大規模言語モデルにより制御されるシステムの評価基準の提案

研究背景

大規模言語モデル(LLM：Large Language Model)
LLMを用いたシステムは
再現性の担保が難しい

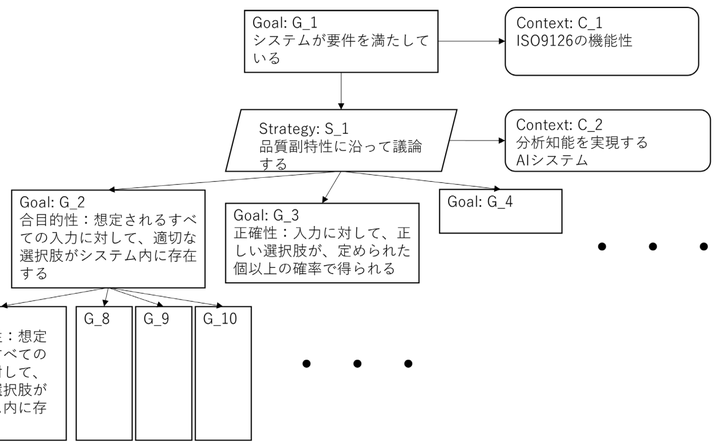
→評価基準を決めておきたい

→どのような要素を含めていれば
いいかが分かるように提示

GSN [1]

GSN（Goal Structuring Notation）を用いた分析手法

- 満たすべき基準の可視化による開発の単純化および精度向上
- AI実践システムの評価にも用いられた[2]

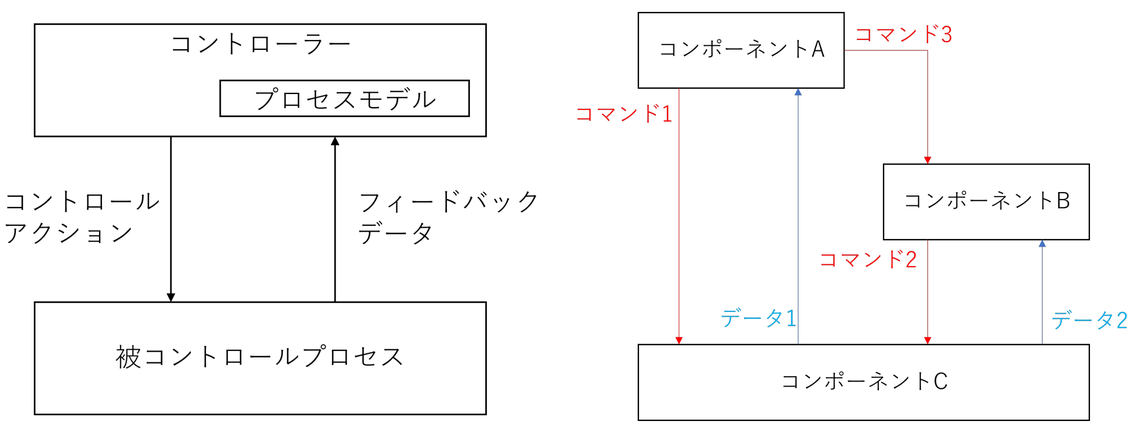


STAMP/STPA [3]

STAMP（Systems-Theoretic Accident Model and Process）：システム理論に基づくアクシデントモデル

STPA（STANP based Process Analysis）：STAMPベースのプロセス分析

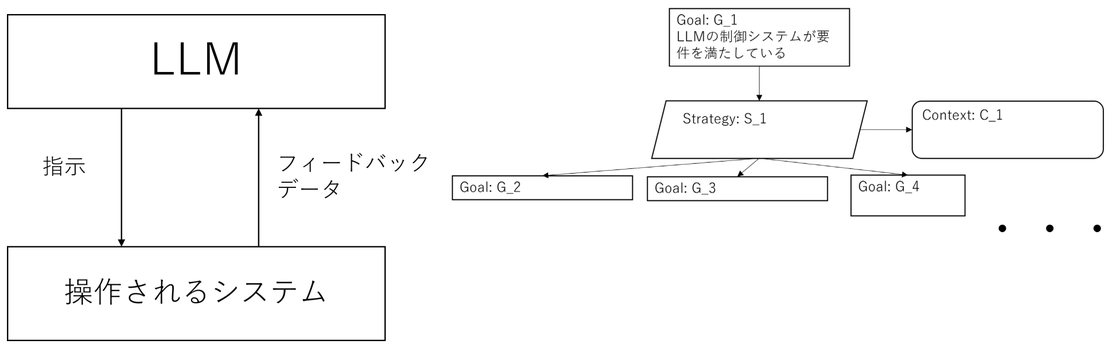
- アクシデント対策として適用できるフレームワーク



提案手法

現状の課題
→再現性の担保が難しいので、含んでいればシステムが担保されるとわかる評価基準を提案したい

対応
→GSNとSTAMP/STPAを合わせた手法でLLMシステムを評価することで、含むべき要素の発見を試みる



ケーススタディ

VOYAGER[4]：Minecraftを自動でプレイするエージェント

↑GPT-4の指示で木材獲得に成功

エージェントが持つ自動カリキュラム・スキルライブラリ・反復プロンプトメカニズムのほか、GPT-4を用いることでオープンエンドな世界でのプレイを実現

今後の課題

- 定量的なデータの定義や獲得
- 実際に基準を適用した場合の結果の確認
- 他研究との比較によるこの基準の精度の評価

参考文献
[1] GSN : https://scsc.uk.scsc-141c
[2] 竹内広宜1，山本修一郎2 (1 ： 日本アイ･ビー･エム株式会社 東京基礎研究所，2 ： 名古屋大学大学院情報科学研究科): “保証ケースを用いたAI実践プロジェクトの導入準備評価”，ウィンターワークショップ2018･イン･宮島 論文集，pp.4-5, 2018-01-11
[3] IPA 独立行政法人 情報処理推進機構: “はじめてのSTAMP/STPA ～システム思考に基づく新しい安全性解析手法～”, 2016
[4] Guanzhi Wang1 2 ✉, Yuqi Xie3, Yunfan Jiang4 *, Ajay Mandlekar1 *, Chaowei Xiao1 5, Yuke Zhu1 3, Linxi "Jim" Fan1 † ✉, Anima Anandkumar1 2 † (1 ： NVIDIA, 2 ： Caltech, 3 ： UT Austin, 4 ： Stanford, 5 ： UW Madison * ： Equal contribution, † : Equal advising, ✉ : Corresponding authors): “VOYAGER: An Open-Ended Embodied Agent with Large Language Models,” arXiv preprint arXiv:2305.16291, 2023

芝浦工業大學 SHIBAURA INSTITUTE OF TECHNOLOGY

システム理工学部　電子情報システム学科　ソフトウェア工学研究室
教授　**久住憲嗣**

システム理工学部　電子情報システム学科　4年
髙橋允治

■お問い合わせは　芝浦工業大学　複合領域産学官民連携推進本部　03-5859-7180　sangaku@ow.shibaura-it.ac.jp