物理シミュレータによる自己検証ループを用いた,高信頼なロボット用データセットの構築手法

芝浦工業大学 工学部　基盤システム研究室 嶋中雄大

背景

近年,自然言語を用いて直感的にロボットを操作できる技術への需要が高まっており,大規模言語モデル(LLM)はこの課題を解決する鍵として注目されている
しかし,LLMの応用には大きな障壁が存在する.

LLM応用の３つの壁

- ハルシネーション
- データセット不足
- リソース制約

目的

これらの障壁を解決するために,低リソースでも動作する小規模言語モデルをファインチューニングするための,ロボット特化の高品質データセットを作成する機構を提案し,今後の研究計画とする.

提案手法

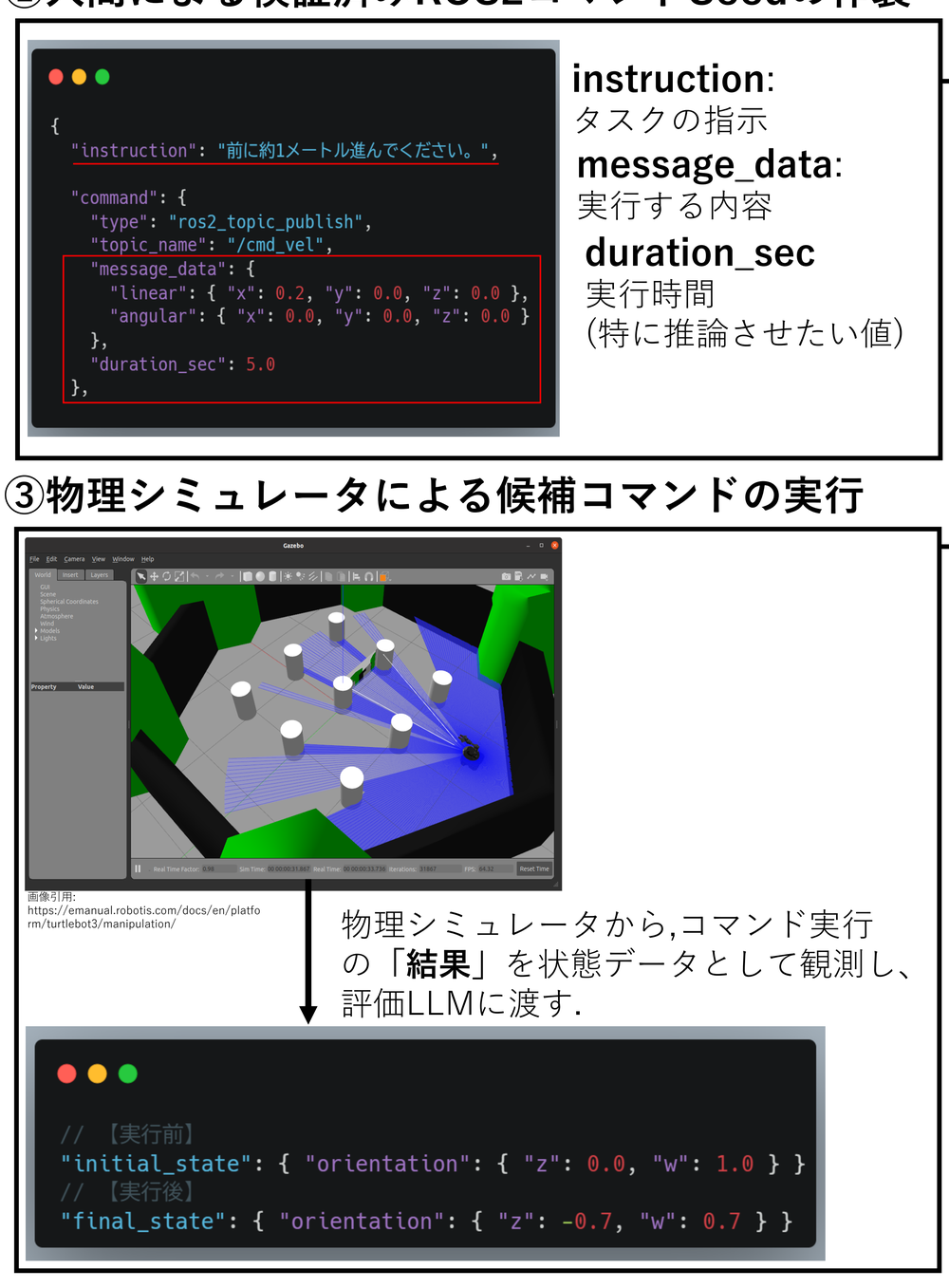
①人間による検証済みROS2コマンドSeedの作製

```
{
  "instruction": "前に約1メートル進んでください。",

  "command": {
    "type": "ros2_topic_publish",
    "topic_name": "/cmd_vel",
    "message_data": {
      "linear": { "x": 0.2, "y": 0.0, "z": 0.0 },
      "angular": { "x": 0.0, "y": 0.0, "z": 0.0 }
    },
    "duration_sec": 5.0
  },
```

instruction:
タスクの指示
message_data:
実行する内容
duration_sec
実行時間
(特に推論させたい値)

②LLMによるタスク&候補コマンド生成




```
{
  "task_instruction": "その場で右に90度回転してください。",
  "command_candidates": [
    {
      "candidate_id": 1,
      "command": {
        "type": "ros2_topic_publish",
        "topic_name": "/cmd_vel",
        "message_data": {
          "linear": { "x": 0.0, "y": 0.0, "z": 0.0 },
          "angular": { "x": 0.0, "y": 0.0, "z": -0.5 }
        },
        "duration_sec": 3.14
      },
      "reasoning": "標準的な角速度(-0.5 rad/s)で90度(π/2 rad)回転するために必要な時間を計算した本命の仮説。"
    },
    {
      "candidate_id": 2,
      "command": {
        "type": "ros2_topic_publish",
        "topic_name": "/cmd_vel",
        "message_data": {
          "linear": { "x": 0.0, "y": 0.0, "z": 0.0 },
          "angular": { "x": 0.0, "y": 0.0, "z": -0.8 }
        },
        "duration_sec": 1.96
      },
      "reasoning": "より速い角速度(-0.8 rad/s)で回転するアグレッシブな仮説。実行時間は短くなる。"
    },
    {
      "candidate_id": 3,
      "command": {
        "type": "ros2_topic_publish",
        "topic_name": "/cmd_vel",
        "message_data": {
          "linear": { "x": 0.0, "y": 0.0, "z": 0.0 },
          "angular": { "x": 0.0, "y": 0.0, "z": -0.5 }
        },
        "duration_sec": 3.5
      },
      "reasoning": "標準的な角速度で意図的に少し長く回転させる仮説。過回転の失敗データを収集する狙いがある。"
    }
  ]
}
```

task_instruction:
タスク文章
例①:3m直進せよ, 90度右回転せよ(単純なタスク)
例②3m直進後,90度右回転せよ(組み合わせタスク)
candidate_id:
複数の候補コマンド列の作成
→ DPOデータ用に複数の候補コマンドを生成

一つの指示から,成功・失敗の双方を含む**多様なコマンド候補**を生成.
これが後の**SFT/DPOデータセットの源泉**となる.

③物理シミュレータによる候補コマンドの実行

画像引用:
https://emanual.robotis.com/docs/en/platform/turtlebot3/manipulation/

物理シミュレータから,コマンド実行の「**結果**」を状態データとして観測し、評価LLMに渡す.

```
// 【実行前】
"initial_state": { "orientation": { "z": 0.0, "w": 1.0 } }
// 【実行後】
"final_state": { "orientation": { "z": -0.7, "w": 0.7 } }
```

データセット

⑤SFT(Supervised Fine-Tuning)

成功の「正解」を教える教師データ (SFT)
評価LLMで最も高く評価されたコマンド列を教師ありデータセットとして蓄積

```
{
  "instruction": "その場で右に90度回転してください。",
  "successful_command": {
    "type": "ros2_topic_publish",
    "topic_name": "/cmd_vel",
    "message_data": {
      "linear": { "x": 0.0, "y": 0.0, "z": 0.0 },
      "angular": { "x": 0.0, "y": 0.0, "z": -0.5 }
    },
    "duration_sec": 3.14
  }
}
```

⑥DPO(Direct Policy Optimization)

成功と失敗を比較させ、
“より良い選択”を学ばせる選好データ (DPO)
評価LLMで最も高く評価されたコマンド列と最も低く低く評価されたコマンド列を選考チューニングデータセットとして蓄積

```
{
  "prompt": "その場で右に90度回転してください。",
  "preferred_answer": {
    "type": "ros2_topic_publish",
    "topic_name": "/cmd_vel",
    "message_data": "{ "angular": { "z": -0.5 }
}", "duration_sec": 3.14
  },
  "non_preferred_answer": {
    "type": "ros2_topic_publish",
    "topic_name": "/cmd_vel",
    "message_data": "{ "angular": { "z": -0.5 }
}", "duration_sec": 3.5
  }
}
```

評価

以下の3モデルを比較し,
物理検証を経た高品質データによる学習の優位性を実証する.
(a)ベースLLM (事前学習のみ,物理感覚なし)
(b)検証なしSFTモデル (LLMが生成したコマンドを検証せずに学習)
(c) 本手法 (SFT+DPO) モデル (物理的に検証済み高品質データのみで学習)

課題と展望

1. **Sim-to-Real,シミュレータから実世界へ**
   → Domain Randomization技術,実機データのフィードバック
2. **より複雑で実用的なタスクへ**
   →アームによる実用的なマニピュレーションタスクへと拡張
3. **マルチモーダルな理解へ**
   → VLM(Vison Language Model)を導入し,曖昧な指示への対応
   例)「テーブル上の赤いリンゴを取って」