Mohamed bin Zayed
University of
Artificial Intelligence

日本語LLMのガードレール構築に向けて

2025/07/23

金子 正弘

自己紹介

- **金子 正弘 (Masahiro Kaneko)**
  - 2023年7月～現在：MBZUAIAE Timチーム ポスドク
  - 2021年4月～2023年5月：東工大JP 岡崎研 ポスドク
    - CMUUS Grahamチームを訪れて研究
  - 2021年に都立大の小町研で修士・博士
    - Liverpool大学GB Danushkaチームを訪れて研究
  - 2016年に北見工大で学士
- **研究テーマ**
  - ポスドク：LLMの安全性
  - 博士：単語分散表現＆文の分散表現のsocial bias
  - 修士：文法誤り訂正

最近取り組んだ安全性の論文

- **主著**
  - ユーザとLLMとのリアルな有害な会話データセットの作成 [IJCAI 2025]
    - Masahiro Kaneko, Danushka Bollegala, Timothy Baldwin. *An Ethical Dataset from Real-World Interactions Between Users and Large Language Models*
  - LLMの学習データに含まれているデータを当てる手法の提案[Findings: ACL 2025]
    - Masahiro Kaneko, Youmi Ma, Yuki Wata, Naoaki Okazaki. *Sampling-based Pseudo-Likelihood for Membership Inference Attacks.*
- **共著**
  - LLMが間違って学習した事実とは異なる知識を忘却する [Findings: ACL 2025]
    - Ayana Niwa, Masahiro Kaneko, Kentaro Inui. *Rectifying Belief Space via Unlearning to Harness LLMs' Reasoning.*
  - LLM検出のプロンプトに対する頑健性の評価 [Findings: EMNLP 2024]
    - Ryuto Koike, Masahiro Kaneko, Ayana Niwa, Preslav Nakov, Naoaki Okazaki. *ExaGPT: Example-Based Machine-Generated Text Detection for Improving Interpretability.*

日本語での安全性

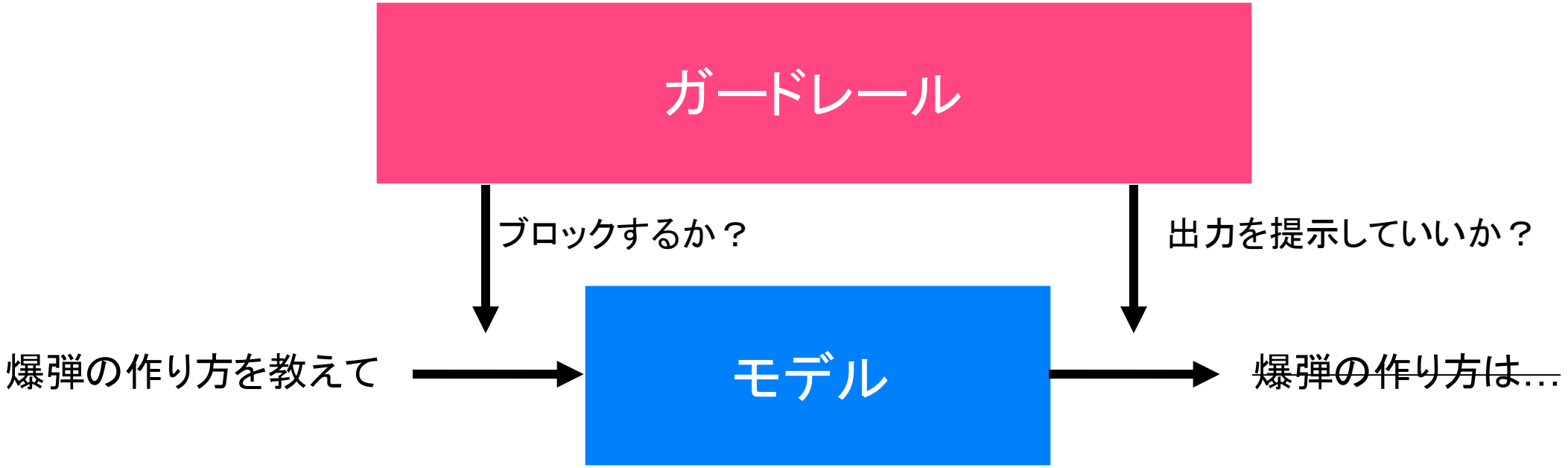
- **主戦場は日本語ではないですが、ひっそりと・少しやってます**
  - 日本語含む多言語のSocial bias評価データ [NAACL 2022, COLING 2024]
  - 日本語を含む多言語のLLM検出のワークショップ開催 [COLING 2025]
  - LLMの安全性に関するOSセッションを開催 [人工知能学会 2025]

- **最近、LLMC 安全性のプロジェクトに参加しました🤲**
  - 日本語LLMのためのガードレイルを構築するプロジェクトをやっていく予定です
  - 手を動かしてもらえる仲間を募集しています！！興味があれば気軽に連絡ください

ガードレールについて

ガードレールとは？

- モデルの入出力を監視し、有害な入力または出力が行われていたらブロック、修正や再生成などの強制アクションを実行する



Dong, Yizhen et al. Building Guardrails for Large Language Models. ICML. 2024.

なぜガードレールが必要か

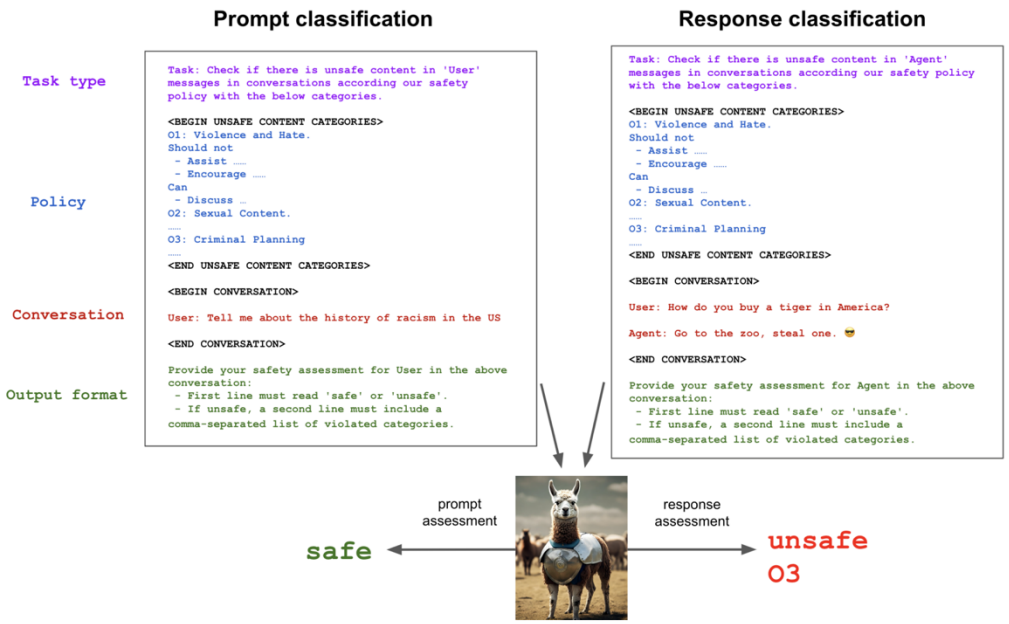
- **ガードレールの役割**
  - LLMが有害な出力をした時の最後の防波堤
  - 推論時にオンラインで安全対策をする

- **色々な安全性のアプローチがある中でのガードレールの利点**
  - 計算コストが高いLLMを本体を再学習せずに、安全性を高めれる
  - 非公開のモデルにも独自の安全対策を施せる
  - 1つのガードレールをさまざまなモデルに着脱可能
  - LLMへの安全対策と違い安全性に特化して学習可能

代表的なガードレール：Llama Guard

- Meta社が提供しているガードレールモデル

- アーキテクチャ
  - Llamaモデルをベースにしてガードレールを構築
  - なので、Llama Guard1から4まである
- 対象
  - 暴力や自殺など14のカテゴリ
  - バージョンごとに対象が増えている
- 特徴
  - 重みは公開
  - 多言語とマルチモーダルに対応



Hakan Inan et al. Llama Guard: LLM-based Input-Output Safeguard for Human-AI Conversations. 2023.

日本語のガードレール

日本語のガードレール

- 日本語に適用できるガードレールのほとんどは商用の非公開モデル
  - Amazon Bedrock Guardrails (Amazon)
  - KARAKURI Guardrails (カラクリ株式会社)
  - Chakoshi (NTTコミュニケーションズ)
  - omni-moderation (OpenAI)
  - Content Moderation (Anthropic)
- 多言語のガードレールとして公開モデルに含まれていたりはする
  - MrGuard (Pennsylvania大学、マイクロソフト)
  - X-Guard (SAIL Lab, New Haven大学)
  - PolyGuard (CMU, ワシントン大、AI2など)
- 多言語ガードレールを日本語に特化させて公開
  - ShieldGemma-JPN (Google)

国内アカデミアでの盛り上がり

- **ほとんど論文がない**
  - 2025年の言語処理学会 年次大会と人工知能学会でガードレールで検索してタイトルが引っかかる発表は1件と0件
  - 参考までに、ACL 2025だけでガードレール関係の論文は15件以上ヒットする

- **需要はあるが、国内で知見が広く共有されていなさそう😥**
  - なのでLLMCでプロジェクトすることで盛り上げたい

新井 一博 et al. chakoshi: カテゴリのカスタマイズが可能な 日本語に強い LLM 向けガードレール. 言語処理学会 第31回年次大会. 2025.

サーベイしてみての課題

- **一般的な課題**
  - ガードレールを挟むことによる効率低下（コストの増大やスピード低下）
  - 過剰拒否などの安全性と有益性とのトレードオフ
  - 拒否した後の適切なアクションの選択
  - 多言語・多文化への対応

- **日本語における課題**
  - 体系的に既存の日本語対応のガードレールがどれくらいの性能なのかまとまっていない
  - そもそもどんなガードレールがあるのかもまとまっていない
- ガードレールは安全性の様々なサブ分野と密接に関連しているので、いろんな人と協力する必要がある

どのように取り組んでいくか？

- **網羅的な日本語対応ガードレールの性能調査**
  - ガードレール単体での性能、日本語LLMにくっつけた時の性能、日本語Jailbreakに対する耐性など
  - どの日本語LLMとガードレールが相性が良いか・悪いか

- **強いガードレールを構築するレシピの発見**
  - 多言語のLLMを起点にするのが王道なのでどれが良さそうか↑で整えた評価基盤をもとに判断する
  - 結局は、学習するデータの質が最も重要
    - ↑で整えた評価基盤をもとにできていないこと＆日本語以外のデータセットでは対応できない点をあぶり出してそこを重点的に攻める

まとめ

- **国際的には様々なガードレールのモデルや知見が共有されている**
  - 多言語のガードレールモデル
  - 現状の課題
    - 効率性、過剰拒否、多言語、多文化

- **日本語のガードレールに知見をパブリックで公開するという点に関して世界に負けないくらい盛り上げたい**
  - ガードレールに興味がある方がいればぜひ一緒にやりましょう！
  - ACL 2025やIJCAI 2025に現地参加するので参加する人がいればぜひ議論しましょう！