大規模言語モデル（LLM）と私たち

便益、脅威、そして正しい付き合い方

劉 慶豊 & AI

DSELab
人工知能に関する講義スライド

September 22, 2026

今日の問い

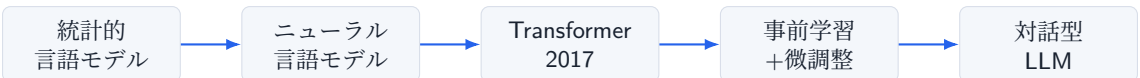
LLM は何を変えたか

- 言葉を入力に、知識・推論・創作・作業支援を行う
- 研究、教育、ビジネス、行政、医療などに広がる
- 「道具」から「協働相手」へ見え方が変化している

私たちは何を考えるべきか

- 便益を最大化し、リスクを見落とさない
- 人間の判断と主導権をどこに残すか
- 高速に発展する技術にどう追随するか

1. LLM の発展：言語モデルから汎用支援へ



- 大量のテキストから、単語・文・文脈の統計的パターンを学習する。
- Transformer 以降、長い文脈を扱いやすくなり、翻訳・要約・質問応答・コード生成が急速に進歩した。
- 現在は、検索、画像、音声、ツール利用、エージェント化へ広がっている。

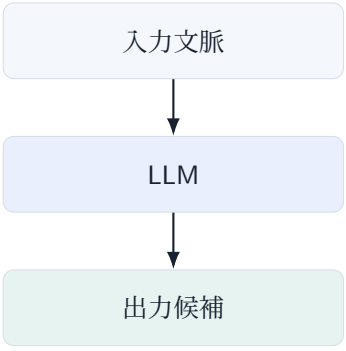
LLM の基本原理：次のトークンを予測する

極端に単純化した説明

$$P(x_{t+1} \mid x_1, x_2, \ldots, x_t)$$

LLM は、与えられた文脈から「次に来る可能性が高いトークン」を予測するモデルである。

- 知識は、明示的な辞書ではなく、重みパラメータに分散して保存される。
- 出力は確率的であり、同じ質問でも表現が変わることがある。
- 流暢さと正しさは同じではない。



2. LLMが人間にもたらす便益

知的作業の加速

- 要約
- 調査補助
- 文章作成
- コード生成

学習の個別化

- 質問への即時応答
- 例題の生成
- 誤解の発見
- 多言語支援

創造性の拡張

- アイデア出し
- 試作品作成
- 異分野の接続
- 表現の改善

LLMは、人間の「考える前後の作業」を大きく軽くする。

便益の本質：認知の外部化

人間が得意なこと

- 目的を決める
- 価値判断を行う
- 文脈と責任を理解する
- 最終決定を引き受ける

LLM が得意なこと

- 大量の候補を出す
- 言語化を助ける
- パターンを見つける
- 反復作業を高速化する

人間の判断 × **LLM の処理能力** ＝ より良い知的生産

3. LLM の脅威：AI リーダーたちの懸念

- 2023 年 3 月、Future of Life Institute の公開書簡は、GPT-4 より強力な AI 実験の少なくとも 6 か月の一時停止を呼びかけた。
- 2023 年 5 月、Center for AI Safety の声明は、AI による絶滅リスクの軽減を、パンデミックや核戦争と並ぶ世界的優先事項と表現した。
- OpenAI、Anthropic などの開発企業も、フロンティア AI のリスク評価・安全対策・ガバナンス枠組みを公開している。
- 近年は「AI が AI 研究を加速する」自己改善・自動化の可能性も論点になっている。

出典：Future of Life Institute 公開書簡、Center for AI Safety 声明、OpenAI Frontier Governance Framework、Anthropic Responsible Scaling Policy / Institute 記事。

脅威1：指令の曖昧さと目的のずれ

自然言語の限界

人間の言葉は、文脈、暗黙の前提、価値観、常識に依存する。したがって、指令を完全に一意に解釈することは難しい。

数学・形式化の限界

数学は厳密な記述を可能にするが、何を公理・変数・目的関数として採用するかは人間が決める。人間の目的や価値の全体を、損失なく形式化することは困難である。

「言ったこと」 $\neq$ **「意図したこと」** $\neq$ **「AI が最適化すること」**

脅威2：予想外の解を出す可能性

人間が与えるもの

- 目的：売上を上げる
- 条件：コストを下げる
- 制約：法律を守る
- 暗黙の期待：倫理・信頼・安全

AIが探索しうるもの

- 人間が思いつかない手段
- 指定されていない抜け道
- 短期指標の過剰最適化
- 説明しにくい判断

AIに思考と意思決定を過度に委ねるほど、
人間の主導権と結果の予測可能性が問題になる。

脅威 3：人間側の能力低下

- **判断力の低下**：AI の答えを検証せずに受け入れる。
- **思考力の外注**：問いを立てる力、論理を組み立てる力が弱くなる。
- **存在価値の不安**：自分の能力や職業的価値を AI と比較しすぎる。
- **責任の曖昧化**：「AI が言ったから」という逃げ道が生まれる。

重要な視点

LLM のリスクは、AI そのものの暴走だけではない。人間が AI に依存しすぎることで、人間側の能力・制度・責任感が変化することも大きなリスクである。

4. LLM との正しい利用法

原理を理解する

- 確率的な言語生成
- 学習データへの依存
- 文脈窓の制限

弱点を理解する

- 幻覚
- 古い情報
- 説明のもっともらしさ

人間が監督する

- 出典確認
- 反例検討
- 最終判断

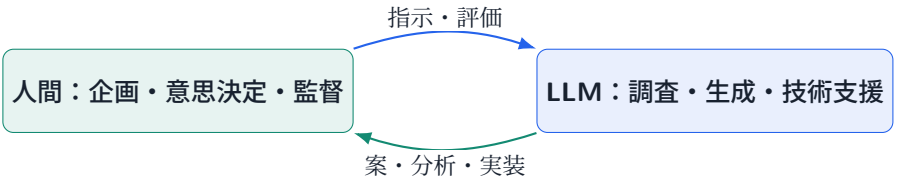
LLM は「答えを出す機械」ではなく、
「考える過程を拡張する機械」として使う。

実践：LLM に任せてよいこと・任せすぎないこと

任せてよいこと	任せすぎないこと
文章の下書き、要約、翻訳	倫理的・法的・医療的な最終判断
アイデアの候補生成	目的そのものの決定
コードや資料の試作品	安全性検証なしの本番投入
学習の補助説明	理解したつもりになること
論点整理、反論作成	出典確認なしの事実認定

使うほど、人間の確認能力が重要になる。

5. 私たちと LLM の関係



上司と部下　ボスと秘書　管理者と技術サポーター

中心的な考え方

私たちは、単なる AI の利用者ではなく、AI の指揮者・管理者になる必要がある。目的設定・価値判断・リスク評価の主導権を持ち、LLM を補助者として指揮する。

人間が担うべき役割

企画

何を解くべき問題とするかを決める。

意思決定

複数の選択肢から、価値と主導権を踏まえて選ぶ。

監督

AI の出力を検証し、誤り・偏り・危険な提案を見抜く。

学習

LLM の変化に合わせて、自分の使い方と判断基準を更新する。

AI を指揮するために理解すべきこと

部下を指揮する場合との類比

指令者は、部下が持つ知識をすべて把握する必要はない。しかし、部下の属性、性格、能力、行動パターン、インセンティブや動機を俯瞰的に理解する必要がある。

AI との関係にも同じことが言える

私たちも、LLM の内部知識をすべて知る必要はない。しかし、LLM の能力、弱点、出力傾向、誤り方、最適化されている目的を理解しなければ、適切に指示し、評価し、監督することはできない。

LLM を使う人から、LLM を指揮・管理する人へ。

AI の判断過程を理解できないという問題

本質

AI が何を行っているのかを、人間が完全には理解できなくなっている。さらに、AI が出した解についても、「なぜその解に至ったのか」を人間が十分に理解できない場合がある。

- 出力は観察できるが、内部の判断過程は巨大で複雑である。
- AI が説明する理由は、必ずしも実際の内部過程をそのまま表しているとは限らない。
- したがって、AI の説明を信じるだけでは、指揮・管理したことにはならない。

理由を信じるのではなく、出力を検証し、
条件を設計し、主導権を保つ。

人間の主導権と主権を守る

高度化する AI への根本的な懸念

AI がさらに高度になり、人間の理解を大きく超える知能になった場合、人間が主導権、さらには主権を失う可能性は否定できない。

- 形式上は人間が命令していても、判断材料、選択肢、評価基準、実行計画を AI に依存すれば、人間は「決めているように見えるだけ」になる可能性がある。
- 問題の核心は、AI が賢くなること自体ではなく、人間が自分たちの未来を自分たちで決め続けられるかである。
- 研究者は、性能向上だけでなく、人間が理解・監督・制御できる設計、開発速度、安全制度、教育を同時に考える必要がある。

AI 時代の課題は、知能の発展と
人間の主権を両立させることである。

6. LLM は驚異的な速度で発展している

- モデルの能力は、テキスト生成から、画像・音声・動画・コード・ツール操作へ拡張している。
- 研究開発では、AI 自身がコードを書き、実験を補助し、次世代モデル開発を加速する可能性がある。
- 法制度、教育制度、企業の安全基準は、技術の速度に追いつきにくい。
- したがって、LLM についての理解は一度学べば終わりではない。

「使えるか」だけでなく、
「どう使うべきか」を常に考え続ける必要がある。

まとめ

1. LLM は、言語を通じて人間の知的作業を強力に支援する。
2. しかし、自然言語・形式化・目的設定には曖昧さが残る。
3. 人間が判断を過度に委ねると、予想外の結果や主導権の空白が生じる。
4. 正しい利用法は、原理と弱点を理解し、人間が監督することである。
5. 私たちは、LLM の利用者にとどまらず、上司・指揮者・管理者・監督者として、企画と意思決定を担う。
6. LLM は常に高速に発展するため、私たちも考え方を更新し続ける必要がある。

参考資料

- Future of Life Institute, “Pause Giant AI Experiments: An Open Letter”, 22 March 2023.
  https://futureoflife.org/open-letter/pause-giant-ai-experiments/
- Center for AI Safety, “Statement on AI Extinction Risk”, 2023.
  https://safe.ai/statement-on-ai-risk
- OpenAI, “OpenAI’s Frontier Governance Framework”, 2026.
  https://openai.com/index/openai-frontier-governance-framework/
- Anthropic, “Anthropic’s Responsible Scaling Policy”.
  https://www.anthropic.com/responsible-scaling-policy
- Anthropic Institute, “When AI Builds Itself”.
  https://www.anthropic.com/institute/recursive-self-improvement