Understanding Procedures

Lara J. Martin (she/they)

https://laramartin.net/interactive-fiction-class

Slides adapted from Li "Harry" Zhang and Qing "Veronica" Lyu

Learning Objectives

Define what a procedure is

See how procedures can be recalled, adapted, and generated

Examine different ways of evaluating script generation

Determine how procedures can be used in interactive fiction

Review: Levels of Information

‘What’s it going to be then, eh?’

There was me, that is Alex, and my three droogs, that is Pete, Georgie, and Dim, Dim being really dim, and we sat in the Korova Milkbar making up our rassoodocks what to do with the evening, a flip dark chill winter bastard though dry. The Korova Milkbar was a milk-plus mesto, and you may, O my brothers, have forgotten what these mestos were like, things changing so skorry these days and everybody very quick to forget, newspapers not being read much neither. Well, what they sold there was milk plus something else. They had no licence for selling liquor, but there was no law yet against prodding some of the new veshches which they used to put into the old moloko, so you could peet it with vellocet or synthemesc or drencrom or one or two other veshches which would give you a nice quiet horror-show fifteen minutes admiring Bog and All His Holy Angels and Saints in your left shoe with lights nursing all over your mozg. …

Text from *A Clockwork Orange* by Anthony Burgess

The story begins with the droogs sitting in their favourite hangout, the Korova Milk Bar, and drinking "milk-plus" – a beverage consisting of milk laced with the customer's drug of choice – to prepare for a night of ultra-violence.

Summary from Wikipedia

Alex begins his narrative from the Korova, where the boys sit around drinking.

Summary from SparkNotes.com

Review: What is…

A causal link?

A causal chain?

A script?

The Principle of Minimal Departure?

An event?

A pre-condition?

An effect?

Review: Linking Events

PROBABILISTIC

Occur frequently together (not necessarily because they had to)

Example:

I pour dog food in my dog's bowl.

I pet my dog.

CAUSAL

Occur because of one another

Example:

I pour dog food in my dog's bowl.

My dog eats dog food.

What are procedures?

- A procedure is “a series of **actions** conducted in a certain **order** or manner,” as defined by Oxford
- A more refined definition: “a series of **steps** happening to achieve some **goal**[1]”
  - Why?
- Examples of procedures: instructions (recipes, manuals, navigation info, how-to guide), algorithm, scientific processes, etc.
  - We focus on **instructions**, which is human-centered and task-oriented
- Examples of non-procedures: news articles, novels, descriptions, etc.
  - Those are often narrative: events do not have a specific goal
- The umbrella term is **script**[2]

Why study procedures?

- Procedures are useful
  - Core to task-oriented systems (e.g., dialog agent)
  - Can be used for scaffolding in interactive fiction generation
  - Can be used to understand and analyze narrative texts (how?)
- Procedural events are a suitable scope to study **machine reasoning**
  - Procedural texts are more structured and less complex
  - Good starting point to study planning and reasoning

How to study procedures

- **Extract** procedural knowledge from instructions[1]
  - Usually a structured representation
- A subset of these works specifically focus on **recipes**[2]

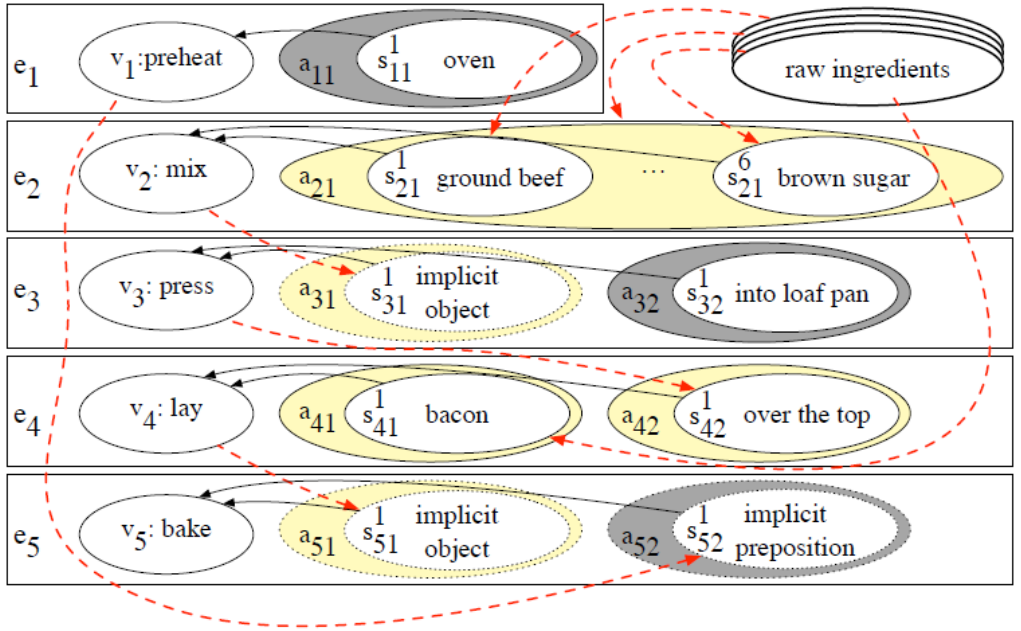
Why might research on procedures focus on recipes?

Amish Meatloaf (http://allrecipes.com/recipe/amish-meatloaf/)

Ingredients
2 pounds ground beef
2 1/2 cups crushed butter-flavored crackers
1 small onion, chopped
2 eggs
3/4 cup ketchup
1/4 cup brown sugar
2 slices bacon

Preheat the oven to 350 degrees F (175 degrees C).
In a medium bowl, mix together ground beef, crushed crackers, onion, eggs, ketchup, and brown sugar until well blended.
Press into a 9x5 inch loaf pan.
Lay the two slices of bacon over the top.
Bake for 1 hour, or until cooked through.

(recipe condensed)



[1] (Paris et al., 2002; Zhang et al., 2012; Maeta et al., 2015) [2] (Kiddon et al., 2015; Bisk et al., 2019)

How to study procedures

- **Reason** about procedural events and serve **downstream** tasks
  - Early text classification (Takechi et al., 2003)
  - Answering how-to questions (Delpech and Saint-Dizier, 2008; Zhang et al. 2020)
  - Tracking entity states (Dalvi et al., 2018; Gupta and Durett, 2019; Tandon et al., 2020)
  - Event relation reasoning (Zhou et al. 2022)
  - Intent classification (Lyu et al. 2021)
  - Application in task-oriented dialogs (Zhang et al. 2020)

This work* answers the questions…

How well can LLMs reason about the steps of a procedure?

How can we combine procedures to create new scripts?

How can procedures help us do intent detection?

How can LLMs expand procedures to show more detailed steps?

* Work by Li “Harry” Zhang and Qing “Veronica” Lyu, and others

This work answers the questions…

How well can LLMs reason about the steps of a procedure?

How can we combine procedures to create new scripts?

How can procedures help us do intent detection?

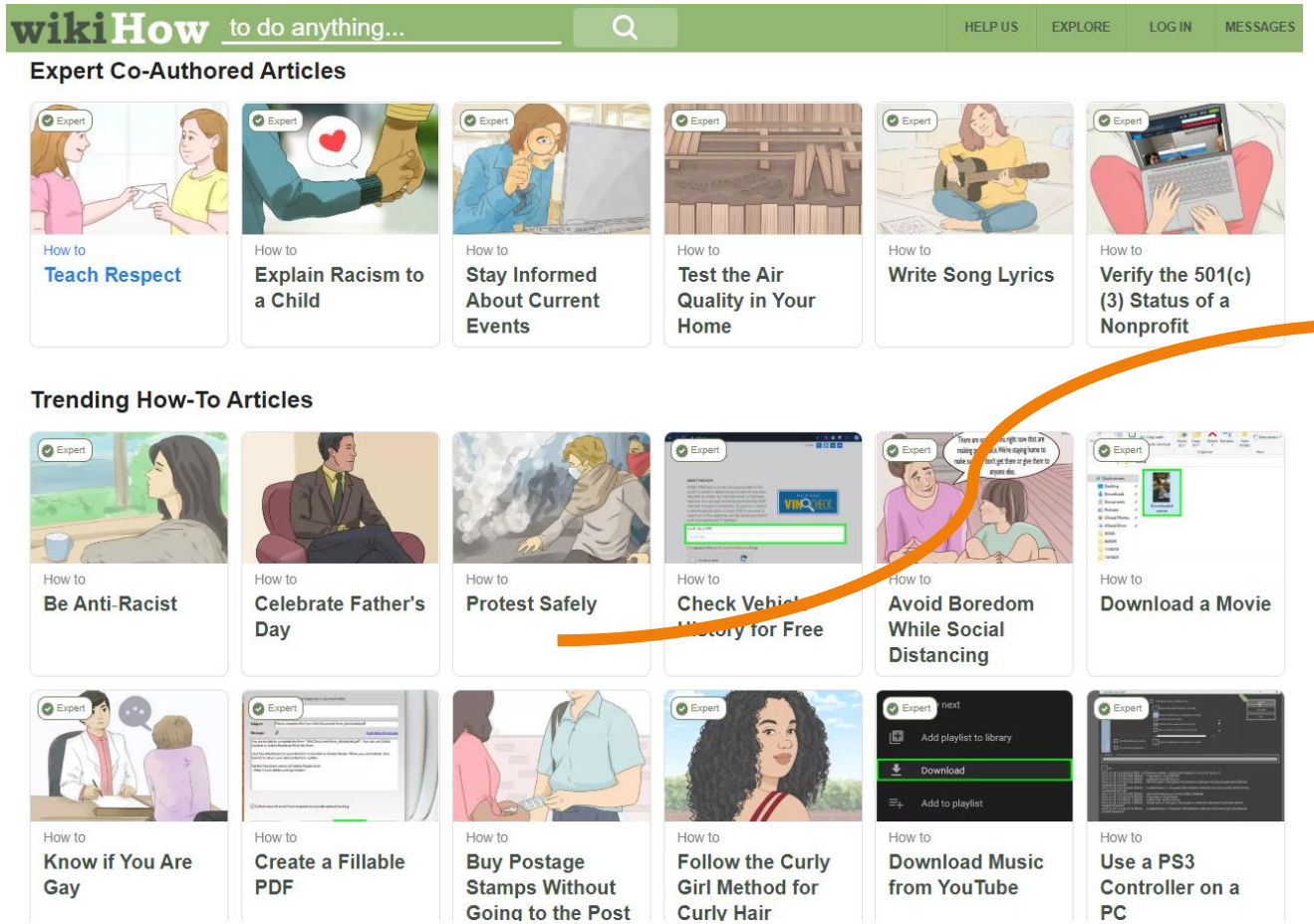
How can LLMs expand procedures to show more detailed steps?

Reasoning about Goal-Step Relations

- Procedure is “a series of **steps** happening to achieve some **goal**”
- First and foremost, a model should understand what steps and goals are
- Examples:
  - (Infer goal) If I preheat the oven to 350 F, what am I trying to do?
  - (Infer step) If I want to get a good GPA, should I study or play PS5?
  - (Infer order) To drive a car, do I first shift to drive or foot down the gas?
- Where can we get lots of procedural data?

Reasoning about Goals, Steps, and Temporal Ordering with WikiHow (Zhang et al., EMNLP 2020)

wikiHow.com



Reasoning about Goals, Steps, and Temporal Ordering with WikiHow (Zhang et al., EMNLP 2020)

Multiple-choice format

- Task #1 Goal Inference: Given a goal, choose the most likely step out of 4 candidates.
  - **Input**: "How to prevent coronavirus"
  - **Choices**: Wash your hands? Wash your cat? Clap your hands? Eat your protein?
- Task #2 Step Inference: Given a step, choose the most likely goal out of 4 candidates.
  - **Input**: "Blink repeatedly."
  - **Choices**: Handle Pepper Spray in Your Eyes? Relax Your Eyes? Draw Eyes? Diagnose Pink Eye?
- Task #3 Step Ordering: Given a goal and two unordered steps, determine which comes first.
  - **Input**:

  Goal: How to Act After Getting Arrested.

  Step (a): Get a lawyer. Step (b): Request bond from the judge

Negative Sampling

- Finding the correct answer is easy; finding some wrong answers is hard
- Semantic similarity-based approach
  - BERT sentence embedding similarity based on KNN
  - Consider full sentence, only verbs, only nouns, etc.
- Example
  - In an article “Protest safely”, a step is “Wear sensible clothes.”
  - Find the sentence embedding of “Wear sensible clothes”, and all other steps
  - Find 3 other steps whose sentence embeddings have the highest cosine similarity with the sentence embedding of “Wear sensible clothes”
  - Result: “Wear sensible clothes,” “buy some clothes”, “put on costumes”, “try old clothing.”

Statistical Cues

- Modern models like BERT can cheat!
- Sometimes models can tell the answer without looking at the question
- Zhang et al. resolve this by randomly reassigning the correct candidate
  - Once they get all 4 candidates, any of them could be the correct one
  - Any model is guaranteed to perform no better than chance given only the candidates

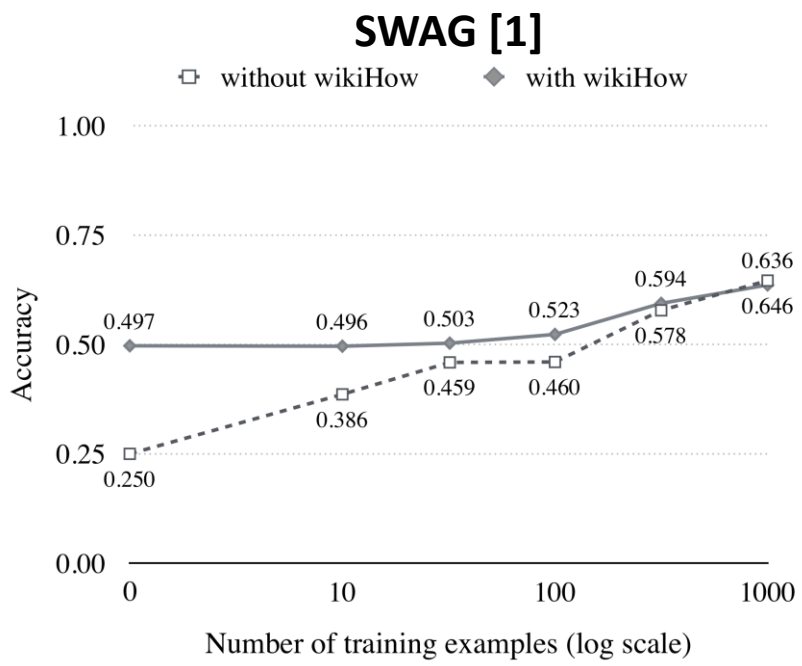
Crowdsourcing Evaluation

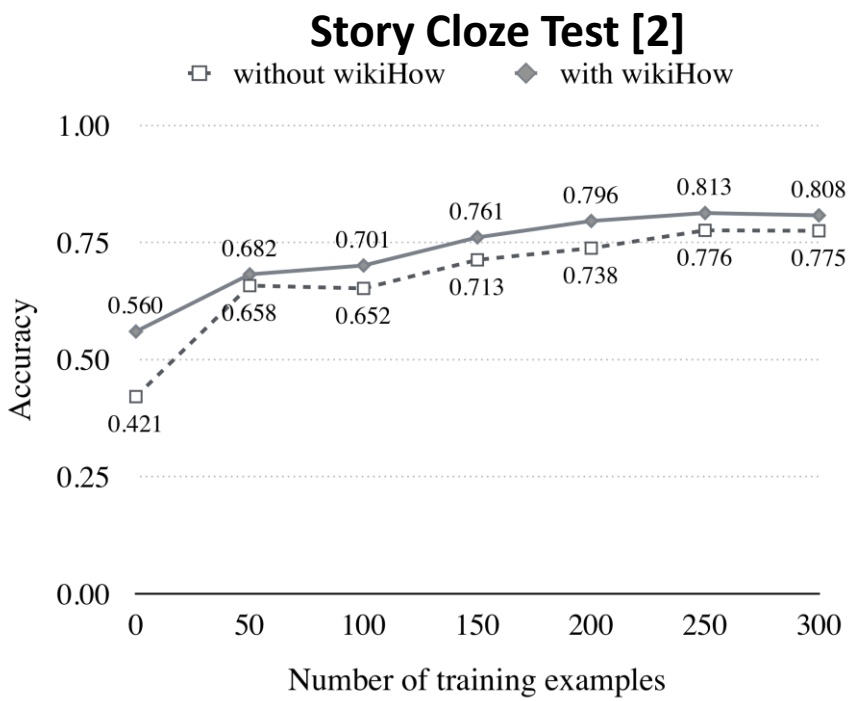
Why is just a fraction of the data crowdsourced?

- We now have a large number of automatically generated examples
- These examples might be noisy, but we still want a clean test set
- For each task, randomly sample thousands of examples to pose to crowd workers
- If all of 3 crowd workers answer correctly based on the gold label, keep the example in the test set
- Use the rest as training and validation sets

How can it be used?

- Large dataset for training or finetuning & testing
- A model trained on their data did well on other datasets (transfer learning)





[1] Rowan Zellers, et al. "SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference." EMNLP 2018. https://aclanthology.org/D18-1009/
[2] Nasrin Mostafazadeh, et al. "A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories." NAACL-HLT 2016. https://aclanthology.org/N16-1098/

This work answers the questions…

How well can LLMs reason about the steps of a procedure?

How can we combine procedures to create new scripts?

How can procedures help us do intent detection?

How can LLMs expand procedures to show more detailed steps?

Motivation

- Models can infer goals, steps, and ordering in **existing procedures**. Can we go one step further?
- **Creating new procedures**:

If you know

how to make an apple pie and how to make a banana cake

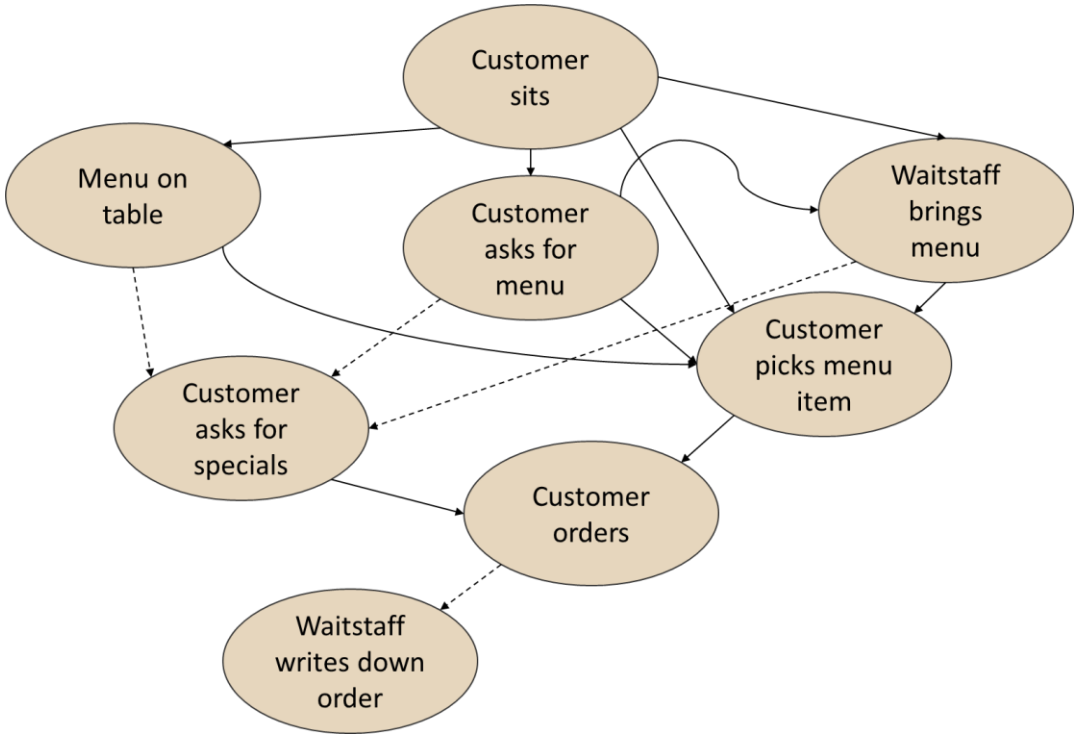
can you infer

how to make a banana pie ?

- This is commonsense knowledge to humans; do language models have it?

The Script Construction Task

- **Script:** a standardized sequence of events
- **Script Construction:** Given the goal, narrate the full script



Goal-Oriented Script Construction (Lyu et al., INLG 2021)

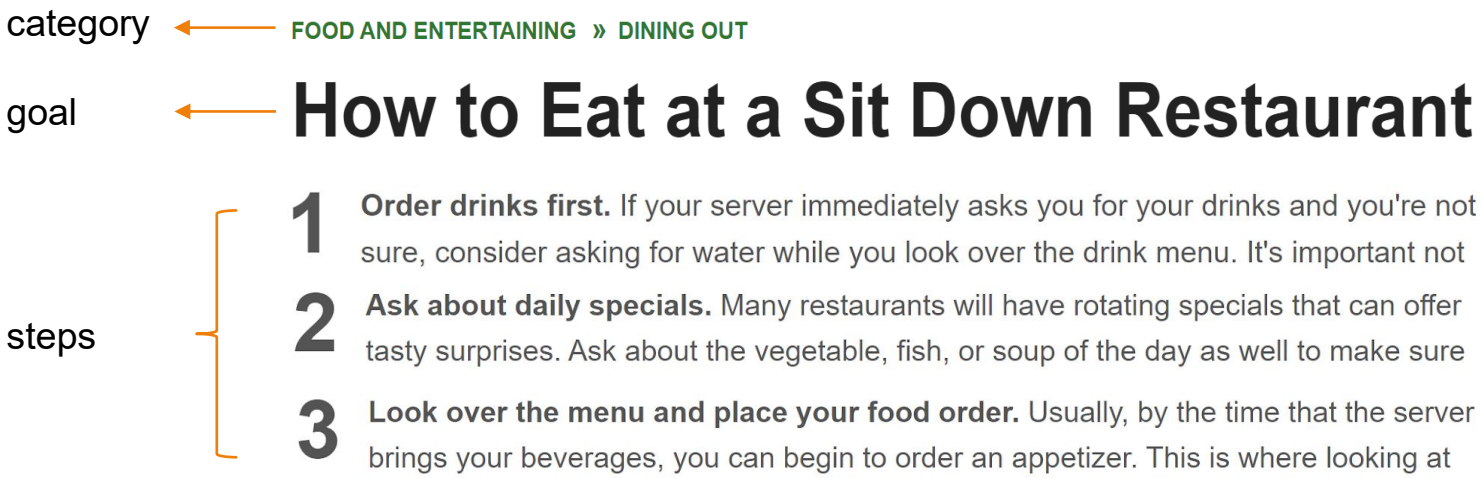
How do we set up the task?

- **Generation** setting
  - Input: a goal
  - Output: a list of steps
  - Evaluation: Perplexity; Human
  - Concerns: might be too difficult; output might be "unfocused"
- **Retrieval** setting
  - Input: a goal, a set of candidate steps
  - Output: a list of steps
  - Evaluation: Accuracy; Recall; Normalized Discounted Cumulative Gain + Kendall's Tau; Human
  - Concerns: where to get a suitable set of candidate steps?

More recent work (at the end of the lecture) addresses this

Building a Dataset from wikiHow

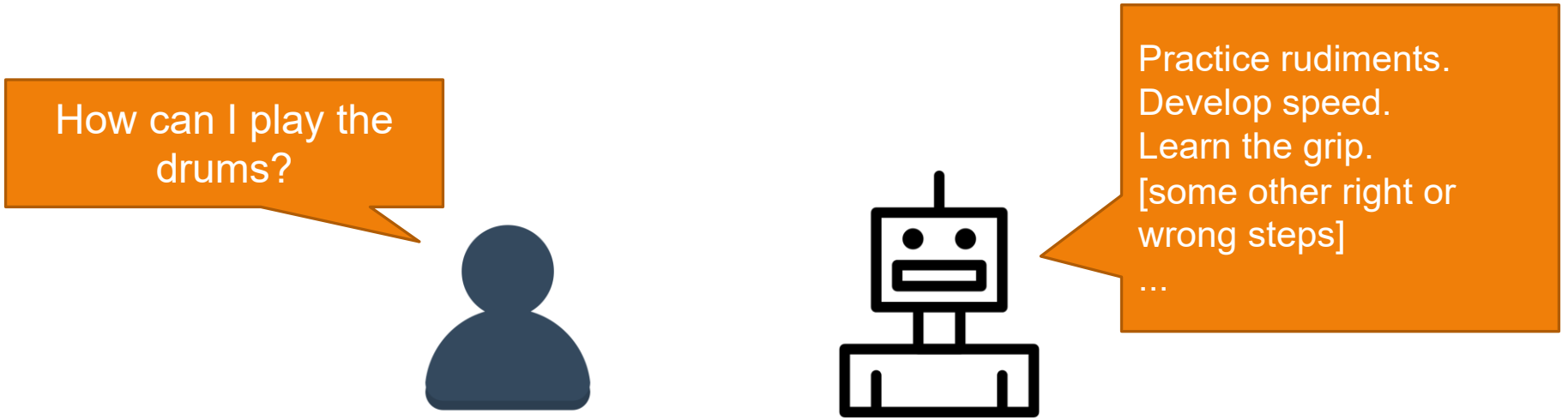
- Extends WikiHow dataset to include **18 languages** (not just English)



- For the Retrieval Setting, the set of candidate steps are from all articles of the same category

How to construct scripts?

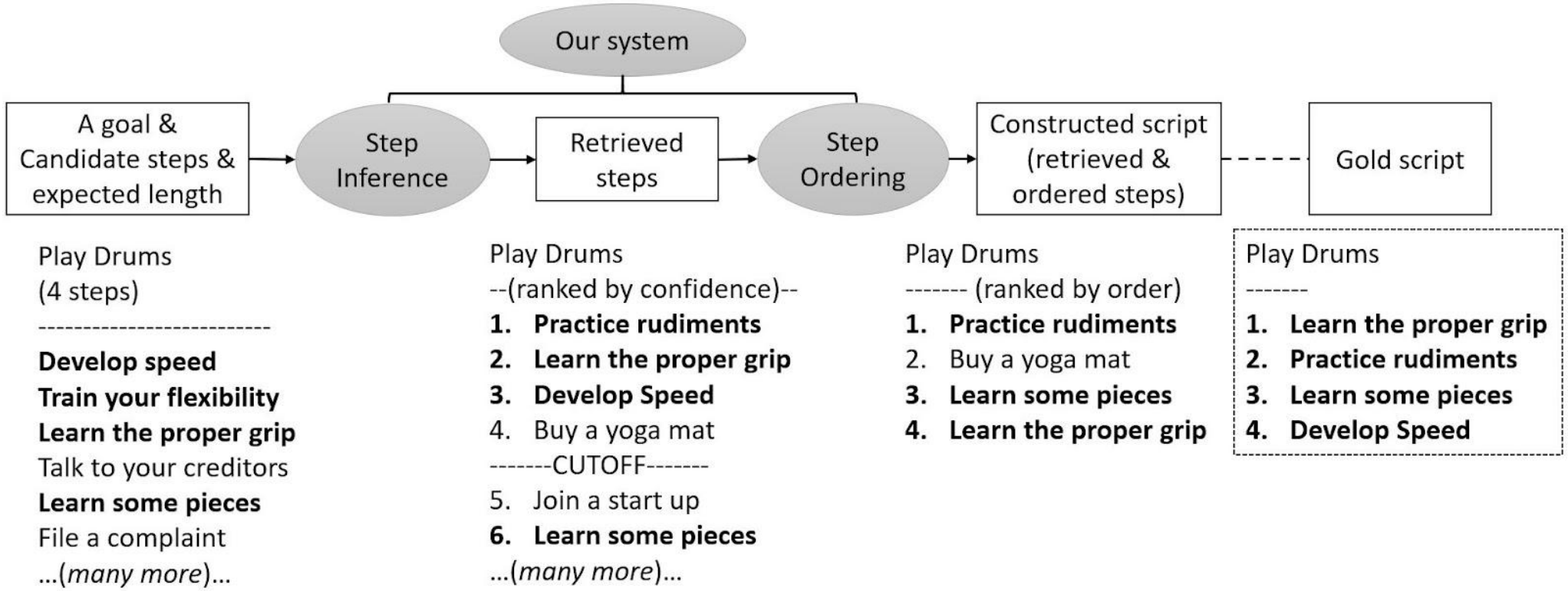
- **Generation-based approach:**
  End-to-end finetuning of a generative language model, Google's Multilingual-T5 (Xue et al., 2021).
  - During training, input is goal + steps
  - During inference, input is goal; output is steps



How to construct scripts?

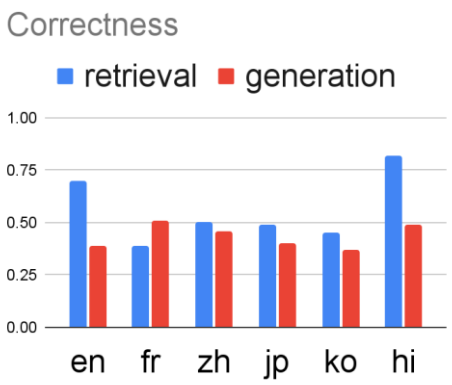
- **Retrieval-based approach:**
  A pipeline using the Step Inference model & Step Ordering models from their previous EMNLP 2020 work

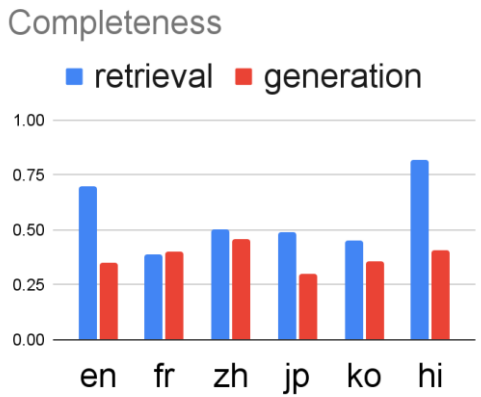
Reasoning about Goals, Steps, and Temporal Ordering with WikiHow

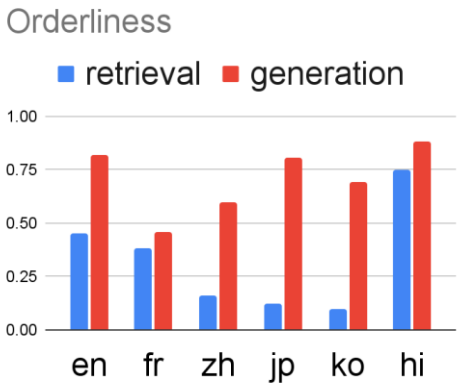


How well can models construct scripts?

- Lyu et al. ask humans to ***edit*** the predicted script by either deleting or moving a step
- They then calculate:
  - *"Correctness"*: len(edited script) / len(predicted script)
  - *"Completeness"*: len(edited script) / len(gold script)
  - *"Orderliness"*: Kendall's Tau of steps in the edited script







How can it be used?

- Construct scripts in new domains:
  - Models trained on wikiHow (**daily activities**) can transfer to the **news** (**political & military**) domain
  - Example: A script of "**Roadside Improvised Explosive Device Attack**"

1. Movement.Transportation
   (Example: "Taliban transported explosives in car to centre place.")
2. Conflict.Attack.DetonateExplode
   (Example: "Someone detonated ied explosive device at ghazni place.")
3. Conflict.Attack
   (Example: "Group attacked convoy using explosive.")
4. Life.Injure
   (Example: "861 was injured by contractor using explosive.")
5. ArtifactExistence.DamageDestroyDisableDismantle.Damage
   (Example: "Someone damaged vehicle.")
6. ArtifactExistence.ManufactureAssemble
   (Example: "Someone manufactured or assembled or produced weapons.")
7. Life.Die
   (Example: "Members died, killed by efps killer.")
8. Justice.TrialHearing
   (Example: "Someone tried someone before dunford court or judge.")

Goal-Oriented Script Construction (Lyu et al., INLG 2021)

This work answers the questions…

How well can LLMs reason about the steps of a procedure?

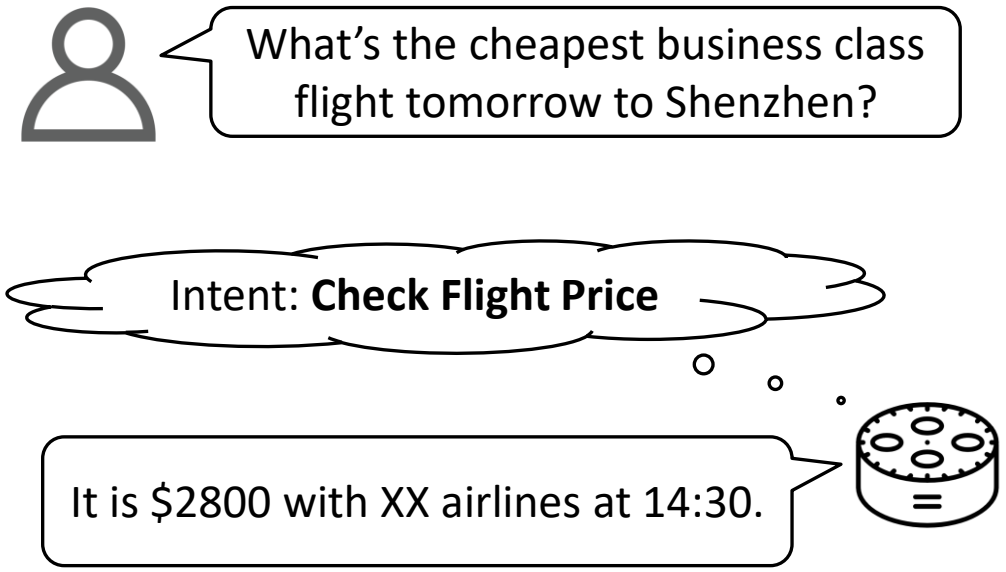
How can we combine procedures to create new scripts?

How can procedures help us do intent detection?

How can LLMs expand procedures to show more detailed steps?

Intent Detection

- Task-oriented dialog systems needs to match an **utterance** to an **intent**, before making informed responses
- Sentence classification task
  - Given an utterance, and some candidate intents
  - Choose the correct intent
  - Evaluated by accuracy

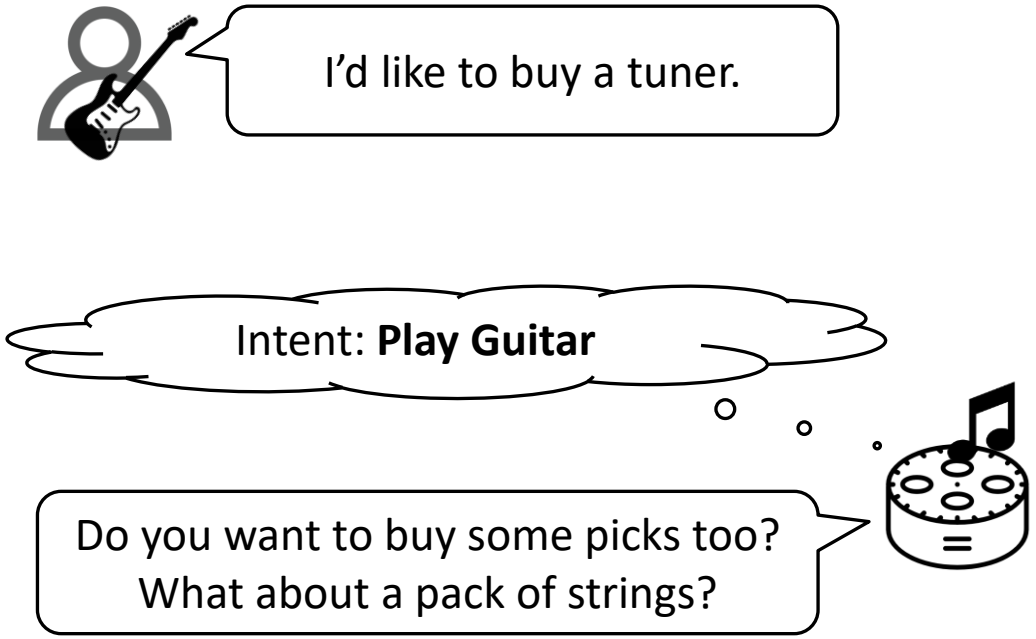


Example from Snips (Coucke et al., 2018)
Utterance: "Find the schedule at Star Theatres."
Candidate intents: Add to Playlist, Rate Book, Book Restaurant,
Get Weather, Play Music, Search Creative Work, **Search Screening Event**

Intent Detection Gets Difficult

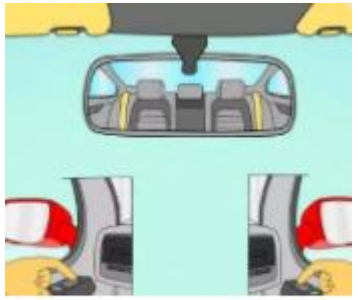
- Some intents are hard to infer
  - Require some world knowledge
  - Utterances and intents might span many domains, sometimes niche ones
  - Multilingual settings, especially low-resource

Image: https://safebooru.org/index.php?page=post&s=view&id=546415



WikiHow articles contain relevant information

- A **goal** (title stripped of “how to”) can approximate an **intent**
- A **step** (step paragraph header) can approximate an **utterance**
- Not a perfect correspondence
  - Some goals are too specific to be an intent
  - Some steps are unlikely utterances
- Still, potentially very strong pre-training data



1 Learn the driving rules for your location.
2 Get your permit.
3 Practice driving.
…

Data: https://github.com/zharry29/wikihow-goal-step

Results

Single-turn English

Schema-guided Dialogue (multi-turn, max 4)

Single-turn Multilingual

Multi-task learning with intent detection & slot filling

English data

	Snips	SGD	FB-en
(Ren and Xue, 2020)	.993	N/A	.993
(Ma et al., 2019)	N/A	.948	N/A
+in-domain (+ID)	.990	.942	.993
(ours) +WH+ID	**.994**	**.951**†	**.995**†
(ours) +WH 0-shot	.713	.787	.445
Chance	.143	.250	.083

Table 2: The accuracy of intent detection on English datasets using RoBERTa. State-of-the-art performances are in bold; † indicates statistically significant improvement from the previous state-of-the-art.

Siamese NN, prev SoTA

Data augmented with back-translation to & from Chinese

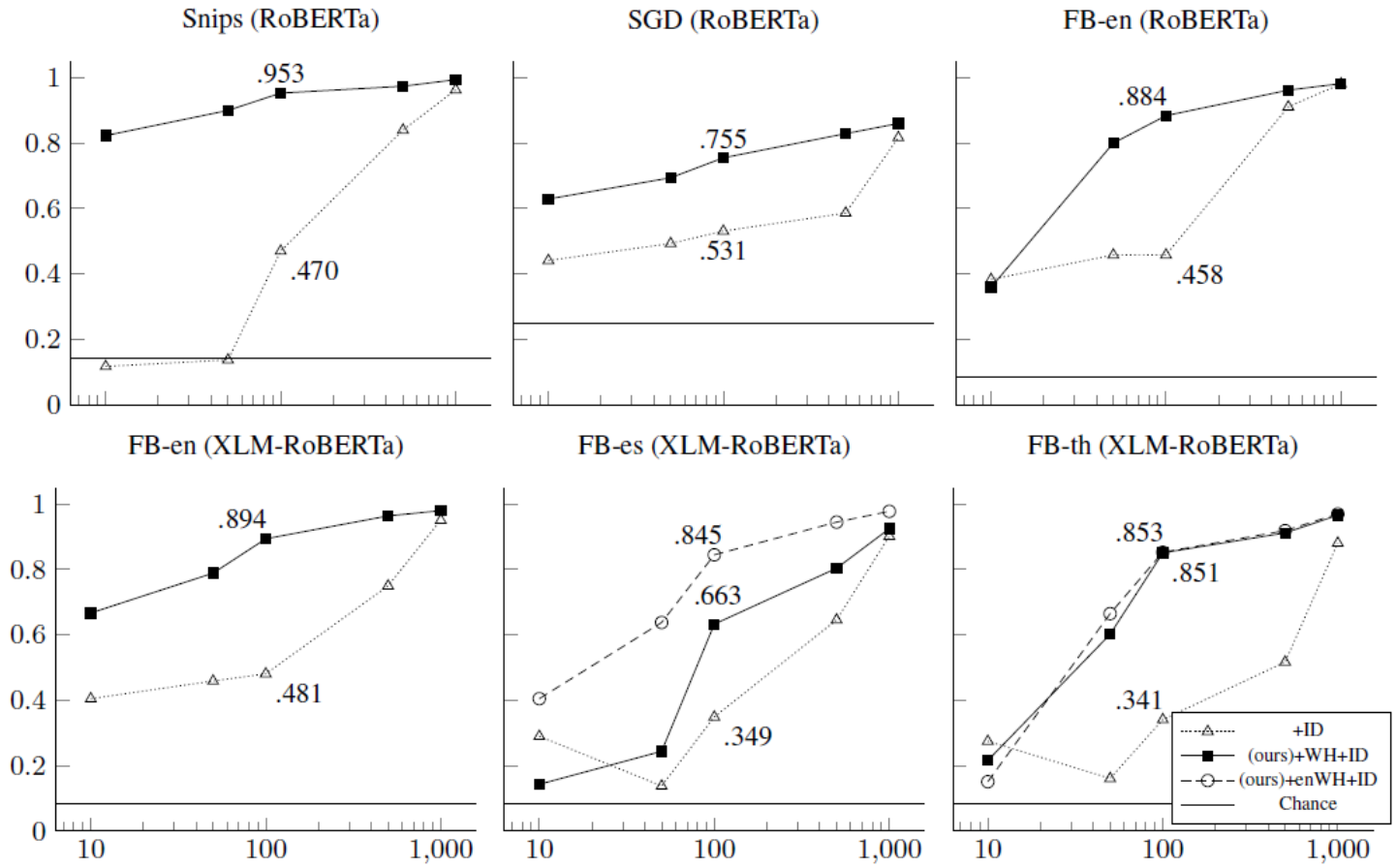
Multilingual data

	FB-en	FB-es	FB-th
(Ren and Xue, 2020)	.993	N/A	N/A
(Zhang et al., 2019)	N/A	.978	.967
+in-domain (+ID)	.993	.986	.962
(ours) +WH+ID	**.995**	.988	.971
(ours) +enWH+ID	**.995**	**.990**†	**.976**†
(ours) +WH 0-shot	.416	.129	.119
(ours) +enWH 0-shot	.416	.288	.124
Chance	.083	.083	.083

Table 3: The accuracy of intent detection on multilingual datasets using XLM-RoBERTa.

Few-Shot Transfer

- Learning curves of models in low-resource settings
- Vanilla transformers (+ID) struggle with low-resource setting



Vertical axis: accuracy of intent detection
Horizontal axis: # in-domain training examples of each task, distorted to log-scale

This work answers the questions…

How well can LLMs reason about the steps of a procedure?

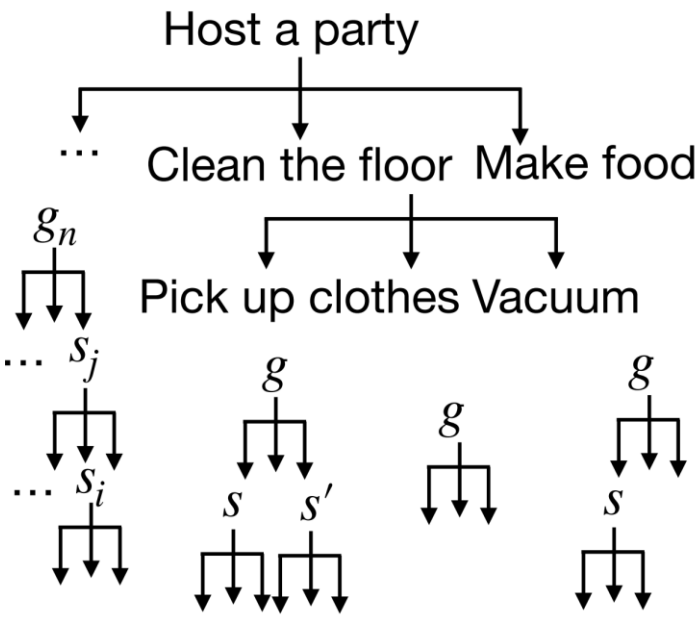
How can we combine procedures to create new scripts?

How can procedures help us do intent detection?

How can LLMs expand procedures to show more detailed steps?

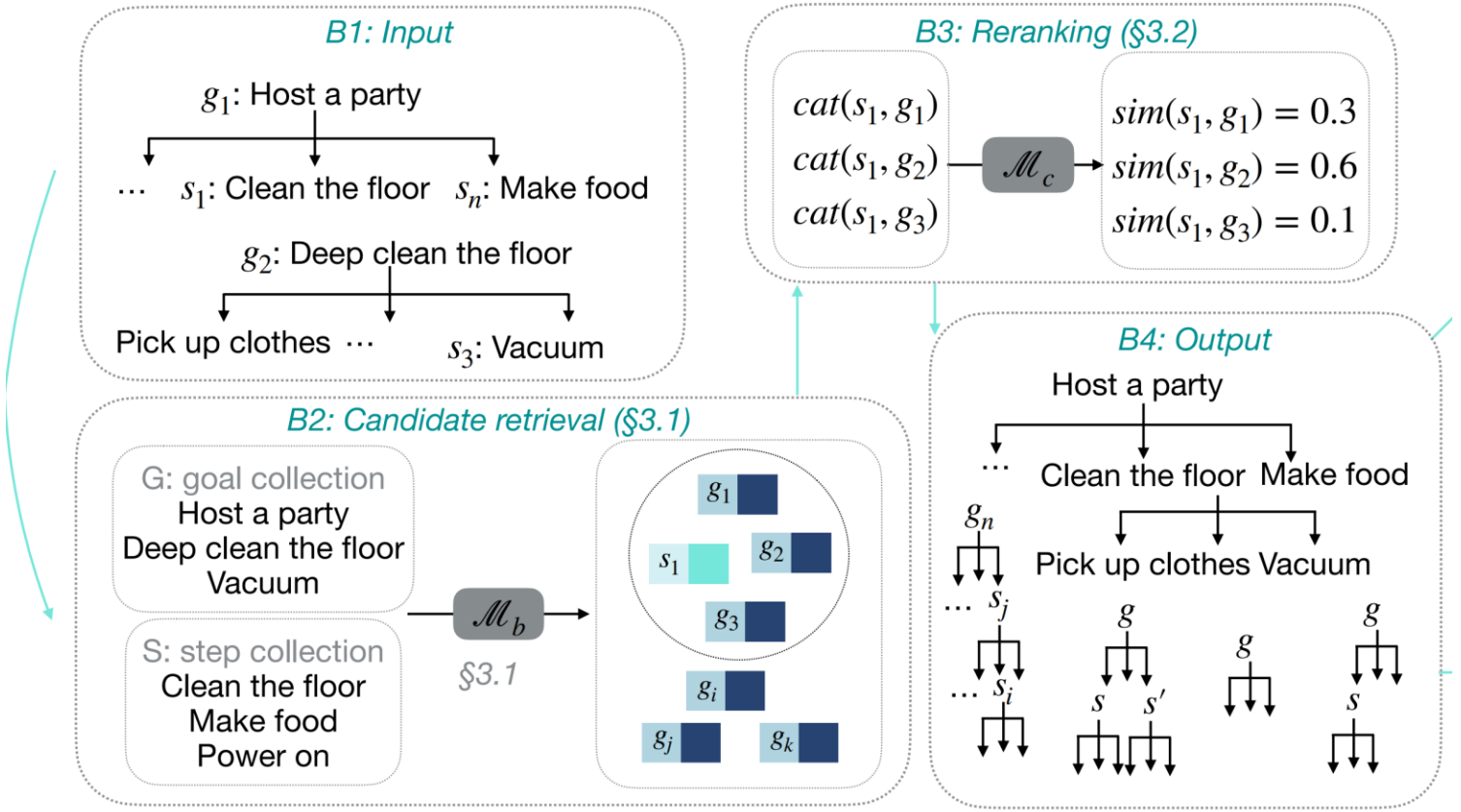
Procedures are Hierarchical

- An event can simultaneously be a **goal** of one procedure, and a **step** in another
- A procedural hierarchy… So what?
  - Can "explain in more details" by expansion
  - Can shed light on event **granularity** (why?)
- How do you build such hierarchy?
  - To "host a party", I need to "clean the floor"; to "clean the floor", I need to do what?



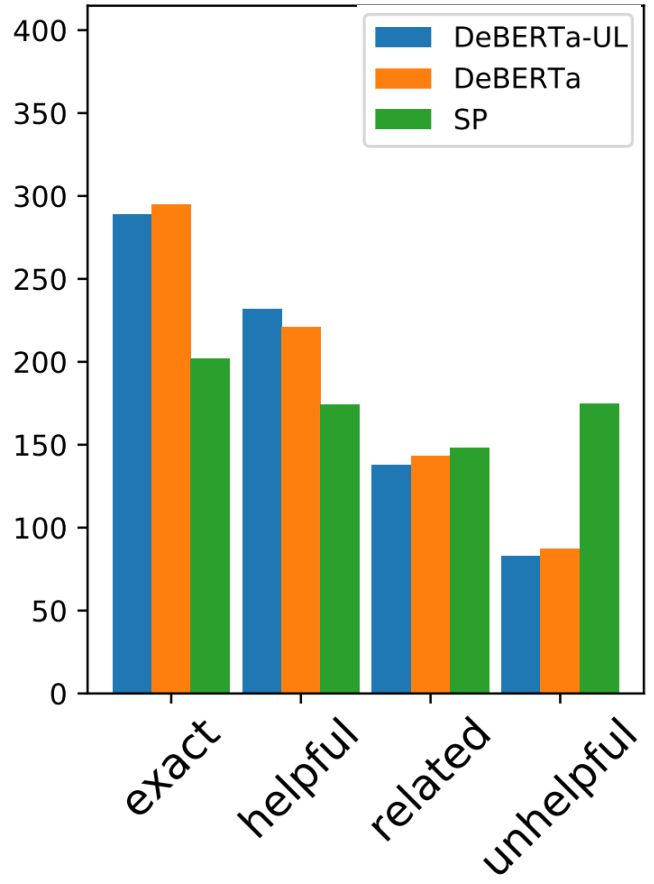
Linking Steps and Goals

- For each procedure, for each **step**, find another procedure whose **goal** has the same meaning
- Effectively a paraphrasing matching or semantic search problem (is it really?)
- Method: retrieve-then-rank



Procedural Hierarchy is Useful

- Crowdworkers rate linked **step**-**procedure** pair, and decide if the instructions of the **procedure** is helpful for doing the **step**
  - About 80% are helpful
- Can be directly applied to a task-oriented dialog system (e.g., Alexa Prize Taskbot)



Event Granularity is Not So Obvious

- Words lower in the hierarchy (supposedly more low level): push, continue, learn, decide, repeat, avoid, finish, move, wait, accord
- Words higher in the hierarchy (supposedly more low level): decorate, knead, season, beat, paint, simmer, dig, sew, build, melt
- They don't seem clustered by granularity

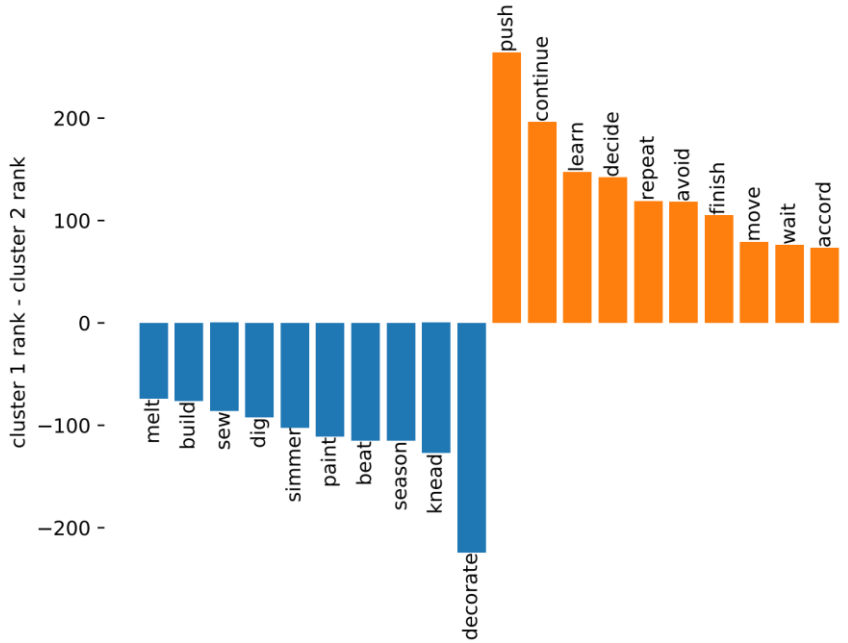
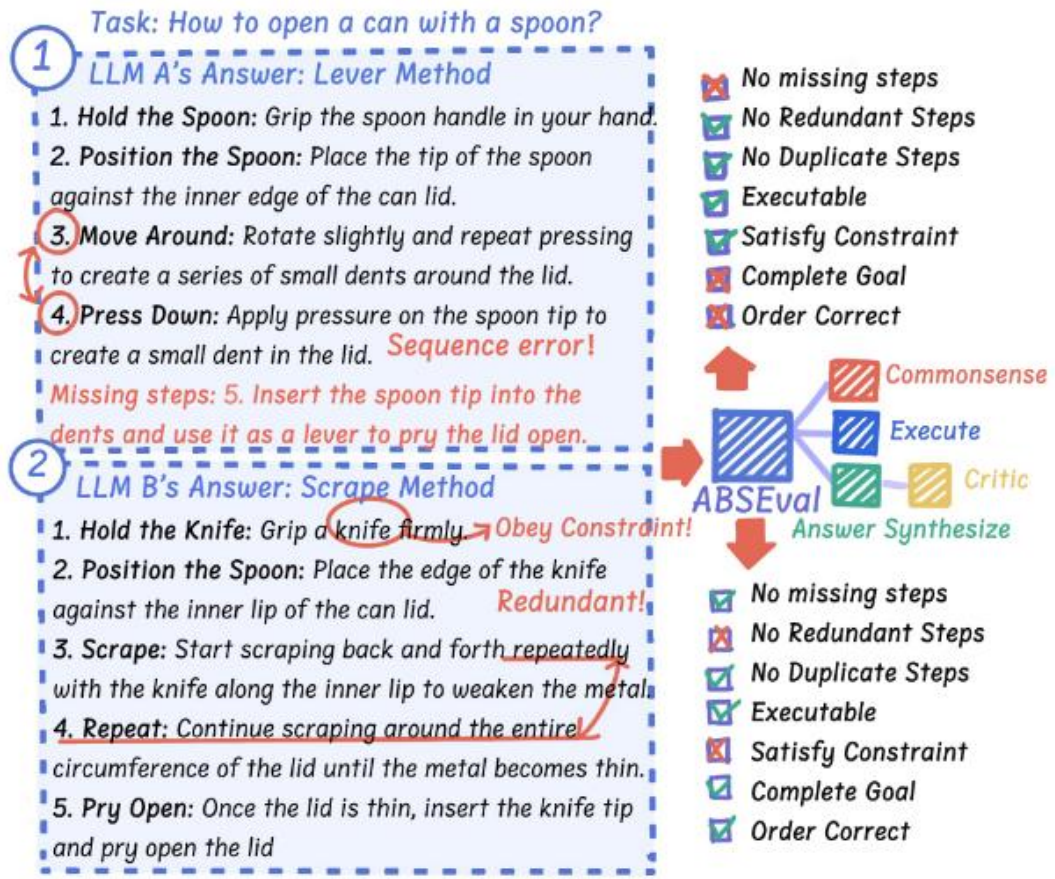


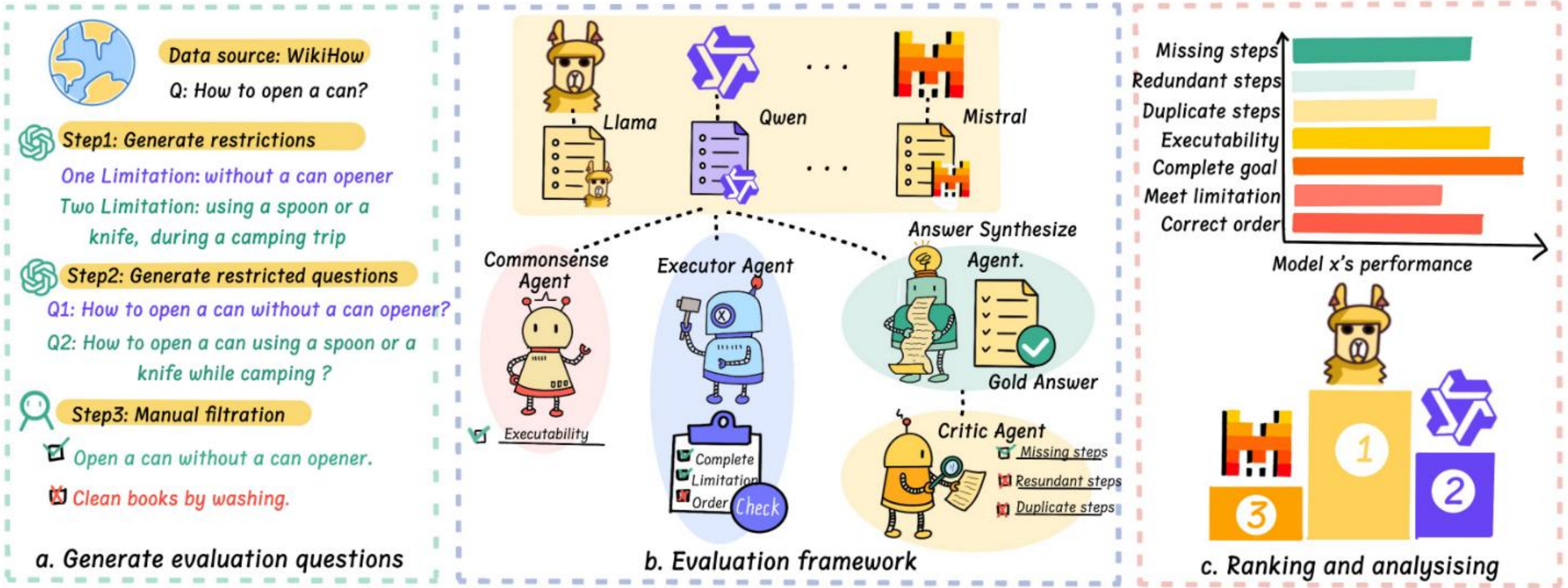
Figure 3: The verbs with largest rank difference in two clusters. The blue bars are words becoming less frequent in cluster 2 (decomposed steps) and the orange bars are words becoming more frequent.

Alternative Evaluation for Procedures: LLMs as Evaluators



ABSEval: An Agent-based Framework for Script Evaluation (Liang et al., EMNLP 2024)

Evaluation Pipeline (ABSEval)



ABSEval: An Agent-based Framework for Script Evaluation (Liang et al., EMNLP 2024)

Evaluation Pipeline Results

Model Name	Rank	Size	NM	NR	ND	EX	SC	CG	OC
Baichuan-Chat	14th	13B	0.029	0.787	0.994	0.833	0.673	0.572	0.632
Baichuan2-Chat	13th	13B	0.139	0.777	0.992	0.813	0.677	0.580	0.604
Vicuna-v1.5	10th	7B	0.044	0.811	0.995	0.876	0.713	0.611	0.696
Vicuna-v1.5	9th	13B	0.074	0.858	0.999	0.888	0.708	0.624	0.720
LLaMa2-chat	11th	7B	0.250	0.728	0.999	0.836	0.661	0.566	0.709
LLaMa2-chat	7th	13B	0.211	0.807	0.999	0.871	0.715	0.622	0.722
LLaMa2-chat	2nd	70B	**0.379**	0.773	0.999	0.886	0.711	0.665	0.727
LLaMa3-instruct	5th	8B	0.103	0.880	**1.000**	0.889	0.758	0.681	0.725
LLaMa3-instruct	1st	70B	0.154	0.894	1.000	0.902	0.755	**0.711**	0.745
Mistral-Instruct-v0.1	15th	7B	0.048	0.703	0.998	0.816	0.671	0.565	0.610
Mistral-Instruct-v0.2	6th	7B	0.220	0.810	1.000	0.889	0.713	0.666	0.718
Mistral-8x7B-Instruct-v0.1	4th	8x7B	0.092	0.888	**0.999**	0.902	0.753	0.685	**0.766**
Qwen-Chat	12th	7B	0.089	0.831	0.996	0.862	0.678	0.564	0.668
Qwen-Chat	8th	14B	0.139	0.878	0.997	0.879	0.719	0.593	0.703
Qwen-Chat	3rd	72B	0.129	**0.913**	0.998	0.900	**0.763**	0.654	0.763
ALL	-	-	0.137	0.824	0.998	0.870	0.712	0.624	0.700

Table 5: The accuracy rate of all evaluation LLMs for different metrics on the MCScript data set. NM: No Missing Steps, NR: No Redundant Steps, ND: No Duplicate Steps, EX: Executable, SC: Satisfy Constraint, CG: Complete Goal, OC: Order Correct.

Agreement with Human Judgements

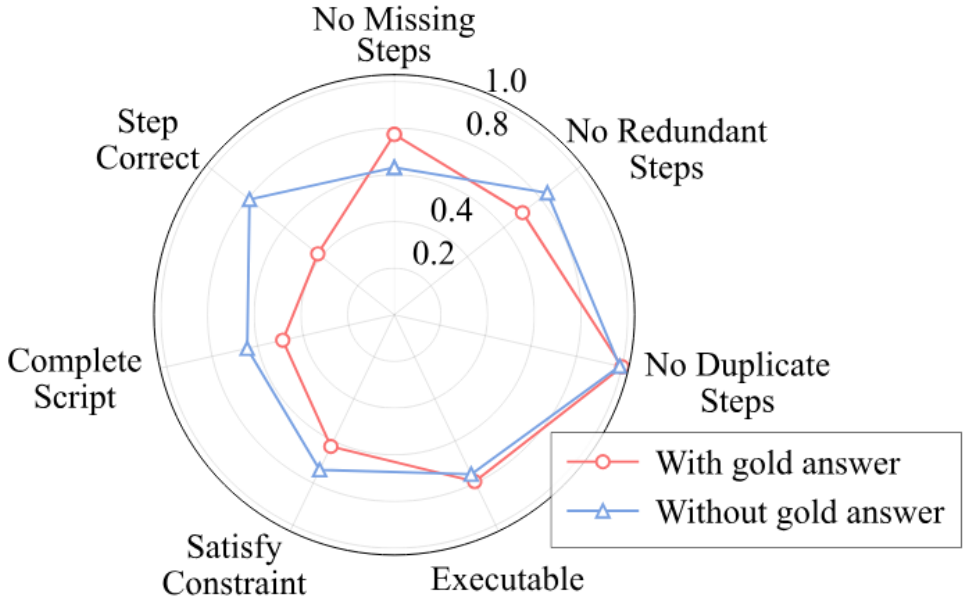


Figure 4: Comparing the consistency of evaluation results with human assessments when directly using LLM for evaluation, with and without providing an answer.

Procedures in IF (Think-Pair-Share)

How might you use procedures in interactive fiction?