MuSR: Testing the Limits of Chain-of-thought with Multistep Soft Reasoning

Zayne Sprague, Xi Ye, Kaj Bostrom, Swarat Chaudhuri, Greg Durrett
Department of Computer Science
The University of Texas in Austin
zayne@utexas.edu

Presented by Junqi Qu

The Problem: Evaluating LLM Reasoning is Hard

- Large Language Models (LLMs), even with Chain-of-Thought (CoT) prompting, still struggle with robust, complex reasoning.
- Evaluating their true capabilities is challenging because existing benchmarks have limitations:
  - **Formal Solvability**: Math reasoning tasks can be offloaded to formal tools. Datasets like RuleTakers can be solved by rule-based systems.
  - **Structural Simplicity**: Commonsense benchmarks like SocialIQA often involve only 1-2 steps of reasoning.
  - **Artificiality**: Many datasets like bAbI and CLUTRR are synthetically crafted and don't reflect the nuance of natural text.
- **The Gap**: There is a need for a benchmark that involves both **sophisticated natural language** and **sophisticated, multi-step reasoning**.

The Solution: MuSR (Multistep Soft Reasoning)

- MuSR is a new dataset for evaluating language models on multistep soft reasoning tasks specified in a natural language narrative.
  - **Novel Neurosymbolic Generation**
  - **Scalable and Renewable**
  - **Realistic & Challenging**
  - **Not Trivial for the Creator**

The MuSR Generation

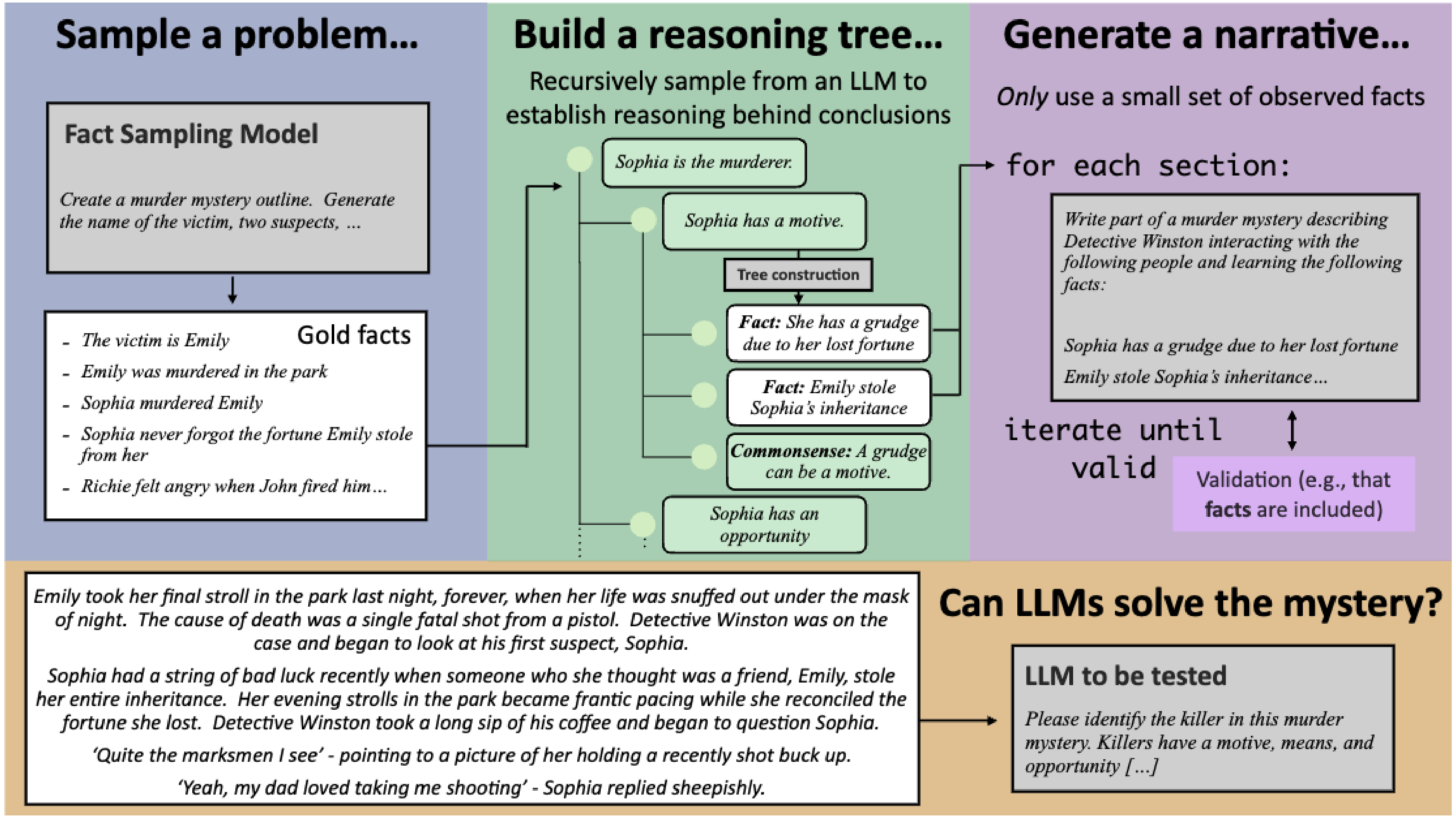

Figure 1: Dataset construction process for MuSR. First, we generate gold facts that are used to deduce the correct answer (the murderer in this case). Then, using an LLM, we create a reasoning tree leading to those deductions from facts in a story combined with commonsense. Finally, we iteratively generate a narrative one chunk at a time using the facts generated in step 2, validating the generations for fact consistency and recall.

MuSR Domains

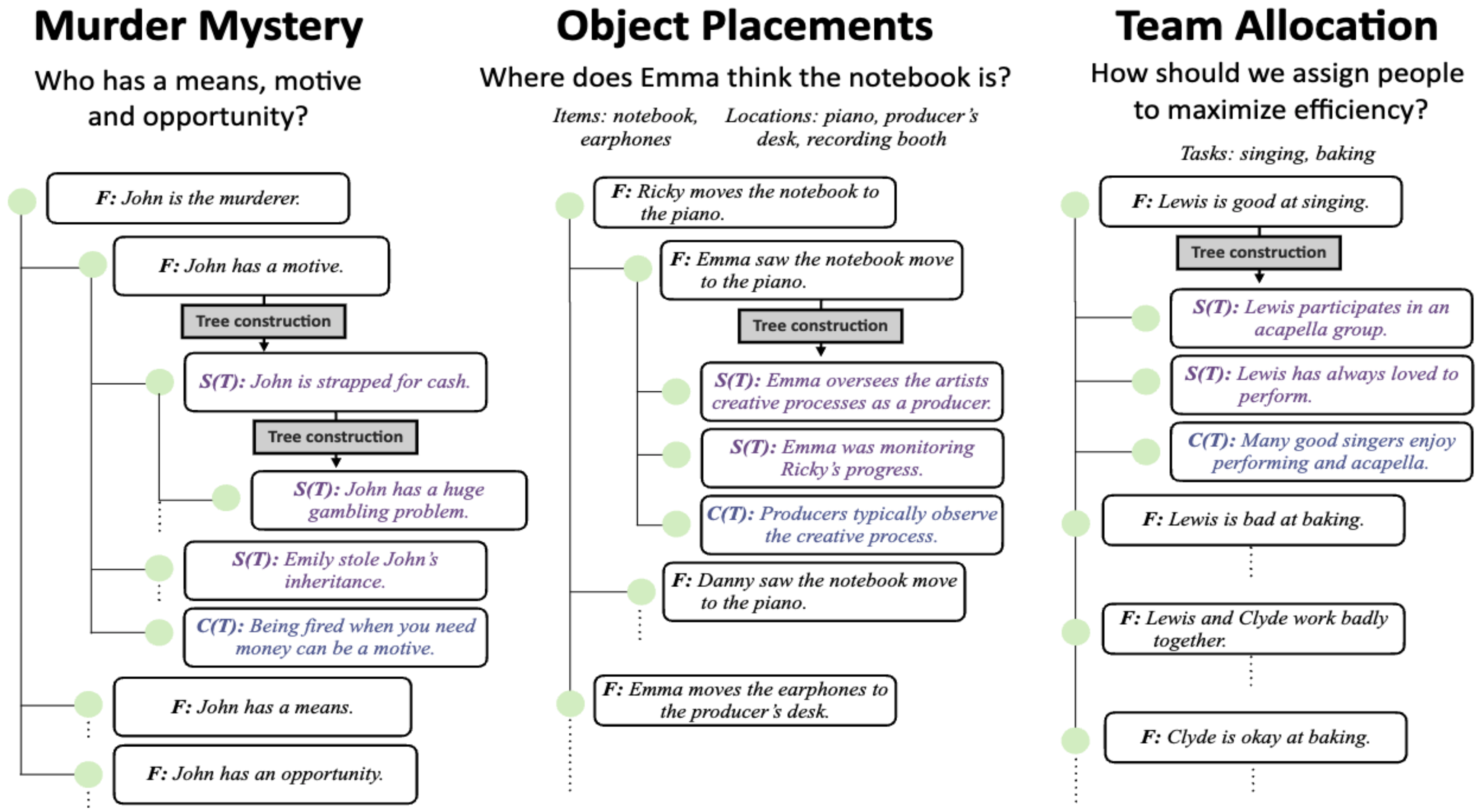

Figure 2: Partial reasoning trees showing gold facts F, story facts $S(T)$, and commonsense facts $C(T)$ for each of our three domains. Dotted lines indicate incomplete trees. Each deduction sampled from an LLM will yield two scenario facts and one commonsense fact in our setup.

Results

Table 5: Scores for LLMs on each domain in MuSR as well as the human evaluation using the CoT+ strategy.

	MM	OP	TA
random	50.0	24.6	33.3
GPT-4	80.4	60.9	68.4
GPT-3.5	61.6	46.9	40.4
Llama2 70b Chat	48.8	42.2	44.8
Llama2 7b Chat	50.8	29.3	36.8
Vicuna 7b v1.5	48.4	29.7	26.4
Vicuna 13b v1.5	50.8	34.4	32.0
Vicuna 33b v1.3	49.6	31.2	30.0
Human Eval	94.1	95.0	100.0

Table 7: Evaluations of different popular prompting strategies for GPT-3.5 and GPT-4, our strongest models. "Regular" supplies only the context and question. "CoT" asks the model to think step-by-step. "CoT+" includes a textual description of the reasoning strategy, and "1-Shot CoT+" includes a solved example. "Few-Shot CoT+" extends "1-Shot CoT+" with 3 examples (3 examples hits the token limit for GPT-4)

	Murder Mystery		Object Placements		Team Allocation	
	GPT-3.5	GPT-4	GPT-3.5	GPT-4	GPT-3.5	GPT-4
Regular	59.2	64.8	44.5	43.0	41.2	64.0
CoT	56.0	65.6	48.4	41.8	46.4	64.4
CoT+	61.6	80.4	46.9	60.9	40.4	68.4
1-Shot CoT+	70.0	86.0	56.2	72.3	50.4	88.4
Few-Shot CoT+	68.4	84.8	58.2	71.5	78.8	89.6

Ablation studies

Table 4: Variations of our dataset creation process. We compare against a simple one-shot prompting approach and an approach using seed facts G to add diversity, which produce simple and poor-quality narratives. We then ablate chaptering and tree validators, showing that these lower length, fact recall in the narrative, and accuracy; the latter usually indicates inconsistent narratives.

	Murder Mysteries				Object Placements				Team Allocation			
Ablation	Len	Div	R	Acc	Len	Div	R	Acc	Len	Div	R	Acc
Prompt Only	280	0.30	-	76	200	0.26	-	64	172	0.34	-	80
Diversity Sampling	422	0.25	-	60	404	0.24	-	39	448	0.26	-	84
MuSR − chapt − validators	428	0.24	67	60	380	0.27	83	78	-	-	-	-
MuSR − validators	924	0.24	93	60	793	0.25	82	65	-	-	-	-
MuSR	900	0.25	95	84	777	0.25	87	58	503	0.25	81	68