Robots that ask for help: Uncertainty Alignment for Large Language Model Planners

keyword

Introduction

LLMs for planning (기존 방법론들)

(+) leveraging the vast amount of prior knowledge and rich context
(-) tendency to
hallucinate, high degree of *ambiguity*

⇒ Provide a way to clarify

(i) calibrated confidence : seek sufficient help
(ii) minimal help
: narrowing down possible ambiguities

⇒ collectively refer to these sufficiency and minimality conditions as
uncertainty alignment

KNOWNO - know when you don't know
: framework for aligning the uncertainty of LLM-based planners

(1) Conformal Prediction - utilizing this theory

(2) Prove theoretical guarantees on calibrated confidence in singe-step / multi-step planning problems

(3) Evaluation: language-instructed manipulation tasks in simulation and HW

Overview : Robots that Ask for Help

Language-based planners

- the uncertainty of next prediction p(y) → highly sensitive to variable-length k

Planning as multiple-choice Q&A

- LLM로 semantically diff candidate next steps 생성
- normalized scores that can be used by various uncertainty quantification methods (e.g. thresholding, ensemble methods) - CP framework 내에서

Robots that ask for help

- LLM planning (+CP)
- Environment 'e': formulated as a partially observable Markov decision process
- policy의 구성: multiple-choice generation / prediction set generation / human help / low-level control

Goal: uncertainty alignment

- We don't assume knowledge of D
- Uncertainty alignment setting: calibrated confidence, minimal help

---

Limitations and future work

- Limitation: task completion → dependent on the text input to the LLM
- Future work → incorporate uncertainty of the perception module & low-level action policy