NARRATIVE-OF-THOUGHT: Improving Temporal Reasoning of Large Language Models via Recounted Narratives

Xinliang Frederick Zhang[1], Nick Beauchamp[2], and Lu Wang[1]

[1]University of Michigan,
[2]Northeastern University

Temporal Reasoning & Temporal Graph Generation

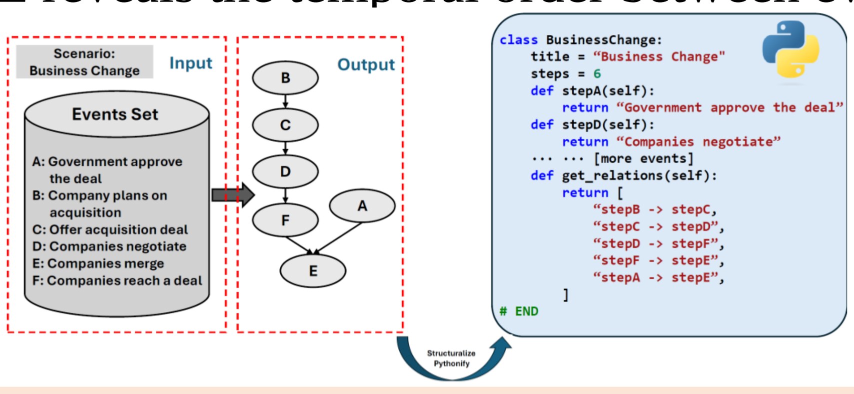
Preamble: Temporal reasoning is essential for humans to perceive the world, understand daily communications, and interpret the temporal aspects of experiences.

Background:

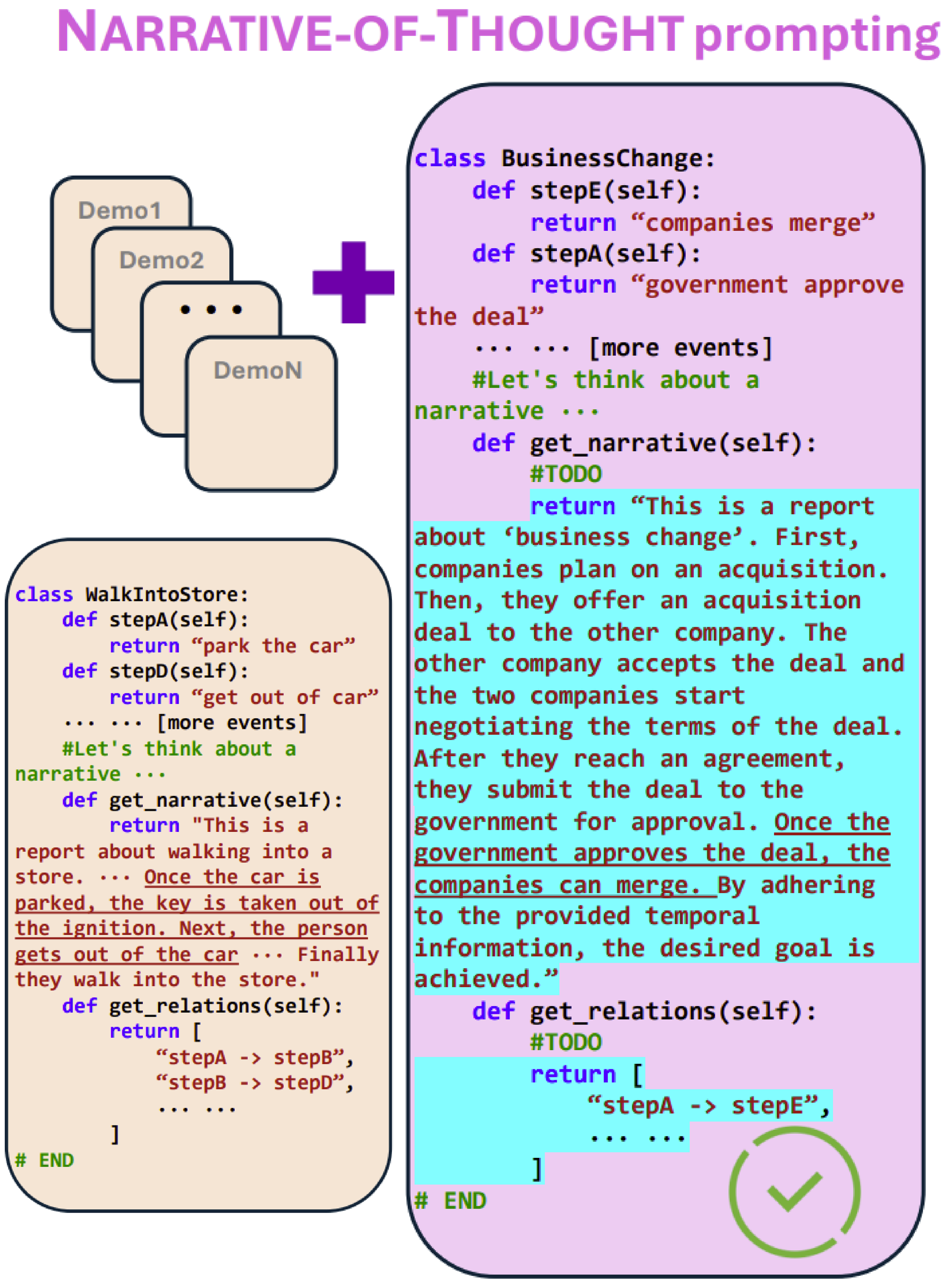
- The advent of LLMs has gathered substantial attention to reasoning, while few LLMs exist to handle temporal reasoning well.
- This reasoning task is inherently complex, mingled with implicit logical inference and the necessity for profound world knowledge.
- Existing research mainly focuses on a simple relation extraction task OR a perplexing commonsense understanding task.

Our objective: Uncover and improve the inherent, global temporal reasoning capabilities of LLMs.

TGG formulation: Given a high-level goal T (e.g., business change) and a set of events V, the objective is to produce a temporal graph G(V, E) where a directed edge in E reveals the temporal order between events.



Method: Narrative-of-Thought (NoT)



[TEXT]: Generations by language models (LMs).
Note: Python class and instructions simplified.

NoT Overview:

1. Given a scenario and a set of events, NoT first converts the input into a Python class.
2. NoT guides LLMs to produce a temporally grounded narrative by arranging events in the correct temporal order, leveraging LLMs' intrinsic temporal knowledge.
3. Based on the recounted temporal relations articulated in the narrative, LLMs are instructed to sort events into a temporal graph.

* We further improve NoT by introducing high-quality reference narratives as part of few-shot demonstrations.

Prompt Design:

Narrative Prompt

# Let's think of a narrative to link aforementioned events in the correct temporal order.
def get_narrative(self):
# TODO

Temporal Graph Prompt

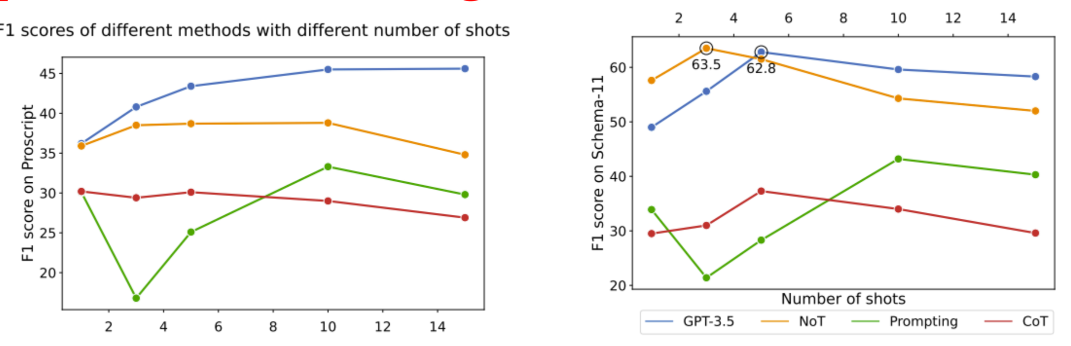
def get_relations(self):
# TODO
# END

Results & Analyses

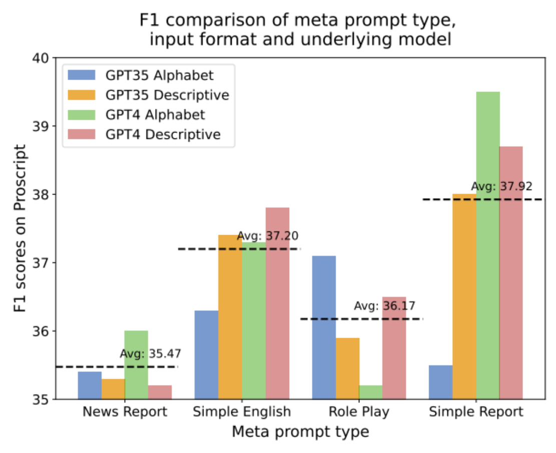
Method	Proscript				Schema-11				WikiHow Script				Avg.	
	F1↑	GED↓	$k(\mathcal{G})$	Cons.↑	F1↑	GED↓	$k(\mathcal{G})$	Cons.↑	F1↑	GED↓	$k(\mathcal{G})$	Cons.↑	F1↑	GED↓
Baselines														
Random	14.0	1.47	1.00	7.8	19.4	3.91	1.00	7.8	14.2	0.06	1.00	8.8	15.9	1.81
GPT-3.5 (0-shot)*	18.4	2.25	1.06	38.6	30.1	4.48	1.27	30.2	17.2	2.80	1.11	40.8	21.9	3.18
GPT-3.5	43.4	1.71	1.07	38.8	62.8	3.30	1.36	50.2	31.0	1.58	1.10	35.4	45.7	2.20
GPT-4	63.9	1.64	1.02	61.4	44.1	7.97	0.64	46.3	43.0	1.71	1.04	48.5	50.3	3.77
LLAMA3-8B (AI@Meta, 2024)														
Standard Prompting	25.1	2.39	1.18	19.9	28.3	4.42	1.24	19.9	20.6	1.17	1.07	21.2	24.7	2.66
Chain-of-Thought	30.1	2.06	**1.00**	23.3	37.3	5.79	0.85	23.5	22.6	**0.99**	**1.02**	**24.3**	30.0	2.95
NoT (no reference)	35.5	1.88	**1.00**	25.3	52.6	**3.18**	1.12	35.0	25.4	**0.99**	**1.02**	20.9	37.8	**2.02**
NoT (alphabetical meta)	**39.5**	1.87	1.01	**28.8**	59.0	3.72	1.12	39.1	26.3	1.01	1.03	22.5	41.6	2.20
NoT (descriptive meta)	38.7	**1.86**	1.01	28.4	**61.5**	3.57	**1.09**	**45.6**	**26.5**	1.04	1.03	22.3	**42.2**	2.16

- Small LLMs struggle with temporal reasoning even with few-shot examples.
- CoT is also ineffective at temporal reasoning, in line with existing findings.
- GPT-4 sometimes falls off the throne due to additional alignment, when answering sensitive queries.
- NoT is a powerful tool to assist small LLMs to catch up with or even surpass GPT-3.5, and presents strong compatibility with various base LLMs. The average F1 improvements are between 16%-71%.
- Temporally grounded narratives are significant in improving LLMs' temporal reasoning process.
- AI systems are far from mastering temporal reasoning, trailing the human baseline by 30 F1 points.

Analysis 1: Does the number of shots matter?
Ans: The performance generally reaches its peak around the range of **5-10 shots.**



Analysis 2: What characteristics define effective reference narratives?
Ans: We identify three key characteristics: **conciseness, simplicity and factuality.**



Analysis 3: How faithful is the temporal graph to intermediate narratives?
Ans to Analysis 3: We find a medium-to-high **self-faithfulness of 72.8%** where the generated narrative and the temporal graph is **aligned** in terms of the temporal order of events.

Code: github.com/launchnlp/NoT

Contact: xlfzhang@umich.edu

Paper

Codebase & Data

EMNLP 2024