# Choice-75: A Dataset on Decision Branching in Script Learning

Zhaoyi Joey Hou, Li Zhang, Chris Callison-Burch





# agenda.

- Problem Statement
- Dataset
- Experiment & Analysis
- Takeaway

#### problem statement.

- Context
  - Script Learning & Event Centric Reasoning (the study of events)
  - Decision Making (daily scenarios)
- "Decision Branching"
  - Non-linear
  - Subtle differences among choices requires commonsense

Goal purchase a plane ticket to see a desert abroad	Option 1 purchase a plane ticket to a major city and take a train to the desert Option 2 purchase a plane ticket to a small city but right next to the desert
Scenarios & Choices	
Scenario 1 [easy] (the person) finds no train the major city to desert of	True: Option 2 Pred: Option 2
Scenario 2 [medium] (the person) has a long-tin living in that major city	ne friend
Scenario 3 [hard] (the person) hates connect	ting flights True: Option 1 Pred: Option 2
Scenario 4 [N/A] (the person) really looks f the first time ever in des	orward to ert True: Either Pred: Option 2
Scenario 5 [easy] [User Profile] Interests: - Enjoy visiting metropolis Financial situation:	na on travel

### problem statement.

- Formally...

- Task: [Goal, Option1, Option 2, Scenario] -> [Optimal Choice]
- "Scenario"
  - verb phrase
  - user profile

<i>Goal</i> purchase a plane ticket to see a desert abroad	Option 1       purchase a plane ticket to a major       city and take a train to the desert       Option 2       purchase a plane ticket to a small       city but right next to the desert
Scenarios & Choices	
Scenario 1 [easy] (the person) finds no train the major city to desert a	route from t that time True: Option 2 Pred: Option 2
Scenario 2 [medium] (the person) has a long-tin living in that major city	True: Option 1 Pred: Option 1
Scenario 3 [hard] (the person) hates connect	ting flights True: Option 1 Pred: Option 2
Scenario 4 [N/A] (the person) really looks fi the first time ever in dese	Drward to Pred: Option 2
Scenario 5 [easy] [User Profile] Interests: - Enjoy visiting metropolis Financial situation: - Comfortable with spendi Occupation: police officer - Hobbies: photography; G	ng on travel

- Structure
- Data Collection
- Difficulty Level
- Human-in-the-loop Hard Scenarios Collection
- Annotator Agreement

- Structure
  - Goal Options
  - Scenario(s)
    - Optimal Choice (#1, #2, either)
    - Difficulty Level (easy, medium, hard, either)

Scenarios & Choices         Scenario 1 [easy] (the person) finds no train route from the major city to desert at that time       True: Option 2 Pred: Option 2         Scenario 2 [medium] (the person) has a long-time friend living in that major city       True: Option 1 Pred: Option 1         Scenario 3 [hard] (the person) hates connecting flights       True: Option 1 Pred: Option 2         Scenario 4 [N/A] (the person) really looks forward to the first time ever in desert       True: Either Pred: Option 2	jor ert nall ert
Scenario 1 [easy]         (the person) finds no train route from the major city to desert at that time         Scenario 2 [medium]         (the person) has a long-time friend living in that major city         Scenario 3 [hard]         (the person) hates connecting flights         Scenario 4 [N/A]         (the person) really looks forward to the first time ever in desert	
Scenario 2 [medium] (the person) has a long-time friend living in that major city       True: Option 1 Pred: Option 1         Scenario 3 [hard] (the person) hates connecting flights       True: Option 1 Pred: Option 2         Scenario 4 [N/A] (the person) really looks forward to the first time ever in desert       True: Either Pred: Option 2	Ø
Scenario 3 [hard] (the person) hates connecting flights Scenario 4 [N/A] (the person) really looks forward to the first time ever in desert True: Option 1 Pred: Option 2 True: Either Pred: Option 2	S
Scenario 4 [N/A] (the person) really looks forward to the first time ever in desert True: Either Pred: Option 2	8
	×
Scenario 5 [easy] [User Profile] Interests: - Enjoy visiting metropolis Financial situation: - Comfortable with spending on travel Occupation: police officer - Hobbies: photography; Gender: male 	8

- Data Collection
  - Goal: sampled 75 "goals" from *ProScript* [1]
  - **Options**: written by a one researcher and verified by two researchers
  - Scenarios: human-authored + machine generated (more details later)
  - **Difficulty Level & Optimal Choice**: annotated by one researcher and verified by two other researchers



- Difficulty Level
  - Based on number of reason steps
  - *Either*: no obvious differences between two options

*Goal*: find out the library's hours *Option 1*: call the library *Option 2*: search online for the library's hours

*Easy Scenario:* have no internet connection *Choice*: Option 1

*Medium Scenario:* have special requests about the book *Choice*: Option 1

Medium Scenario (User Profile): Name: Doe; Interests: American history Special circumstances: has a bad sore throat ... (more details omitted) Choice: Option 2

*Hard Scenario:* is 3 am in the morning *Choice*: Option 2

Table 2: Different levels in the *library hours* case

- Human-in-the-loop Hard Scenarios Collection [2]
  - a) Identify hard examples (rare!)
  - b) Over-generation with few-shot (hard examples) prompting
  - c) Manually filter valid data points

Format	Easy	Medium	Hard	Either
Verb Phrase (Manual)	65	76	36	65
Verb Phrase (Machine)	46	41	22	50
User Profile	53	76	17	73
All	164	193	75	188

Table 1: Counts of scenario in Choice-75.

- Human-in-the-loop Hard Scenarios Collection
  - Verb Phrase machine generated



Figure 2: Hard scenario generation (verb phrase)

- Human-in-the-loop Hard Scenarios Collection
  - User Profile machine generated



Figure 3: Hard scenario generation (user profile)

- Annotator Agreement
  - Subset: 290 [scenarios optimal choices] pair
  - Participants: 3 for each
  - Fleiss' kappa: 0.59 (moderate to substantial agreement)

- Experiment Setup
- Results & Analysis

- Experiment Setup
  - Model(s): text-davinci-003, gpt-3.5-turbo (ChatGPT)
  - Prompt(s): naive prompt, story prompt [zero-shot]

#### Naive Prompt

[Goal]: {step goal} [Option 1]: {option 1} [Option 2]: {option 2}

[Scenario]: {scenario}

[Question]: Given the Scenario, which option above is the better choice in order to achieve the Goal?

1) Option 1, 2) Option 2, 3) Either one, since they have similar effect when it comes to the goal

#### Story Prompt

A person Doe needs to {step goal}. Now there are two options for Doe: we can either {option 1} (Option 1) or {option 2} (Option 2). Suppose Doe {scenario}.

[Question]: Given the Scenario, which option above is the better choice in order to achieve the Goal? 1) Option 1, 2) Option 2, 3) Either one, since they have similar effect when it comes to the goal

[Answer]:

[Answer]:

- Results & Analysis
  - Difficulty Level ~ Model's Performance
  - Story prompt is generally better

Cuoun Duomant		All		Bi	Binary Easy		asy	Medium		Hard		Either	
Group Prompt	003	Turbo	003	Turbo	003	Turbo	003	Turbo	003	Turbo	003	Turbo	
Verb Phrase	naive	0.60	0.63	0.81	0.82	0.91	0.92	0.83	0.80	0.58	0.67	0.05	0.14
(Manual)	story	0.63	0.64	0.86	0.81	0.95	0.88	0.87	0.81	0.69	0.69	0.02	0.18
Verb Phrase	naive	0.56	0.56	0.77	0.80	0.79	0.79	0.77	0.85	0.69	0.75	0.21	0.15
(Machine)	story	0.55	0.55	0.79	0.80	0.79	0.82	0.85	0.81	0.69	0.75	0.15	0.13
User Drofle	naive	0.61	0.59	0.72	0.69	0.78	0.73	0.73	0.69	0.47	0.60	0.40	0.40
User Frome	story	0.50	0.60	0.57	0.73	0.58	0.76	0.60	0.74	0.40	0.60	0.37	0.34
Average		0.57	0.60	0.75	0.77	0.80	0.82	0.77	0.78	0.59	0.68	0.20	0.22

Table 3: Prediction accuracy by difficulty levels. **Binary**: overall performance on binary classification (i.e. Option 1 or Option 2); **All**: overall performance on three-class classification.

- Results & Analysis
  - Human Performance
    - 290 scenarios
    - 2 participants each

Format	Easy	Medium	Hard	Either
Verb Phrase (Manual)	0.94	0.81	0.82	0.62
Verb Phrase (Machine)	0.94	0.77	0.68	0.41
User Profile	0.89	0.78	0.75	0.53
All	0.92	0.79	0.76	0.53

Table 4: Human performance (accuracy) on Choice-75

# takeaway.

- Choice-75
  - A new task of decision branching in event-centric reasoning
  - A dataset with fine-grained annotation and alignment between human perception and LLM(s)
  - Hard scenarios are still challenging for LLM(s)

# references.

[1] Keisuke Sakaguchi, Chandra Bhagavatula, Ronan Le Bras, Niket Tandon, Peter Clark, and Yejin Choi. 2021. *proScript: Partially Ordered Scripts Generation*. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2138–2149, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[2] Alisa Liu, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2022. *WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation*. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6826–6847, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.