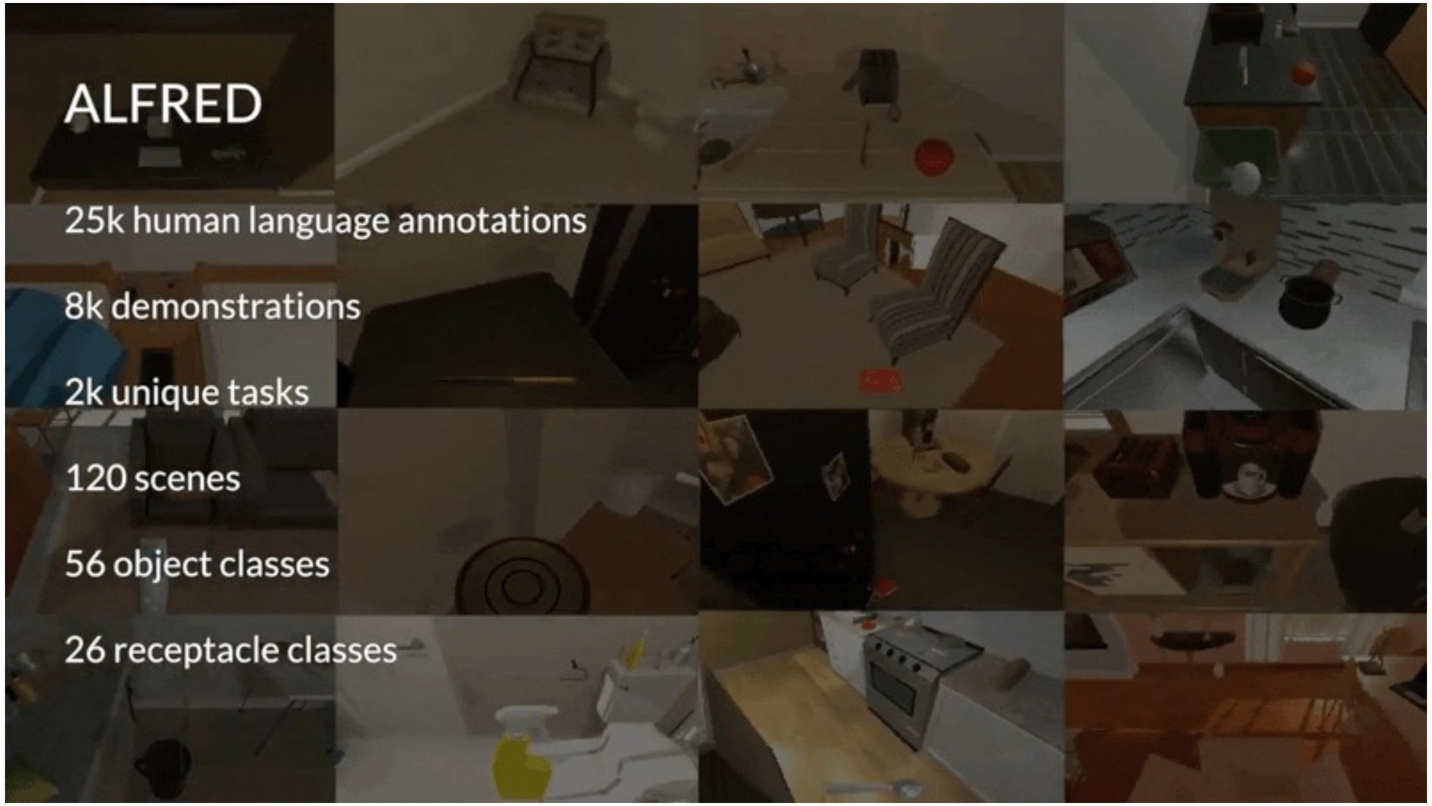


- Pick and place
- Do cleaning
- Prepare coffee
- Cook food ...

Generalize to unseen scenes

Task Reasoning in Simulated Environment

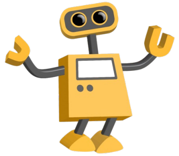
ALFRED (Action Learning From Realistic Environments and Directives)



High-level Goal Directive

Low-level Instructions

⇩



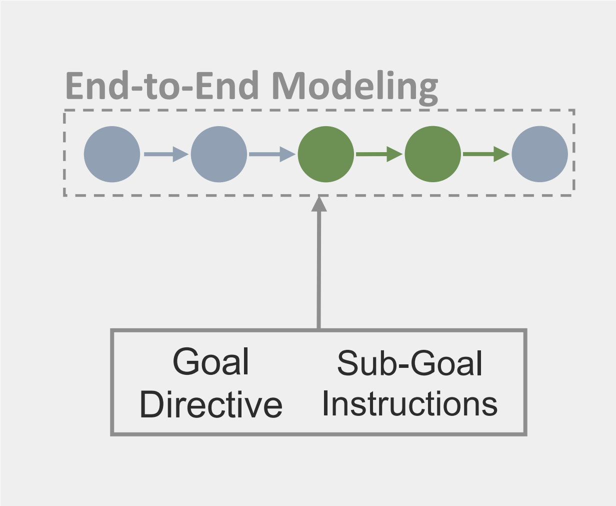
Visual Navigation

Object Interaction

Shridhar et al. ALFRED: A Benchmark for Interpreting Grounded Instruction for Everyday Tasks, CVPR 2020

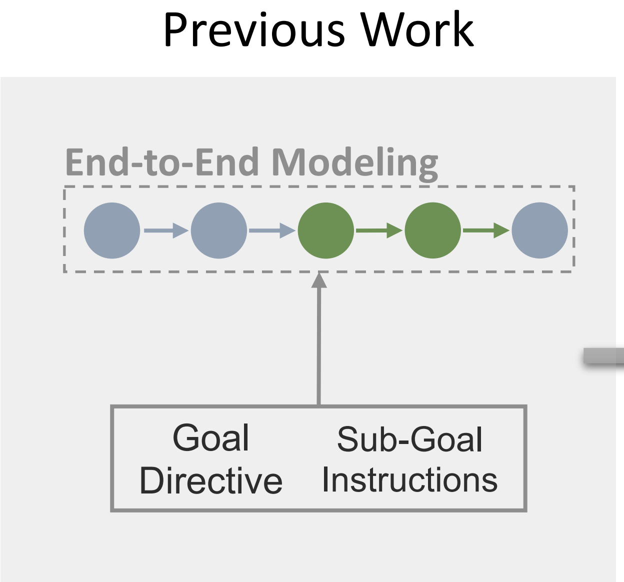
Hierarchical Task Learning with Unified Transformers (HiTUT)

Previous Work

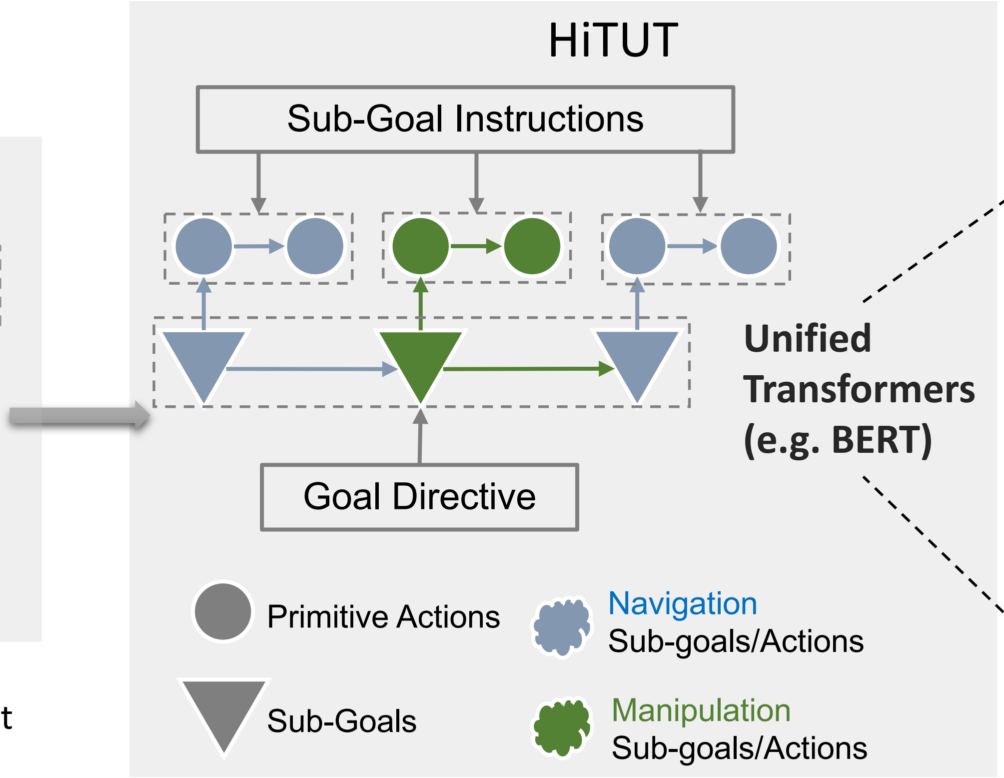


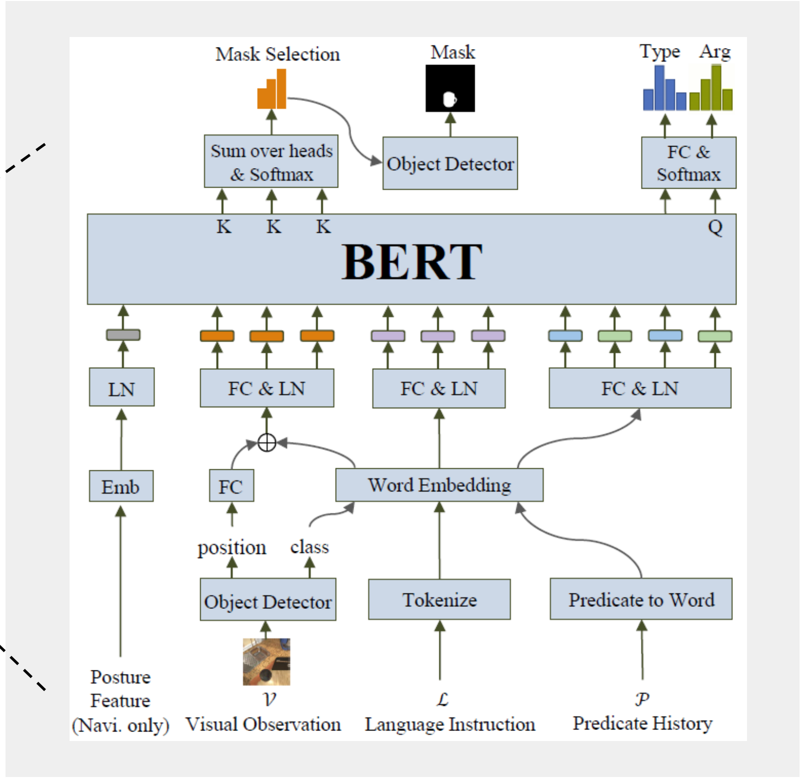
(Shridhar et al., 2020; Pratap Singh et al., 2020; Storks et al., 2021)

Hierarchical Task Learning with Unified Transformers (HiTUT)

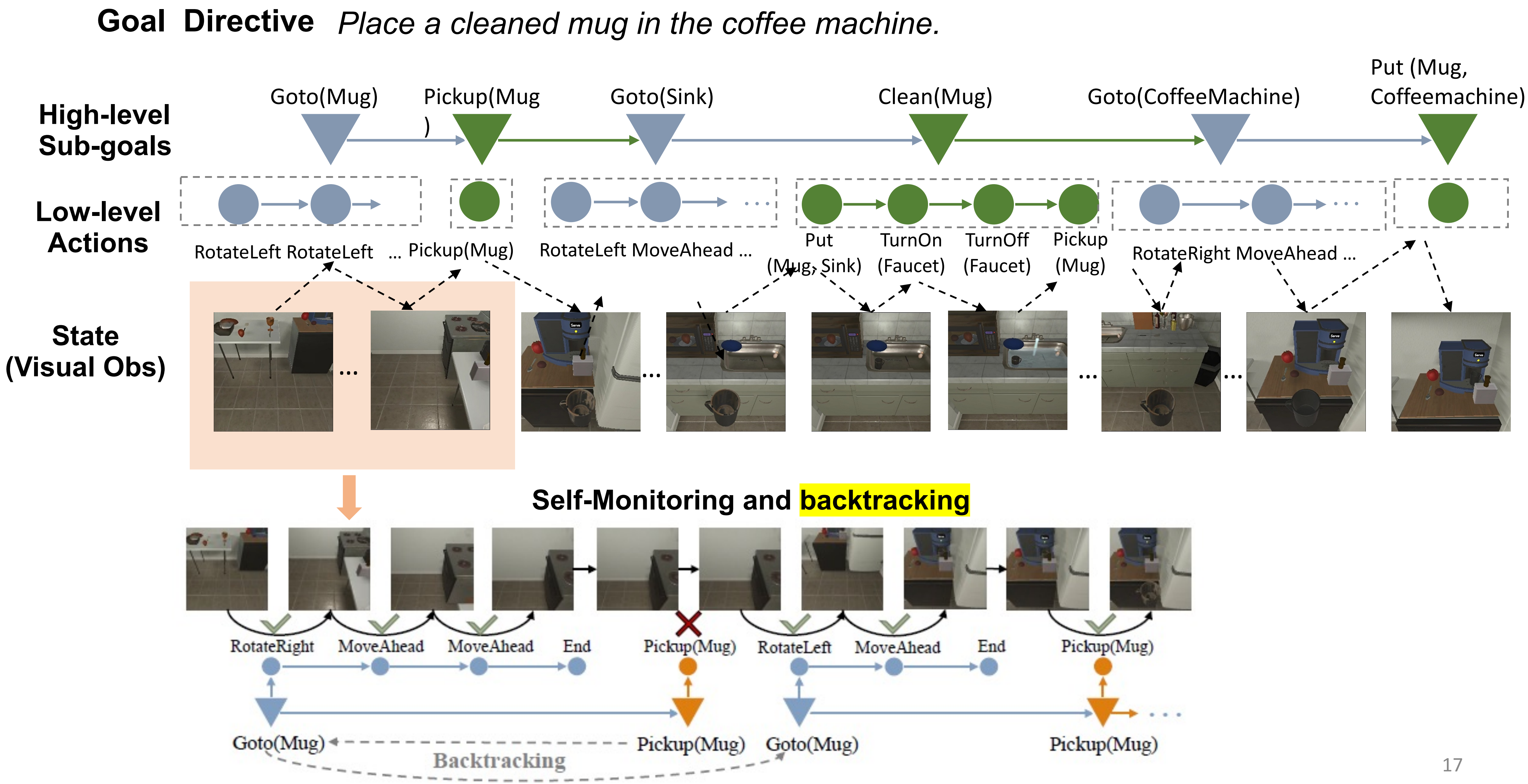


(Shridhar et al., 2020; Pratap Singh et al., 2020; Storks et al., 2021)

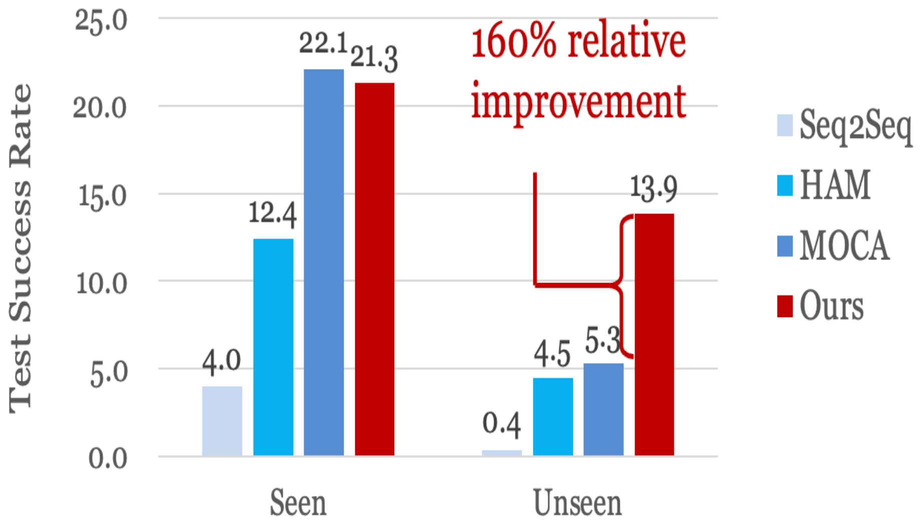




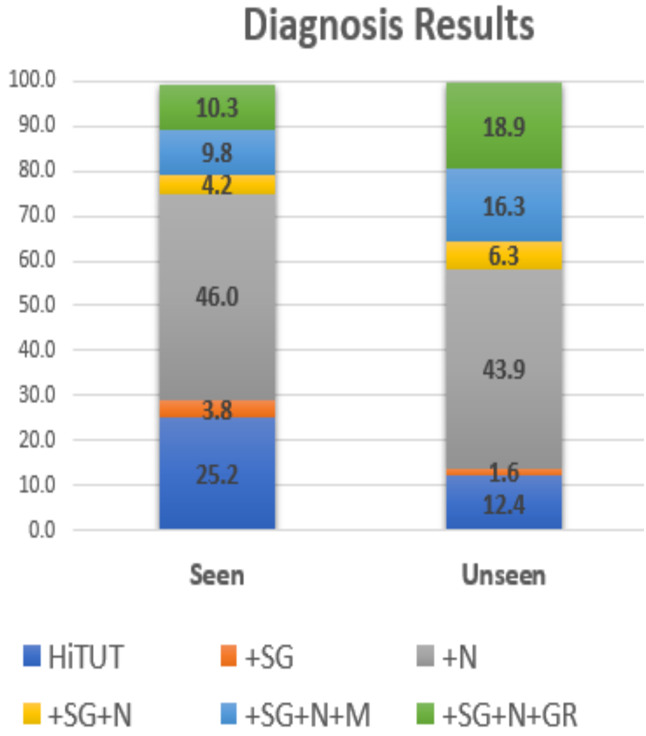
Hierarchical Task Learning with Unified Transformers (HiTUT)



Results: Better Generalization in Unseen Environment



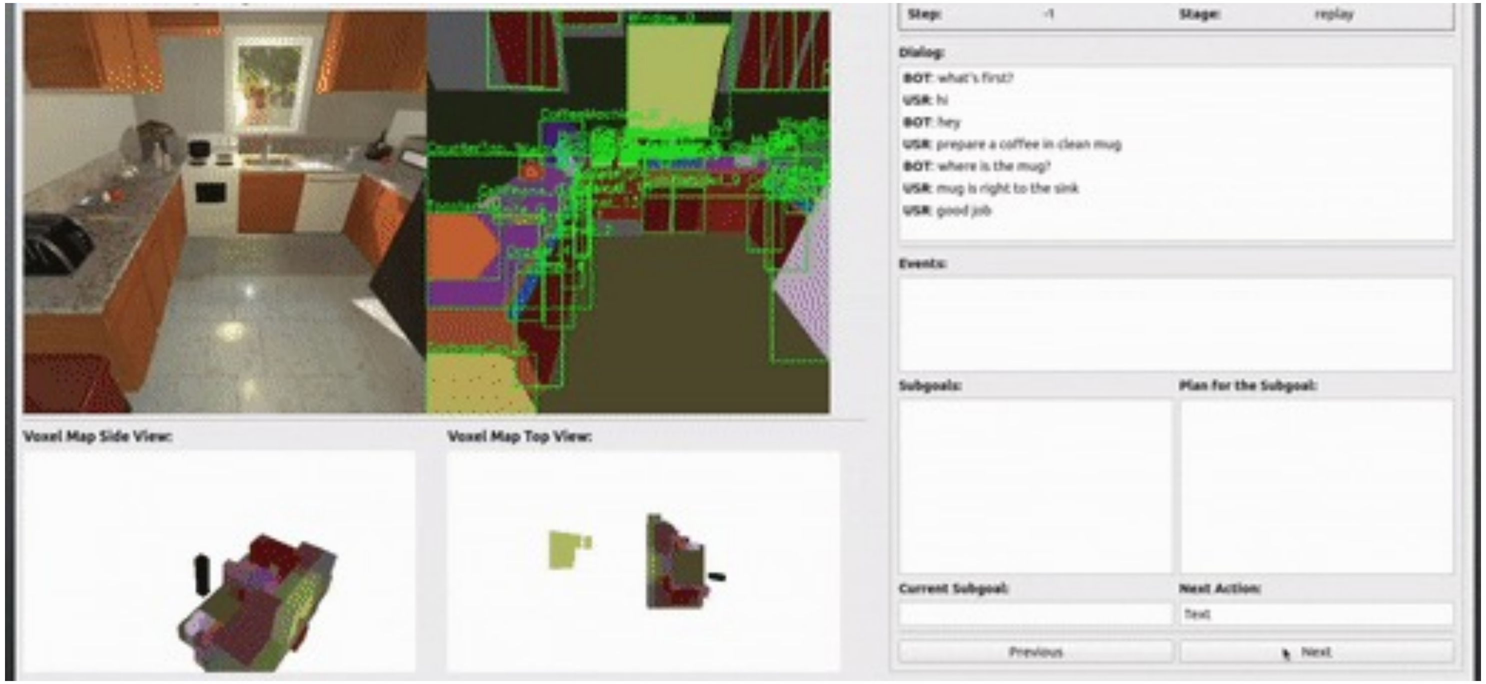
Task Goal:
Put two books on the desk.

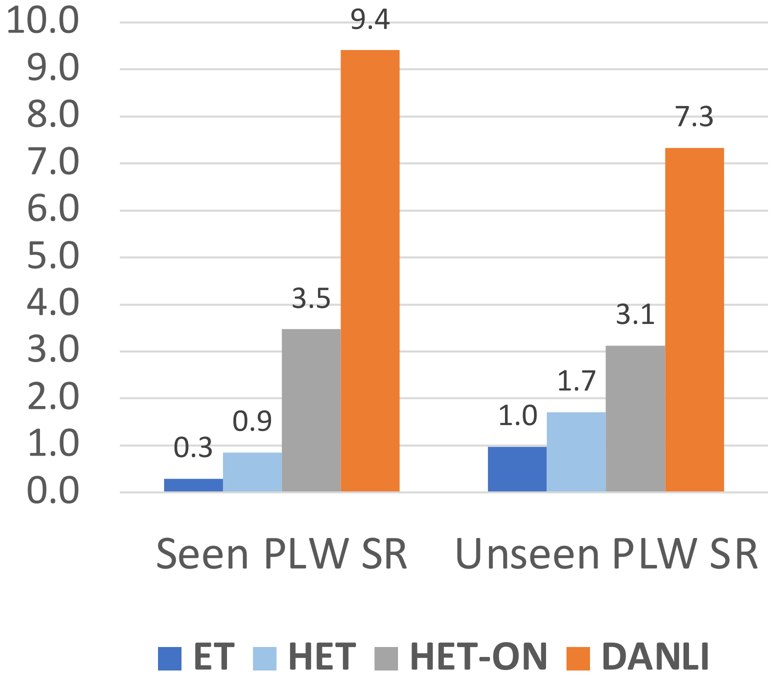


Example of how backtracking helps the agent recover from execution errors.

DANLI: A Deliberative Neuro-symbolic Agent

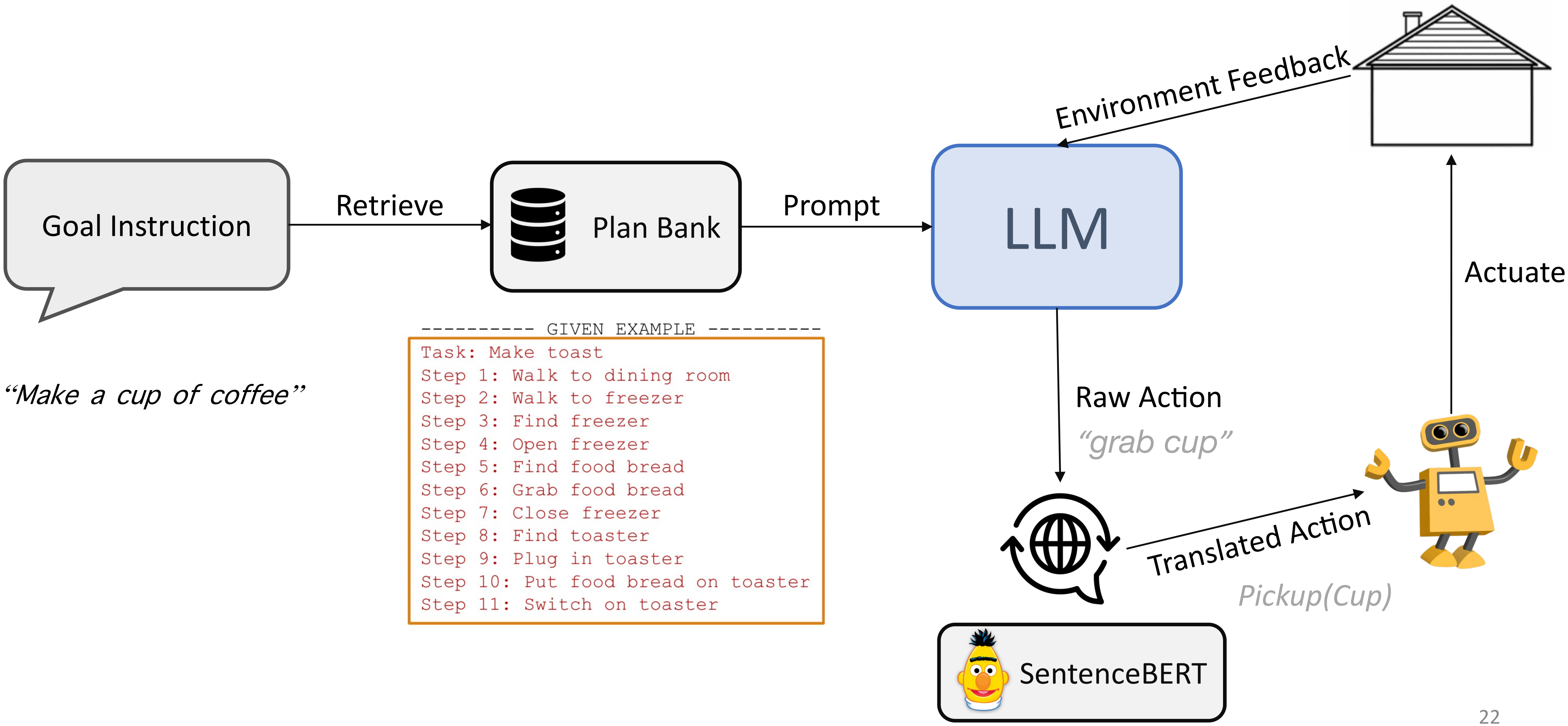
- Leverages neural models to predict subgoals from dialog history, and constructs a 3D voxel map representation from agent's ego-centric vision
- Support symbolic reasoning and proactive planning using a PDDL-based online planning algorithm





DANLI: Deliberative Agent for Following Natural Language Instructions. EMNLP 2022.

LLM for Long-horizon Planning: Pipeline



Evaluation: Zero-Shot Action Prediction

	GPT-2 (1.5B)	Turing-7B
Action Error Rate (↓)	89.4%	**40.6%**
Acc. (Type+Arg) (↑)	1.5%	**3%**
Acc. (Type) (↑)	**25.2%**	23.3%
Acc. (Arg) (↑)	4.5%	**8.7%**

Qualitative Examples: GPT-2 & Turing-7B

Prompt: “<Example_Tasks> Task: Do the laundry, Step 1: _____”

GPT-2 (1.5B)			Turing (7B)		
	GPT-2 action	**ALFRED action**		**Turing action**	**ALFRED action**
0	walk to kitchen	Open drawer	0	walk to bedroom	Go to bed
1	pull out clothes	Move dresser	1	wake up	Pick up alarm clock
2	move closet	Move shelving unit	2	plug in alarm clock	Put alarm clock on box
cut off action	move bookshelf	Move shelving unit	3	go to bathroom	Go to toilet
			4	pick up toilet brush	Pick up toilet paper
			5	put toilet paper on box	Put toilet paper on drawer
			6	pick up towel	Pick up towel
			7	put towel on rack	Put towel on towel holder
			8	pick up washcloth	Pick up cloth
			9	put cloth on clothesline	Put cloth on tv stand
			10	pick up clothes	Pick up cloth
			11	put cloth on coffee table	Put cloth on coffee table
			12	pick up cloth	Pick up cloth
			13	put cloth on kitchen counter	Put cloth on counter top
			14	pick up cloth	Pick up cloth
			15	put cloth on floor	Put cloth on desk
			16	pick up cloth	Pick up cloth
			17	put cloth on floor	Put cloth on desk

	repetitive
	mistranslation

Qualitative Examples: Turing-7B & Turing-530B

Prompt: “<Example_Tasks> Task: Make breakfast, Step 1: _____”

Turing (7B)			Turing (530B)		
	Turing action	**ALFRED action**		**Turing action**	**ALFRED action**
0	walk to dining room	Go to dining table	**0**	make toast	Toggle toaster
cut off action	put food on table	put plate on dining table	**1**	walk to dining room	Go to dining table
			2	find knife	Go to knife
			3	pick up knife	Pick up knife
			4	go to food bread	Go to bread
			5	cut bread	Slice bread
			6	place bread on plate	Put bread on plate
			7	go to cupboard	Put cup on cabinet
			8	pick up cup	Pick up cup
			9	fill cup with water	Fill watering can
			cut off action	water plants	Fill watering can

	repetitive
	mistranslation

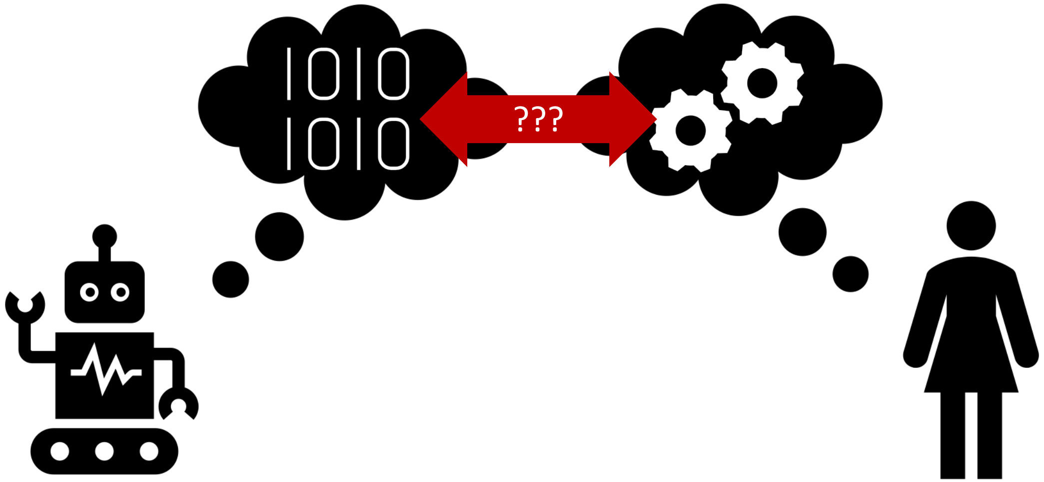
Summary

- While large LMs (such as T-NLR) make some steps toward coherent reasoning for NLU, more work is needed toward neuro-symbolic reasoning pipelines for teaching systems how to reason about the physical world.
- Large generative LMs (such as T-NLG) demonstrates some initial capability of zero-shot task planning, but still has large gap compared to fine-tuned LMs. More work is needed for translating and grounding LLM outputs to unseen task domains.

[Coalescing Global and Local Information for Procedural Text Understanding. COLING 2022.]
[Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. 2022.]
[Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. 2022.]
[Inner Monologue: Embodied Reasoning through Planning with Language Models. 2022.]

Future Work

- Commonsense reasoning with large generative LMs
  - Analogy and relational reasoning
  - Generalized physical commonsense reasoning
- Action planning with large generative LMs
  - Close-loop planning utilizing environmental and interactive feedbacks







Shane Storks | Jianing “Jed” Yang | Qianqi Yan | Wenfei Tang | Joyce Y. Chai

@shanestorks | @jed_yang