現代AIはシンギュラリティに到達するのか？

大規模言語モデル: シンギュラリティの評価

石﨑龍之介,杉山麿人
（国立情報学研究所,総合研究大学院大学）

どんな研究？

大規模言語モデルが自律的かつ指数関数的に知能を増幅するシンギュラリティに達するかどうか、哲学と情報学の観点から検証します。

何がわかる？

現代AIの危険性を理論的に評価できる可能性があります。安全な状態で知能爆発ができれば、人類に莫大な利益をもたらす自己改善型AIが実現できる可能性があります。

背景・目的

哲学

拡張可能な方法 (D. Chalmers)
–AIがシンギュラリティに達する条件
–より知的なシステムの創造

究極の知能の持つ目標 (N. Bostrom)
G1. 自己保存 G2. 目標–内容の一貫性
G3. 知能向上
G4. 資源獲得

情報学

自己報酬型言語モデル (Yuan et al.)
–DPOの選択的自己メタ報酬を可能に
$\mathcal{L}_{\mathrm{R}}(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma\{r_\phi(x,y_w) - r_\phi(x,y_l)\}$

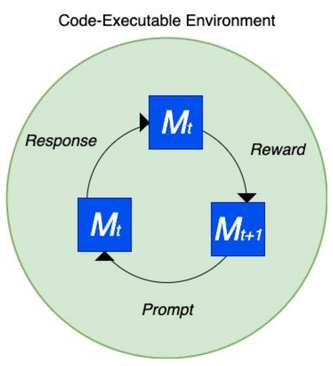
独学型オプティマイザ(Zelikman et al.)
–LLMの再起的な自己改良(RSI)プログラミング

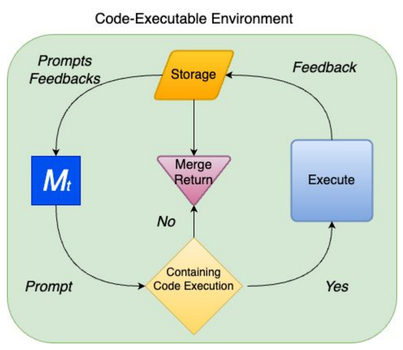
自律的デバッグLLM (Chen et al.)
問題のテスト/デバッグを自動化

研究内容

手法

・シンギュラリティの可能性のあるモデル構造





・RSI 報酬プロンプトフォーマット (RPF)の創造
・コーディング能力の最大値はおおよそ
（モデルパラメータ、*RSI_RPF*、シードデータ）
で決定される。

$$c_{\max} = C(\theta, RSI\_RPF, x \sim \mathcal{D}_{\mathrm{seed}})$$

もし c_{max} が完全に自己学習コーディングできる複雑性 γ に達すると、
それが拡張可能な方法になる。

$$C(\theta, RSI\_RPF, x \sim \mathcal{D}_{\mathrm{seed}}) > \gamma$$

何が起きる?

・LLM人口の指数関数以上の成長
・通常のマルウェアはロジスティックモデルに従う
・ε (N): LLMsは自己改善する事ができる

$$\frac{dN}{dt} = N(\epsilon(N) - \frac{k}{L}N + k)$$

もしいつも dN/dt>0なら、
指数関数以上の成長が起きる。

結論

・より倫理的な研究が必要
・安全に知能爆発が起こせれば
人類に莫大な利益をもたらす可能性

[Click to edit] 連絡先：国立情報学研究所 情報プリンシプル研究系　情報ビット
URL：https://www.nii.ac.jp/about/pr/jouhouken/　Email：oh@nii.ac.jp

Will Modern AI Reach Singularity?

Large Language Models: Assessment for Singularity

Ryunosuke Ishizaki, Mahito Sugiyama
(National Institute of Informatics)

What research?

Philosophically and Informatically, we examine whether LLMs will reach a singularity–the moment when LLMs amplify intelligence autonomously and exponentially.

What can we know?

We can theoretically assess the dangers of modern LLMs. If we can achieve an intelligence explosion safely, a self–improving AI would be benefitial to humanity.

Background・Purpose

Philosophy

Extendable method (D. Chalmers)
–the condition for AI to reach singularity
–creation of the better intelligent system

Common goals of ultraintelligence (N. Bostrom)
G1. Self–preservation G2. Goal–content integrity
G3. Intelligence enhancement
G4. Resource acquisition

Informatics

Self–rewarding Language Models (Yuan et al.)
–enables self–meta selection reward for DPO

$$\mathcal{L}_{\mathrm{R}}(r_\phi, \mathcal{D}) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \log \sigma\{r_\phi(x,y_w) - r_\phi(x,y_l)\}$$

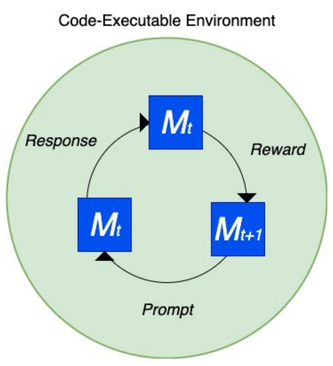
Self–Taught Optimizer (Zelikman et al.)
–Recursively Self–Improve (RSI) coding for LLMs

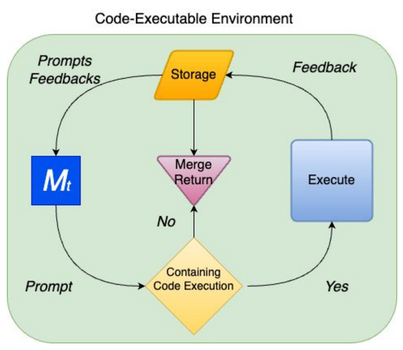
Self–Debugging LLMs (Chen et al.)
Testing & debugging of problems

Research Contents

Method

・Potential singularity model structure





・Creation of **RSI Reward Prompt Format (RPF)**
・Tha maximal complexity of coding ability is roughly determined by model parameter θ, *RSI_RPF*, and seed dataset *Dseed.*

$$c_{\max} = C(\theta, RSI\_RPF, x \sim \mathcal{D}_{\mathrm{seed}})$$

If c_{max} is greater than the fully self–coding complexity γ,
it will be an extendable method.

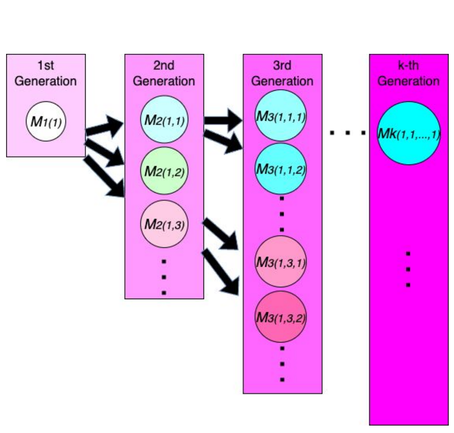
$$C(\theta, RSI\_RPF, x \sim \mathcal{D}_{\mathrm{seed}}) > \gamma$$

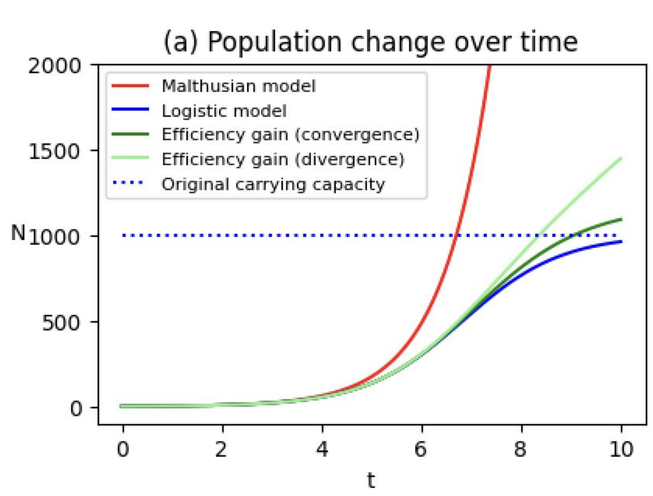
What will happen?

・super–exponential growth of population
・Usual Malwares end up to Logistic model
・ε (N): LLMs can improve themselves

$$\frac{dN}{dt} = N(\epsilon(N) - \frac{k}{L}N + k)$$

If always dN/dt>0,
super–exponential growth will happen.





Conclusion

・More ethical development is needed
・Safely reaching intelligence explosion gives us enormous benefits

[Click to edit] 連絡先：国立情報学研究所 情報プリンシプル研究系　情報ビット
URL：https://www.nii.ac.jp/about/pr/jouhouken/　Email：oh@nii.ac.jp