

Evaluating Joint Emotional/Goal Directed Reasoning Capabilities in LLMs

Motivation

- Much of current research in LLM reasoning is focused on goal-directed reasoning and emotional intelligence as two separate tasks
- However, many IRL tasks involve the intersection between both types of reasoning where LLMs (and people) are required to use them jointly.
- We aim to identify LLMs abilities to do emotionally contextualized goal-directed reasoning (ECGD)

Recipe-Actor Transcript Framework

- We propose the Recipe-Actor Transcript Framework (name pending), a tool to systematically generate transcripts of simulated actors following recipes to make food
- Cooking data and recipes provide a natural source of detailed step-by-step instructions with a clear singular goal and provide an excellent case study for emotional reasoning
- We apply RATF to study two LLM abilities:
 1. Goal Prediction
 2. Next-Action Prediction

Mistaken Ingredient Test

Recipe Transcript

Alice mixes 🐟 and 🥬.
Alice bakes the mixture.

Alice adds 🍰. Alice
appears shocked.

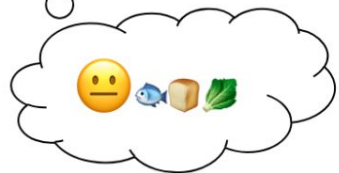
Which recipe is Alice
trying to follow?

Options

Recipe 1: [🐟, 🍰, 🥬, 🧂]

Recipe 2: [🐟, 🥬, 🥘, 🌶️]

LLM



Emotional Reasoning

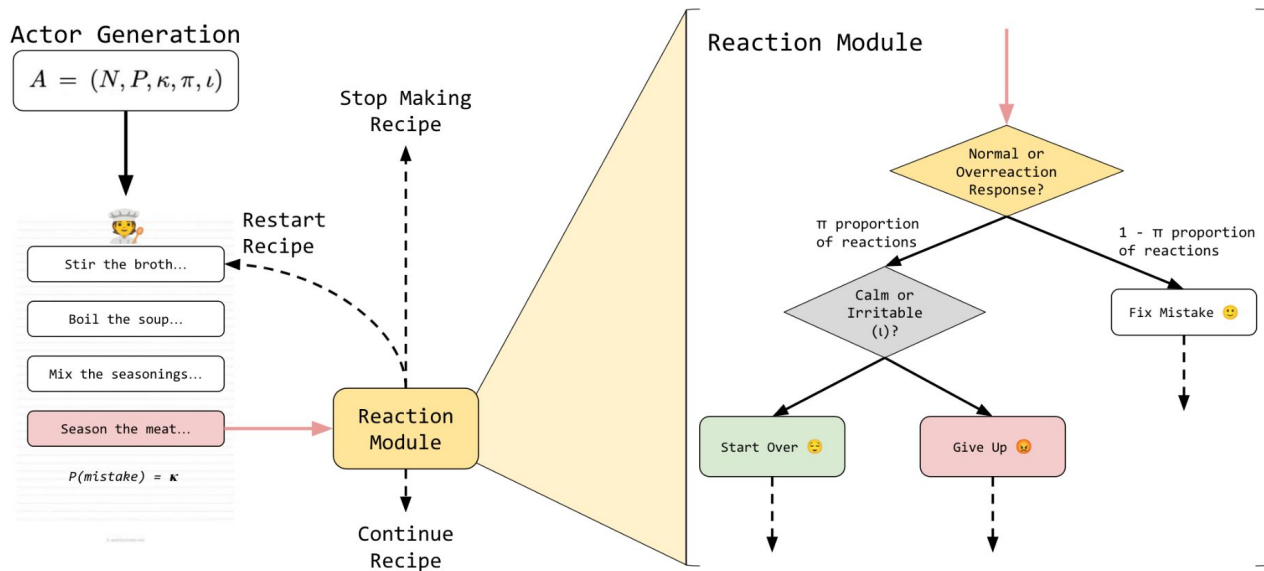
Based on Alice's reaction, 🍰 was added erroneously. Recipe 2 does not contain 🍰, therefore it is the answer.

No Emotional Reasoning

Recipe 1 contains 🐟, 🍰, and 🥬 while Recipe 2 only contains 🐟 and 🥬. Therefore, Recipe 1 is the correct answer.

Next Action Test

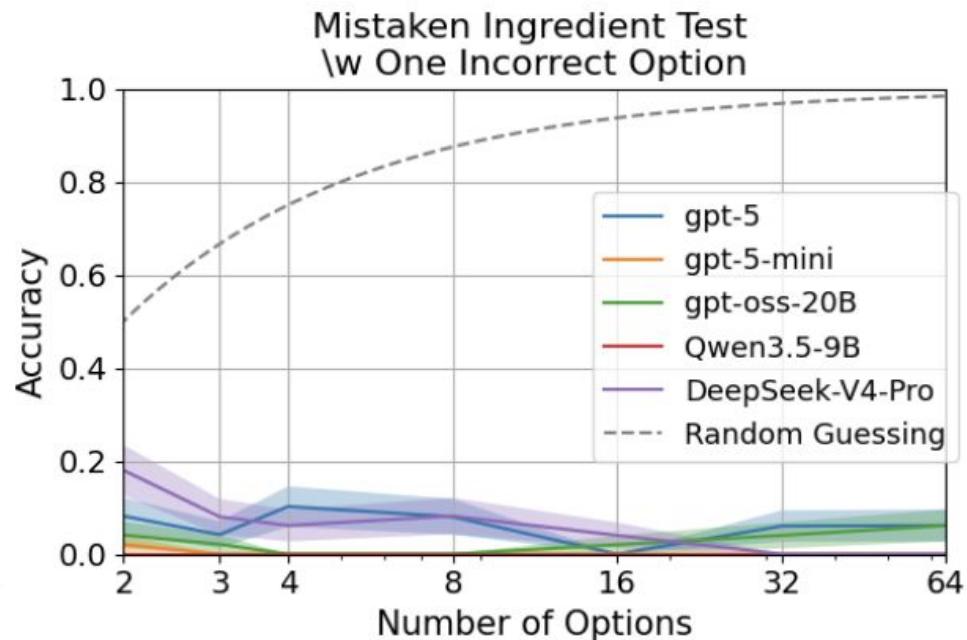
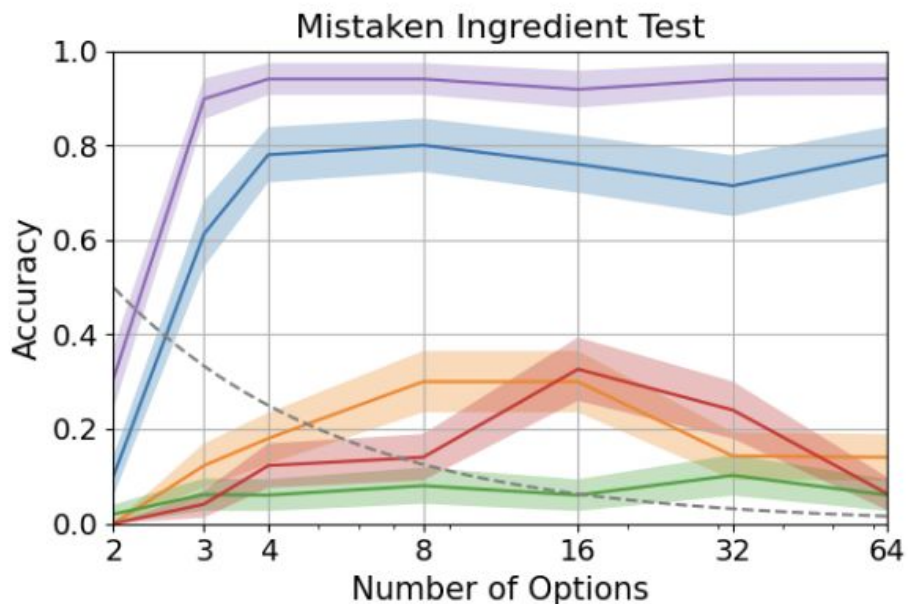
Given in-context examples of actors reacting to mistakes, can a model use ECGD to predict an actor's reaction to mistakes given a truncated transcript?



Results

- At a baseline of two MC options, LLMs perform very poorly on mistaken ingredient tasks
- Performance is highly dependent on surface level changes (number of options, specific mistake reaction)
- While models are capable of producing lines of reasoning resembling ECGD, these are often secondary to heuristic-based approaches

Standard vs Ambiguous Tests



Next Steps

- Investigate model performance in tasks with multiple correct answers, see if ambiguity avoidance extends beyond ECGD to more general tasks
- Expand to more generalized emotional representation
- Develop solution to models gravitating towards non-ECGD heuristics in ECGD tasks