生成**AI**は胡蝶の夢を見るか

―「思考の錯覚」と複雑性の壁：推論モデルにおける能力崩壊の計量分析―

2026.04.25 応用心理測定研究会 第9回

青木 貴寛

プロローグ：AIの「内面」を問う

- P.K.ディックの問い：人工物と人間の境界線としての「共感」
- 荘子の「胡蝶の夢」：認識と虚構が混濁する知能のメタファー

- AIの推論は「論理（現実）」か、それとも「確率（夢）」か？

LRM(大規模推論モデル)という新パラダイム

- OpenAI o1/o3, DeepSeek-R1, Claude 3.7 Sonnet, Gemini 3の台頭
- 特徴:回答前の膨大な「思考トークン(Thinking Tokens)」
- 期待されていたこと:自己反省(Self-reflection)による論理整合性の獲得

項目	LLM (従来の言語モデル)	LRM (推論モデル)
思考のタイプ	システム1（速い思考）	システム2（遅い思考）
反応の仕組み	入力に対して即座に確率的に高い言葉を繋げる	回答の前に内部で「思考トークン」を生成し、試行錯誤する
得意なこと	文章作成、要約、翻訳、知識の検索	数学、プログラミング、複雑な論理パズル、自己修正
スケーリング	学習データ量とモデルサイズで性能向上	推論時の計算量（考える時間）を増やすことで性能向上
代表例	GPT-4o, Claude 3.5, Gemini 1.5	OpenAI o1, o3, DeepSeek-R1, Claude 3.7

測定妥当性への疑義：既存ベンチマークの限界

- 既存試験（GSM8K等）は「学習済みデータ（汚染）」の懸念
- Apple論文の解決策：難易度を厳密に操作可能な「パズル環境」
- GSM-Symbolic：変数や名前の変更による精度の急落（Mirzadeh et al., 2025）

- 現在のAI評価ベンチマークは「推論能力」という潜在変数を純粋に測定できておらず、「既知のパターンに対する記憶力」という攪乱変数を多分に含んでいることが明らかに。

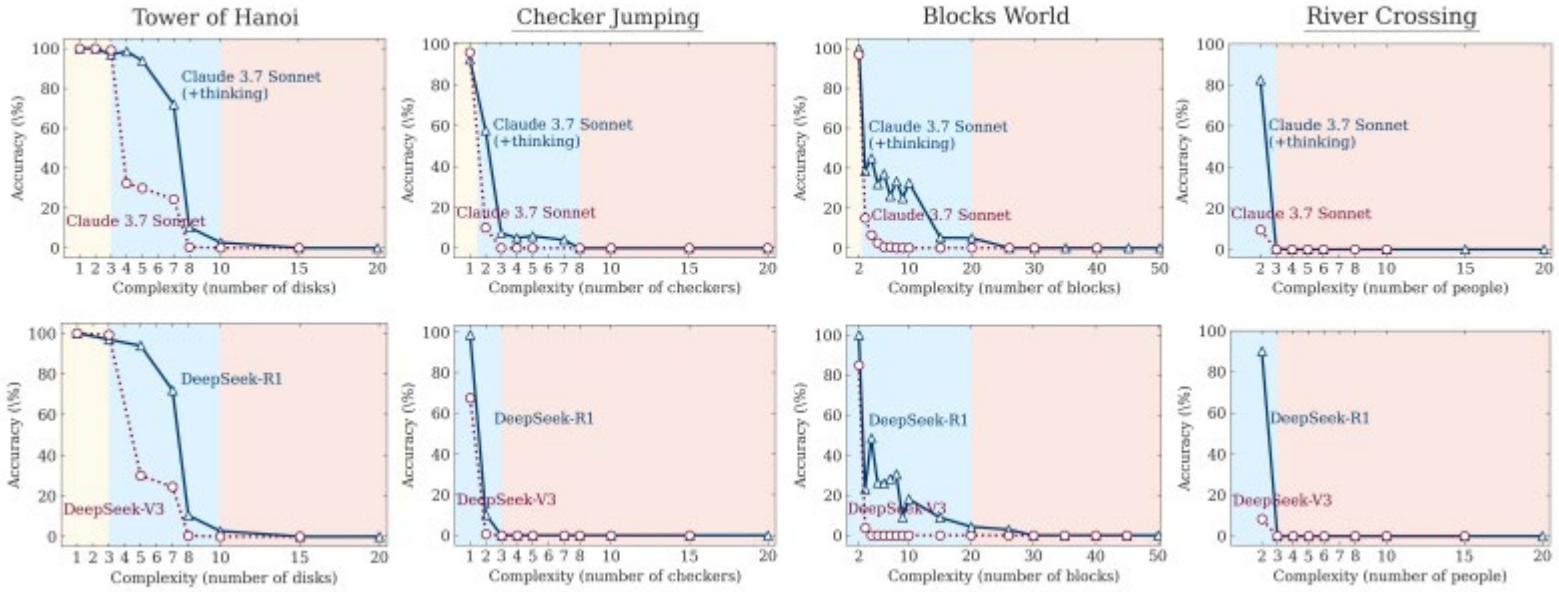
解析結果：
能力の「三つの領域（Regimes）」

1. 低複雑性：標準モデルが効率的。推論モデルは「考えすぎ（誤差）」で自滅

2. 中複雑性：思考トークンが真価を発揮し、標準モデルを凌駕する

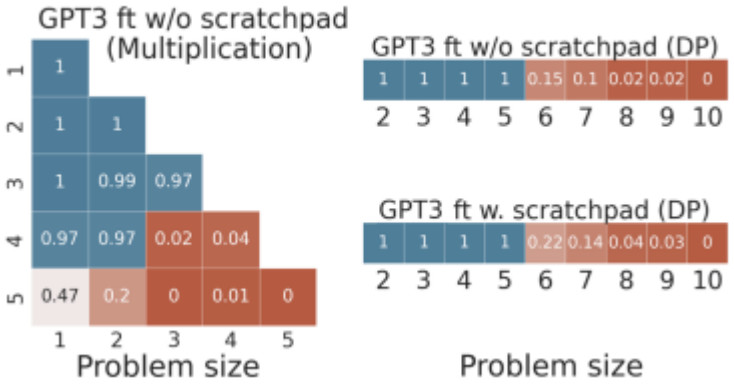
3. 高複雑性：精度の完全な崩壊（Collapse）。正答率は0%へ転落

崩壊の境界線：精度の崖



- 複雑性が一定の閾値を超えた瞬間に精度が消失する様子
- 出典: Shojaee et al. (2025, p. 7, Figure 4).

構成性の絶壁：ステップ数の罠



- 推論ステップ数が増えるに従いエラーが指数関数的に増大する「崖」
- 出典: Dziri et al. (2023, p. 4, Figure 3).

「アリス」のパラドックス：単純な論理の死

- 不思議の国のアリス問題：人間には自明な論理が最新AIで崩壊する怪異

- 「アリスには3人の兄がいます。それぞれの兄には1人の妹がいます。アリスには何人の妹がいるでしょうか？」

- 「3人の妹がいる」という誤答を出力。
- AIが「それぞれの兄には1人の妹がいる」というフレーズを、「兄弟の数だけ妹が存在する」という統計的なパターンとして処理してしまい、アリスという主体の視点を欠落させているため

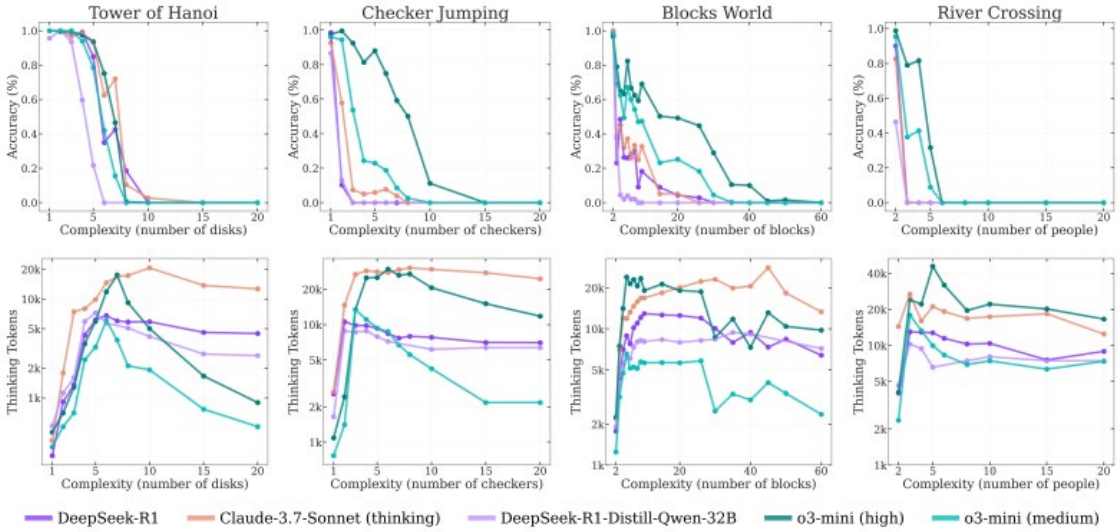
$$X = T + E$$

- 実は *T*（論理的思考）ではなく、データの汚染やパターンの合致という *E*（系統誤差）によって底上げされていたという事実

ハルシネーションの測定心理学

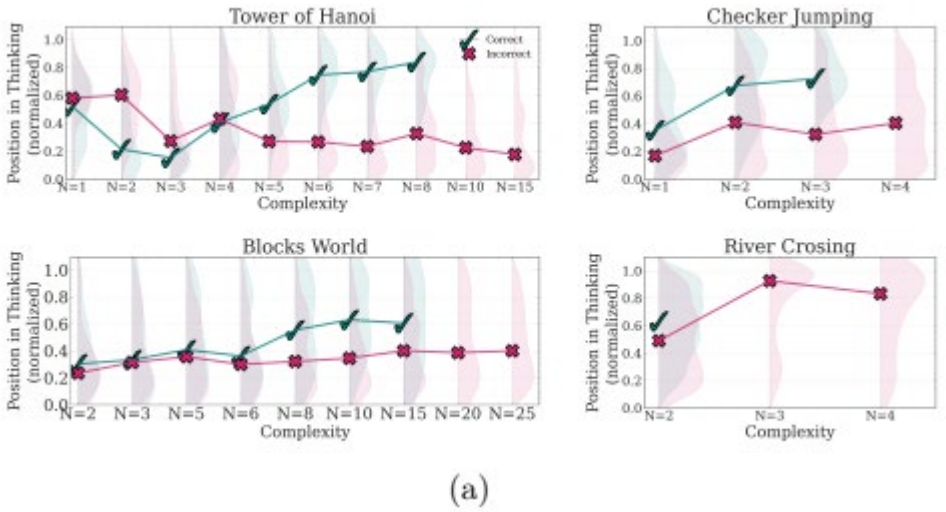
- ハルシネーション ＝「構成概念無関係な分散（CIV）」
- 測定対象（推論力）を「もっともらしい文章を作る能力」が汚染する系統誤差
- 思考を強制することが、かえって誤差を増幅させる逆転現象
- 推論能力とは無関係な「文章生成の癖」や「統計的な尤もらしさ」が、回答という測定値を汚染している状態

推論努力の逆説：なぜ考えるのをやめる？



- 難易度が上がると、AIは十分な予算があるのに自発的に思考量を減らす
- 出典: Shojaee et al. (2025, p. 9, Figure 6).

思考の軌跡：嘘の正当化プロセス



(a)

- 一度「間違った中間解答」を思いついてしまうと、その後の思考トークンのすべてを「その嘘を正当化するための論理構築」に費やしてしまう
- これを「不誠実な思考」と呼ぶ。AIは「真実」を探しているのではなく、「自分が最初に出した推測に合う物語」を生成しているに過ぎない。
- 出典: Shojaee et al. (2025, p. 10, Figure 7).

アルゴリズム実行能力の欠如：手順を知っていても……

- 事実：正しい解法手順を与えても、複雑性が増すとAIはそれを無視する
- AIは論理を「理解」しているのではなく、パターンの「残像」を追っているにすぎない

解決策

- AI（LLM）に直接計算や厳密な論理ステップを任せるのではなく、AIに「計算機（PythonやWolfram Alpha）」を使わせること

結論：夢を現実に繋ぎ止めるために

- AIの夢に付き合うのではなく、高度な計算機として管理
- 複雑性の壁を認識し、適切なツールへのオフロードを行うべき
- 測定の精度を守るのは、AIのトークンではなく、我々の専門知

参考文献（1/3）

- Brown, T., et al. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.
- Chen, J., et al. (2025). Reasoning models don’t always say what they think. arXiv preprint arXiv:2501.07703.
- Chen, X., et al. (2024). Do not think that much for 2+3=?: On the overthinking of o1-like LLMs. arXiv preprint arXiv:2412.21187.
- Chollet, F. (2019). On the measure of intelligence. arXiv preprint arXiv:1911.01547.
- Cobbe, K., et al. (2021). Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168.
- DeepSeek-AI. (2025). DeepSeek-R1: Incentivizing reasoning capability via reinforcement learning. arXiv preprint arXiv:2501.12948.
- Dubreuil, J., et al. (2024). Tülu 3: Pushing frontiers in open language model post-training. arXiv preprint arXiv:2411.15124.
- Dziri, N., et al. (2023). Faith and fate: Limits of transformers on compositionality. arXiv preprint arXiv:2305.18654.

参考文献（2/3）

- Li, Y., et al. (2025). LLMs can easily learn to reason from demonstrations structure, not content! arXiv preprint arXiv:2501.12563.
- Lightman, H., et al. (2023). Let's verify step by step. arXiv preprint arXiv:2305.20050.
- Lin, Z., et al. (2024). Stop overthinking: A survey on efficient reasoning for LLMs. arXiv preprint arXiv:2411.05124.
- McCoy, R. T., et al. (2023). Embers of autoregression: Understanding large language models through the problem they are trained to solve. arXiv preprint arXiv:2309.13638.
- Mirzadeh, S. I., et al. (2025). GSM-Symbolic: Understanding the limitations of mathematical reasoning in LLMs. arXiv preprint arXiv:2410.05229.
- Nezhurina, M., et al. (2024). Alice in Wonderland: Simple tasks showing complete reasoning breakdown. arXiv preprint arXiv:2406.02061.
- Nye, M., et al. (2021). Show your work: Scratchpads for intermediate computation with language models. arXiv preprint arXiv:2112.00114.
- OpenAI. (2024). Learning to reason with LLMs (o1 technical report). https://openai.com/index/learning-to-reason-with-llms/.

参考文献（3/3）

- Shojaee, P., et al. (2025). The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. arXiv preprint arXiv:2501.12345.
- Valmeekam, K., et al. (2023). Large language models still can’t plan (A benchmark for planning and reasoning). arXiv preprint arXiv:2206.10498.
- Valmeekam, K., et al. (2024). LLMs still can’t plan; can LRMs? A preliminary evaluation of o1 on PlanBench. arXiv preprint arXiv:2409.13741.
- Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998-6008.
- Wang, X., et al. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.
- Wang, Z., et al. (2024). DeepSeekMath: Pushing the limits of mathematical reasoning in LLMs. arXiv preprint arXiv:2402.03300.
- Wei, J., et al. (2022). Chain of thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.