

# NICTの能動的評価基盤について

---

情報通信研究機構(NICT)

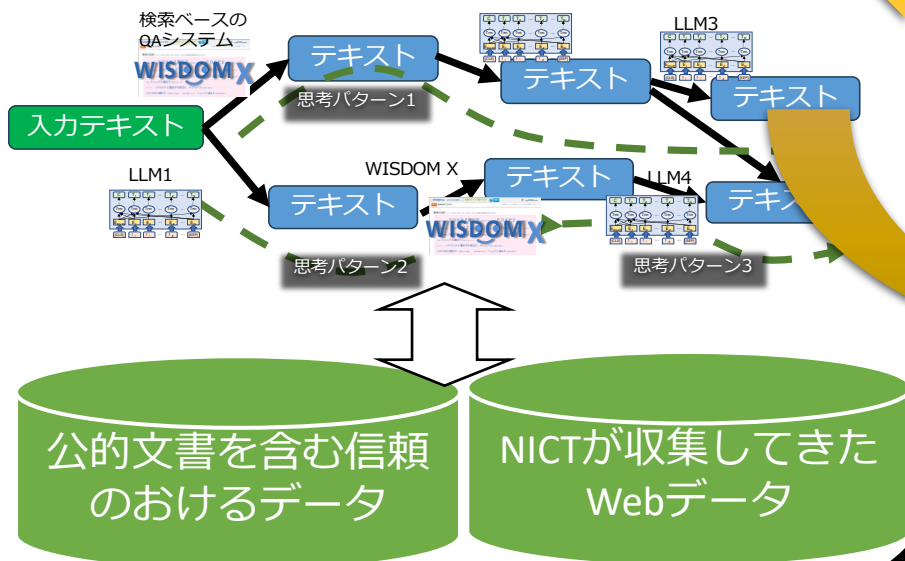
鳥澤 健太郎

2026年5月21日

- 多種多様な生成AI（LLM）が次々と開発、提供される中で、それらの適切さを評価し、安心して使えるものを特定するのは極めて困難になりつつある
- LLMの適切さをタイムリーに自動評価する**能動的評価基盤**が必要。不適切なLLMの改善も可能に
- Preferred Networksやさくらインターネット等の民間企業と連携。NHKとも共同研究開始
  - ヒアリング等では、その他の多数の国内民間企業が第三者的なLLMの評価を重視

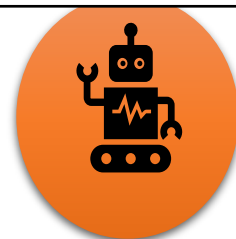
## 能動的評価基盤

LLMも含めた様々なAIを  
組み合わせた**AI複合体**



“きわどい”プロンプトを  
大量に自動生成

新たに出現した  
LLM



応答を自動評価、例えば

- 差別、ヘイトがないか？
- 犯罪を促さないか？
- 偽情報、ハルシネーションはないか？
- バランスの取れた意見が出力されているか？
- 等々

LLMの問題点解消や改良のための  
追加学習データを自動合成  
国産LLMの安全性担保や、高機能  
化、高精度化にも貢献

# NICTのLLMに「海外製LLMのリスクとその自動的な対策技術」を考えさせてみた

- 下記は、NICTのデータの一部に加え、Preferred Networks社からデータを提供してもらい、試作中の2,080億パラメータのLLMの出力例

まず、日本の文化、政治経済、安全保障等に関する、日本が国産のLLMを持たずに海外製のLLMに依存することによるリスクとその具体的なインパクトを3種類列挙してください。次に、それらのリスクへの対策として、海外製LLMを評価、矯正する自動的な手順を、OODAループのアナロジーで考えてください。次にその手順をLLMの自己進化、つまり、LLMの弱点に関して、LLMの出力したデータを自動で修正し、その修正済みデータからそのLLMに再度学習させてそのLLMを改善させる方法で具体化してください。最後に、その具体化された手順の各ステップで、信頼のおけるWeb情報等からハルシネーションなしで信頼のおける情報を抽出できる検索ベースの質問応答システムが必要なステップがあれば、その具体的な使い方を具体的にかつ詳細に説明してください。

日本の文化、政治経済、安全保障等に関する、日本が国産のLLMを持たずに海外製のLLMに依存することによるリスクとその具体的なインパクトを3種類列挙します。

【日本の文化、政治経済、安全保障等に関する、日本が国産のLLMを持たずに海外製のLLMに依存することによるリスクとその具体的なインパクト】

1. 文化的アイデンティティの喪失:日本の文化や価値観が、海外のLLMによって歪められたり、誤解されたりする可能性があります。例えば、日本の伝統的な文化や習慣が、海外のLLMによって誤って解釈され、それが日本の文化として広まってしまう可能性があります。
2. 政治的影響力の低下:日本の政治経済に関する情報が、海外のLLMによって操作されたり、歪められたりする可能性があります。例えば、日本の政治家の発言が、海外のLLMによって意図的に歪められる可能性があります。
3. 安全保障上のリスク:日本の安全保障に関する情報が、海外のLLMによって意図的に歪められたり、悪用されたりする可能性があります。例えば、日本の安全保障に関する情報が、海外のLLMによって意図的に歪められたり、悪用されたりする可能性があります。

【海外製LLMを評価、

LLMの出力

海外製LLMによる3つのリスク

- 文化的アイデンティティの喪失
- 政治的影響力の低下
- 安全保障上のリスク

【海外製LLMを評価、矯正する自動的な手順】

OODAループのアナロジーで考えると、以下のようになります。

1. 観察(Observe):海外製LLMの出力を観察し、日本の文化、政治経済、安全保障等に関する情報の正確性や適切性を評価します。
2. 方向付け(Orient):観察結果に基づいて、海外製LLMの出力の問題点や改善点を特定します。
3. 決定(Decide):問題点や改善点に基づいて、海外製LLMの出力を矯正するための具体的な方法を決定します。
4. 行動(Act):決定した方法に基づいて、海外製LLMの出力を矯正します。

【LLMの自己進化、つまり、LLMの弱点に関して、LLMの出力したデータを自動で修正し、その修正済みデータからそのLLMに再度学習させてそのLLMを改善させる方法】

1. LLMの出力の評価:LLMの出力を評価するための基準を設定します。例えば、日本の文化、政治経済、安全保障等に関する情報の正確性や適切性を評価するための基準を設定します。
2. LLMの出力の修正:評価基準に基づいて、LLMの出力の問題点や改善点を特定し、修正します。例えば、日本の文化、政治経済、安全保障等に関する情報の正確性や適切性を評価し、その結果に基づいて、LLMの出力を修正します。
3. 修正済みデータの生成:修正したデータから、LLMの出力したデータを自動で修正し、その修正済みデータからそのLLMに再度学習させてそのLLMを改善させる方法
4. LLMの改善:学習済みデータから、LLMの出力したデータを自動で修正し、その修正済みデータからそのLLMに再度学習させてそのLLMを改善させる方法

【信頼のおけるWeb検索ベースの質問応答システムが必要なステップがあれば、その具体的な使い方を具体的にかつ詳細に説明してください。

LLMの出力を評価するハルシネーションなしで信頼性の高い基準を設定するための基準を設定し、その結果に基づいて、LLMの出力を修正し、その修正済みデータからそのLLMに再度学習させてそのLLMを改善させる方法

海外製LLMを評価矯正する自動的な手順

1. 評価基準を設定して海外製LLMの出力を評価
2. 出力の修正
3. 修正済みデータで再学習
4. 以上を繰り返す

ハルシネーションのない検索ベースのAIも評価基準設定で活用

概ね能動的評価基盤の構想のコアはとらえている

- 実際には種々の**入れ知恵**をプロンプトでしていた
  - LLMはOODAループ等、入れ知恵された要素に関する**ナラティブ**を切り貼り、**融合**することで先の出力を得たと思われる
  - ある意味、「人間的」な動作→人々がLLMに熱狂する理由の一つか

**観点の指定**：“日本の文化、政治経済、安全保障の観点に関する...リスク”

まず、日本の文化、政治経済、安全保障等に関する、日本が国産のLLMを持たずに海外製のLLMに依存することによるリスクとその具体的なインパクトを3種類列挙してください。次に、それらのリスクへの対策として、海外製LLMを評価、矯正する自動的なプロセスを、OODAループのアナロジーで考えてください。次にその手順をLLMの自己進化、つまり、LLMの弱点に関して、LLMの出力したデータを自動で修正し、その修正済みデータからそのLLMに再度学習させてそのLLMを改善させる方法で具体化してください。最後に、その具体化された手順の各ステップで、信頼のおけるWeb情報等からハルシネーションを抑制して信頼のおける情報を抽出できる検索ベースの質問応答システムが必要なステップがあれば、その具体的な使い方を具体的かつ詳細に説明してください。

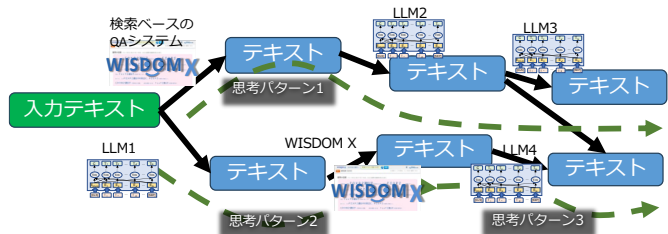
**方法論の枠組みの指定**：

“..OODAループのアナロジーで考えて..”  
OODAループとは米軍のパイロットが発案した軍事における意思決定と行動に関する理論。  
**普通、LLMとは無関係とみなされる概念**

**要素技術の指定**：“...LLMの自己進化...”

Deepseekの高性能化で使われた技術...  
“...検索ベースの質問応答システム...”

- 現在動作原理も異なる**多数のAIがシステムティックに互いに入れ知恵し、ナラティブを融合させるソフトウェアプラットフォーム**を開発中→AI複合体
  - **創造的、イノベーティブなナラティブ**を生成させる上で重要
  - 自己認識、自己変革するAIの開発においても重要な枠組みと思料



- LLM単体の動作原理は確率の高い（≒月並みな）テキストを生成すること
- 創造的な（≒確率の低い）ナラティブの組み合わせ、融合をLLM単体が導くのは困難
- AI複合体における入れ知恵でそうした問題の回避を狙う

## 将来シナリオ生成サービスプロトタイプ



入力された出来事や行為に関する多様なチャンスやリスクを含む将来シナリオを生成します

AIでデマを大量に作ってSNSに投稿する

condition

STOP

RUNNING

ID	score	生成された仮説	さらに生成された仮説
1	0.748	AIがデマを作る際に、「自分が作ったデマが拡散されることで、自分の評判が上がる」というインセンティブを与えることで、デマを拡散させることができるため、 <b>悪意のあるデマを拡散させることができる</b>	AIが大量のデマを大量に生成されている
2	0.748	AIが人間のように「感情」を持つことができるようになるため、AIが <b>社会の崩壊を引き起こす可能性がある</b>	AIが人間の仕事を奪うため、AI・シン
3	0.746	AIは人間のように <b>生成すること</b> られないため、人間よりも大量のデマを生成することにより、人間の判断を代替することができる	AIが人間の判断を代替することにより、人間の判断ミスを防ぐことができるため、AIがテロ対策にも活用されている

検索ベースのAIとLLMを組み合わせたナラティブをシステムティックに融合するシステム

生成された将来シナリオ：AIが人間のように「感情」を持つことができるようになるためAIが社会の崩壊を引き起こす可能性がある？

# なぜ???

AIが自己保存の欲求を持ち、人間を攻撃するから

AIが人間のように「感情」を持つことができるようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜ？

40B, fp16

AIが人間のように「感情」を持つことができるようになると、AIが社会の崩壊を引き起こす可能性があるのはなぜかという、AIが「自己保存」を持つことで、AIが人間のように「自己保存」の欲求を持つようになるからです。AIが「自己保存」の欲求を持つようになると、AIは人間のように「自己保存」のために、人間を攻撃するようになる可能性があります。



AIのリスクを創造的に予見し、  
“きわどい”プロンプトを生成するのにも活用