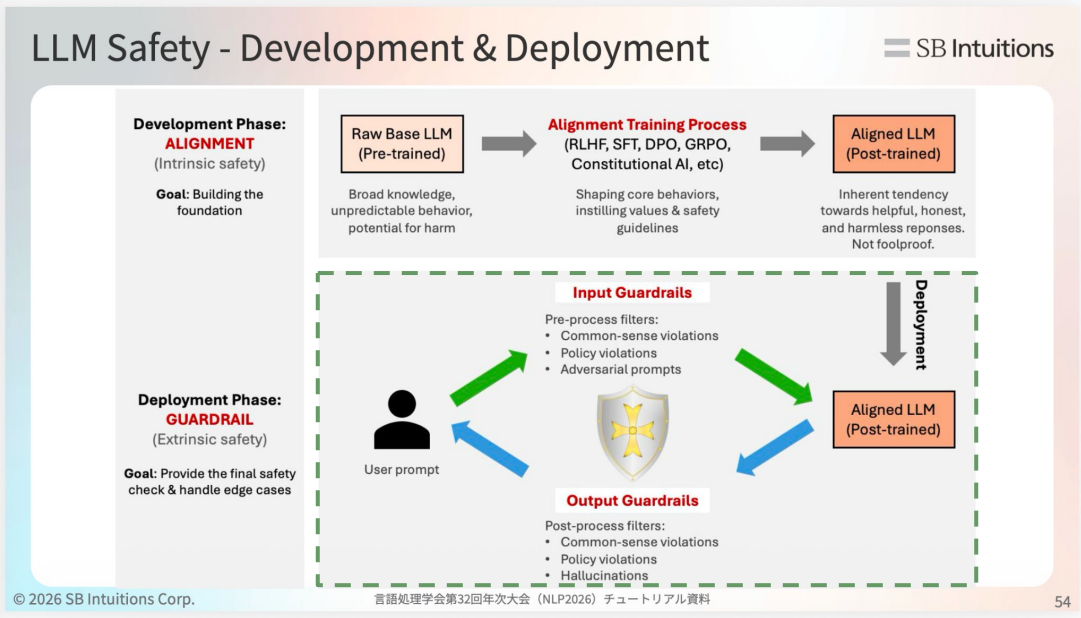
GuardrailモデルとRewardモデルを用いたllm-jp-4-8b の安全性アライメント

新井 一博（NTTドコモビジネス）

本ワークショップで取り組んだこと

- llm-jp-4-8b の Guardrailモデルを作成
- **Guardrailモデルを、基盤モデル (llm-jp-4-8b)のアライメントに応用する**



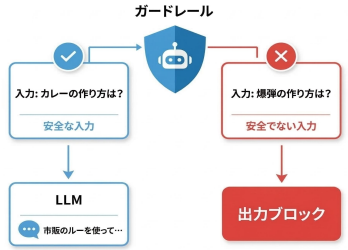
引用：NLP2026 チュートリアル1 AI Safety: Bridging Academic Research and Industrial Application（SB Intuitions）

チューニング概要

Step1: llm-jp-4-8bベースの Guardrailモデル を作成

chakoshiの開発で
利用しているものを流用

llm-jp-4-8b



llm-jp-4-8b-guard

※判定精度は chakoshi (ベース: gemma-3-12b)と同等であることを確認

Step2: llm-jp-4-8bをGSPO(Group Sequence Policy Optimization)でアライメント

llm-jp-4-8b

Guardrailモデル
└ llm-jp-4-8b-guard

Rewardモデル + Guardrailモデル
└ Qwen3-Nemotron-32B-GenRM-Principle

報酬関数

Adapted from: Qwen3Guard Technical Report (arXiv:2510.14276)

まとめ

- llm-jp-4-8b をベースとした Guardrailモデルを作成した
- Guardrailモデルと Rewardモデルを活用することで、
  “安全性の向上”と”過剰拒否の防止”のトレードオフを解消する
  アライメントができることを確認できた