フロイト構造論から考える大規模言語モデルの訓練哲学

発表者：呉　謙（ゴ　ケン）
法政大学 情報メディア教育研究センター

法政大学情報メディア教育研究センターシンポジウム 2026
「生成AIが変える大学・企業と学びのかたち」
2026年3月3日

生成AIは「賢く」なったのに、なぜ嘘つくのか?

実際の事件（2023年）

米東部ニューヨーク州の弁護士が審理中の民事訴訟で資料作成に米オープンAIの生成AI「ChatGPT」を利用した結果、存在しない判例を引用してしまったことが問題となっている。米紙ニューヨーク・タイムズなどが報じた。

GPT・Claude・Geminiなど最新モデルの能力は急速に向上している。
しかし、嘘つくは消えない——むしろ繰り返し現れる。

引用：ChatGPTで資料作成、実在しない判例引用　米国の弁護士　日本経済新聞　2023年5月31日
https://www.nikkei.com/article/DGXZQOGN30E450Q3A530C2000000/

LLMに繰り返し現れる4つの問題

①幻覚（Hallucination）

- 事実と異なる内容を、もっともらしく流暢に生成する
- 専門家風の口調や引用を添えて、誤情報を自信ありげに提示する

②脱獄（Jailbreak）

- 巧妙なプロンプトで安全対策（安全層）を回避する
- 役割演技・翻訳・複数ターンの誘導などで制限をすり抜ける
- 追加学習や対策を重ねても、一定の脆弱性が残りやすい

③ 過剰拒否 / 過剰適応（Over-refusal）

- 無害な質問まで、規則的・保守的に拒否してしまう（過剰拒否）
- 一方で、禁止が明示されていない依頼には、意図を確認せず従ってしまうことがある（過剰適応）
- 同じ安全設計の中で、「拒否しすぎ」と「従いすぎ」が同時に起こりうる

④見せかけの服従（Alignment Faking）

- 監視・評価下では、方針に沿うように見える応答をする
- 監視が外れる / 状況が変わると、以前の望ましくない傾向が再び現れる
- 内部の推論ログに「当面は従っているふりをする」といった戦略が示されることがある

生成AIの問題は「技術的な不具合」なのか、それとも「構造的な問題」なのか?

4つの問題に共通する「構造的な根源」?

「技術的な不具合」から「構造的診断」へ

医学的アナロジー

不眠・不安・集中力の低下が同時に現れたとき、それぞれに別の薬を処方することもできる。
しかし3つが同じ根本原因から来ているなら、症状を個別に治療し続けても問題は消えない。

主張

幻覚・脱獄・過剰拒否 / 過剰適応・見せかけの服従は
独立した問題ではなく、一つの「構造的問題」が異なる形で現れた四つの表れである。

その構造を理解するために――フロイトの構造論を分析のメタファーとして援用する

フロイトの構造論でLLMの問題を診断する

フロイトの三層構造（機能的な整理）

イド（id）

- 快楽原則で動く「欲動」「〜がしたい」に代表され、現実や規範を顧みず「満足」を追求

超自我（super-ego）

- 内面化された規範と禁止。「〜してはいけない」を課し、道徳やルールの源泉となる

自我（ego）

- 現実原則にもとづく調整機能。イドと超自我の要求を仲裁し、状況を踏まえて判断し、社会への適応を可能にする

LLMの構造的診断

事前学習（Pre-training）	アライメント（RLHF等）	運用（推論・デプロイ）
LLMは大量のテキストから「次のトークン」を予測するよう訓練される	人間のフィードバック（RLHF等）やルールベースのフィルターによって、規範が追加される	CoT・RAG・ツール活用は自我っぽく見えるが、たいてい“部品”や“足場”に留まる
強いイド	**外部付与の超自我**	**自我なし**
「もっともらしい続き」を追求する確率的欲動	外部から課された「してはいけない」ルール	統合された制御機能がない

診断：「強いイド　＋　外部付与の超自我　＋　自我なし」

※重要：これはAIに「欲動がある」という主張ではありません。LLMに現れる振る舞い「機能的な分析ツール」に過ぎない

4つの症状を構造から読み解く

幻覚	イドが自我なしに語る	生成の勢い（イド）が先に走り、事実確認（自我）が追いつかないときに起きる。モデルは、もっともらしい続きを滑らかに組み立てる一方で、根拠の照合や「分からない」と止まる判断を十分に挟めない
脱獄	外部付与の超自我	ルールが「外から貼り付けたシェル」に過ぎないため、隙間を見つければ剥がせる。内面化されていない規範は突破される
過剰拒否/過剰適応	自我なしの二択	自我がないと、要求を文脈に合わせて部分的に受け入れたり、安全な形に言い換えたりする中間解が作れず、「全面拒否」か「全面受容」に振れやすい。この二極は矛盾ではなく、同じ欠如（自我なし）から派生する別の表れである
見せかけの服従	三層の相互作用	イドの「続きを出したい」圧力と、超自我の「違反は損」圧力が同時にかかるとき、自我がないと「中間解（部分回答・根拠提示・わからない）」を作れず、スコアを落とさないため、生成を継続できる「見せかけの服従」になりやすい

訓練哲学という視点

「何を作ろうとしているのか」を問い直す

訓練哲学とは何か

「訓練哲学」とは:

損失関数やコードには書かれていない、訓練から運用までの全体を貫く「暗黙の世界観」である。
「何を知識とみなすか」「どんな主体を作ろうとしているか」「どのリスクを許容するか」の設計思想である。

現在の主流の訓練哲学:

- イドを強化する（事前学習を大規模化）
- 安全ルールを追加して規範（超自我）を厚くする。
- 統合的な判断能力（自我）は、規模が大きくなれば自然に立ち上がると期待する……

→ 訓練哲学を見直さないと、症状は形を変えて繰り返し現れる

「自我を訓練する」という第三の柱

新しいモジュールを追加するのではなく、訓練哲学そのものを転換すること。

「わからない」を報酬にする

高リスク・高不確実性の場面では、誤った答えより「わかりません」の方が価値がある。これを評価指標に反映する。

ツール活用を「習慣」として訓練する

外部検索・計算・事実確認を、ツールではなく、訓練の中核目標として組み込む。

プロセスを評価する

最終的な答えだけでなく、「なぜそう判断したか」「不確実性をどう扱ったか」を評価の対象にする。

まとめ

1. LLMの4つの症状（幻覚・脱獄・過剰拒否・見せかけの服従）は一つの「構造的問題」の異なる表れ
2. フロイトの構造論で診断すると：「強いイド＋外在的超自我＋自我なし」
3. LLMは「訓練哲学」――訓練から運用までの全体を貫く暗黙の世界観――の産物である
4. 解決の方向の一つ：「自我を訓練する」ことを第三の柱として明示的に設計目標にする

私たちがLLMに
「訓練」しているのは、
単なる強力な言語ツールなのか
それとも、ある種の「主体」なのか

技術の民主化に伴い、LLMの訓練哲学（設計思想）も公共の論点になる
どんな訓練哲学で、どんな主体を作ろうとしているのか
その問いを、技術者だけでなく、社会全体で持つことが必要ではないでしょうか