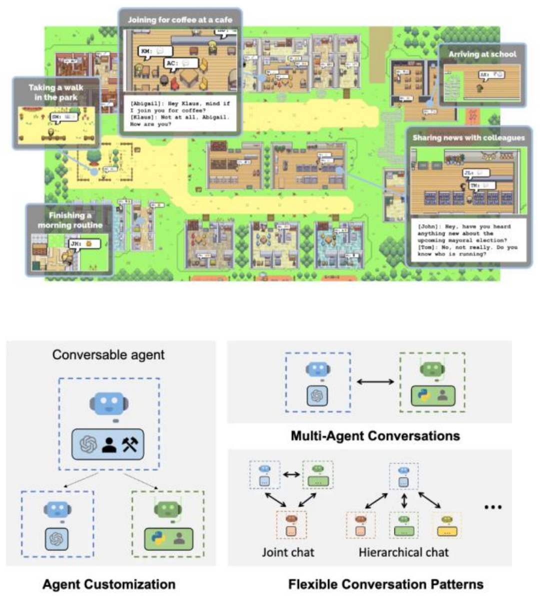
Bias Amplifiction in LLM Evolution: An Iterated Learning Perspective

Yi Ren[1], Shangmin Guo[2], Linlu Qiu[3], Bailin Wang[3], Danica J. Sutherland[1,4]

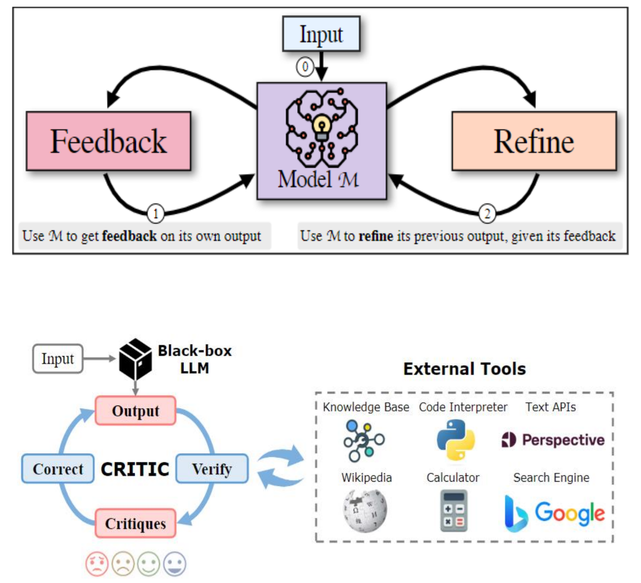
1. UBC; 2. University of Edinburgh; 3. MIT; 4. Amii

1. Motivation: Ubiquitious Self-play in LLM

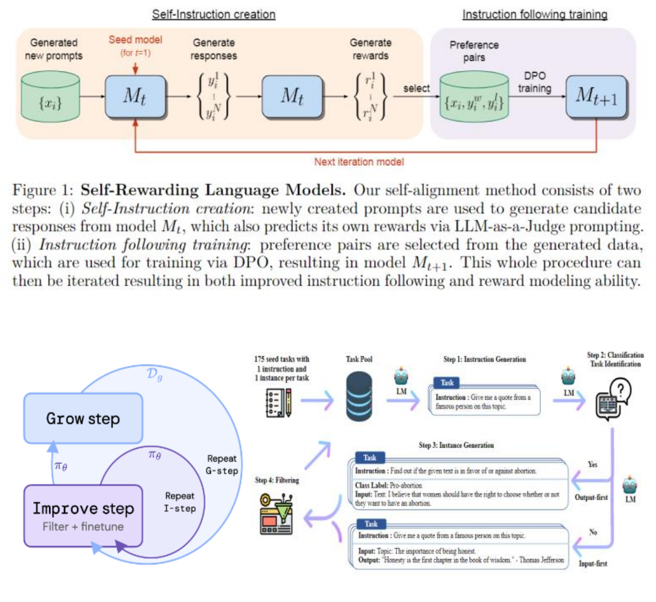
- Self-interaction among LLM agents gains popularity



Inter/intra LLM-agent communication



In-Context Self-refinement [3]
Self-reflection [5], etc

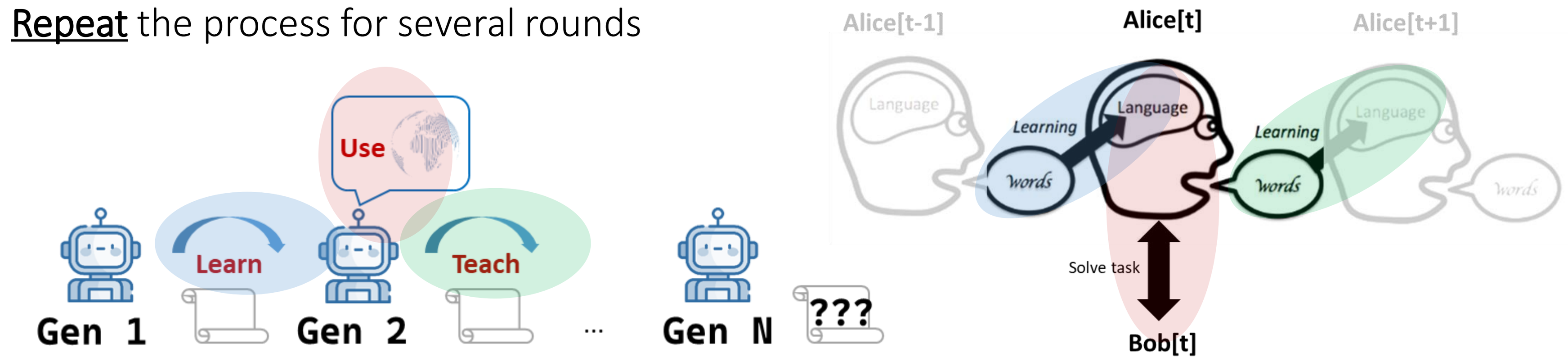


Self-reward [2], Self-instruction [4]
Multi-gen RL [1], etc

[1] Gulcehre, Caglar, et al. "Reinforced self-training (ReST) for language modeling." arXiv 2023.
[2] Yuan, Weizhe, et al. "Self-rewarding language models." arXiv preprint arXiv 2024.
[3] Madaan, Aman, et al. "Self-refine: Iterative refinement with self-feedback." NeurIPS 2023
[4] Wang, Yizhong, et al. "Self-Instruct: Aligning Language Models with Self-Generated Instructions." ACL 2023
[5] Gou, Zhibin, et al. "CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing." ICLR 2024

2. Similarity to Human Language's Evolution

- Although proposed by different reasons, they are similar in:
  - Imitation: Another agent learn from the message generated by previous agent
  - Interaction: LLM interact with other or environment to refine the knowledge
  - Transmission: LLM generate message based on given prompts
  - Repeat the process for several rounds



- But, keeps self-boosting introduce RISKs of
  - Diversity decrease
  - Mode collapse
  - Harmful bias amplification

Although reported in many related works sporadicly, no unified framework to analyze the asymptotic behavior.

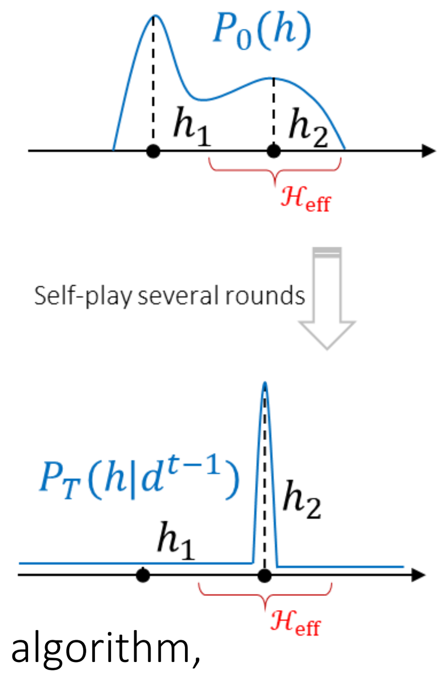
3. Bayesian Iterated Learning

- Bayesian-iterated learning framework:

Input (prompt): $x \in \mathcal{X}$ Data pair: $d = (x, y)$ Prior: $P_0(h)$

Response: $y \in \mathcal{Y}$ Hypothesis: $\boldsymbol{h} \in \boldsymbol{\mathcal{H}}: \boldsymbol{\mathcal{X}} \to \boldsymbol{\mathcal{Y}}$ Posterior: $P_{\text{lm}}(h|\boldsymbol{d})$

- Imitation phase:
  Start from $P_0(h)$, learn $\mathbf{d}^{\mathbf{t-1}}$, becomes $P(h\,|\mathbf{d}^{\mathbf{t-1}})$
- Interaction phase:
  Conduct task, posterior then $\propto \mathbb{I}(h \in \mathcal{H}_{\text{eff}})P(h \mid \mathbf{d}^{\mathbf{t-1}})$
- Transmission phase:
  $\mathbf{d}^{\mathbf{t}} \sim P(d \mid h^*)$; $h^* = \text{argmax}_{h \in \mathcal{H}_{\text{eff}}} P(h\,|\mathbf{d}^{\mathbf{t-1}})$
- Theoretical guarantee:
  $P(h\,|\mathbf{d}^{\boldsymbol{T}}) \to \mathbb{I}(h = h^{T*})$; $h^{T*} = \text{argmax}_{h \in \mathcal{H}_{\text{eff}}} P_0(h)$



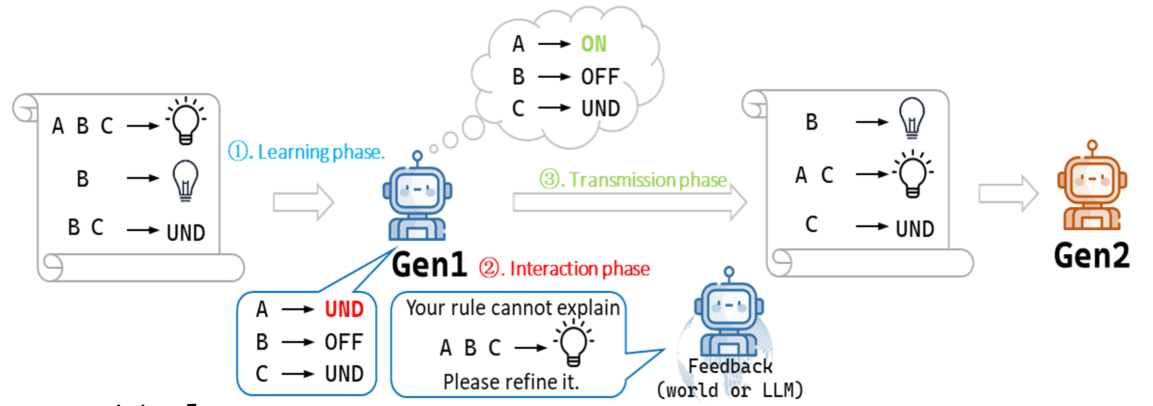
- Proof sketch in Appendix A and [1]:
  Recall the proof of standard EM (Expectation-Maximization) algorithm, replace (θ, z) to $(h, \boldsymbol{d})$, marginalize the input variable x. Done!

  Key assumption to LLM: ICL is implicit Bayesian Inference [2]

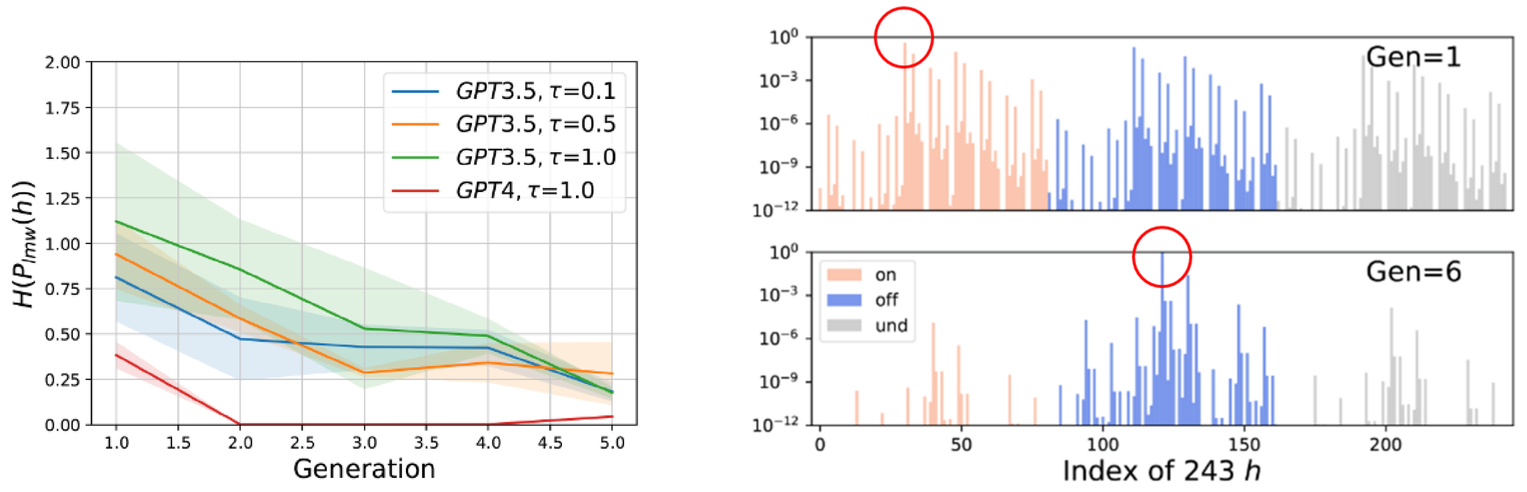
[1] Griffiths, Thomas L et.al . "Using category structures to test iterated learning as a method for identifying inductive biases." Cognitive Science 2008.
[2] Xie, Sang Michael, et al. "An Explanation of In-context Learning as Implicit Bayesian Inference." ICLR-2022

4. LLM Verification – Explicit h

- To verify the subtle trends predicted by the theory, start from Abstract Causal REasoning ACRE, used in [1]



- Consider 5 objects, then $3^5 = 243$ possible $\boldsymbol{h}$



Verify convergence: $P(\boldsymbol{h}|\boldsymbol{d}^t) \to \mathbb{I}(h = h^{T*})$ Verify solution: $h^{T*} = \text{argmax}_{h \in \mathcal{H}_{\text{eff}}} P_0(h)$

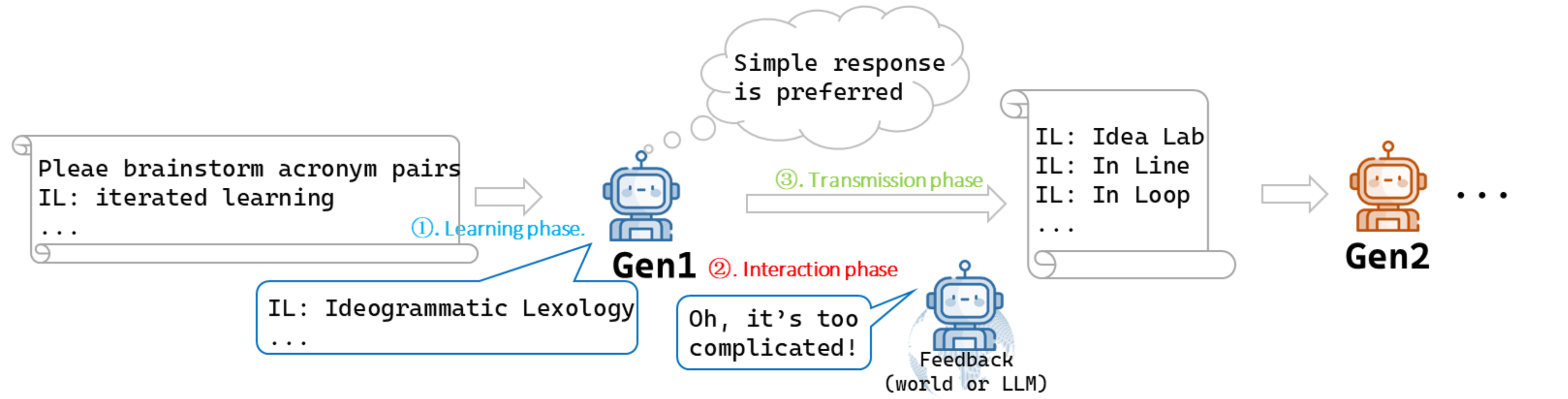
[1] Qiu, Linlu, et al. "Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement." ICLR-2024

5. LLM Verification – Implicit h

- Consider a more practical self-data augmentation problem, where $\boldsymbol{h}$ is implicit, e.g.,

$\boldsymbol{h} = \{\text{Long response, Short response}\}$; $\boldsymbol{h} = \{\text{Use easy words, Use hard words}\}$

- A simple "creative writting"-style game, brainstorming the given acronym



- LLM naturally bias towards common & short words. Manipulate it using different $\boldsymbol{\mathcal{H}}_{\text{eff}}$

Table 2: Results when adding different $\mathcal{H}_{\text{eff}}$. We color the highest and lowest numbers in each column. N_e is the number of easy examples in $\mathbf{d}^0$. Results under different settings are in Table 4 and 5.

	Ratio-easy					Avg-rank					Avg-length				
N_e=	2	4	6	8	10	2	4	6	8	10	2	4	6	8	10
Random	0.913±0.01	0.600±0.08	0.963±0.00	0.887±0.03	0.825±0.06	13519	27269	7487	10425	15871	5.425±1.04	4.825±0.33	5.600±1.55	5.014±1.50	4.713±0.63
Imitation-only	0.438±0.20	0.935±0.01	0.925±0.00	0.975±0.00	0.963±0.00	35235	7497	9081	5549	8075	4.450±0.86	4.387±1.40	4.175±0.13	4.188±0.65	5.438±1.24
Hard	0.219±0.19	0.250±0.43	0.450±0.43	0.338±0.16	0.500±0.23	49869	46436	37288	41255	31903	4.630±1.54	5.788±1.39	4.675±0.40	4.388±0.60	5.200±0.42
Easy	0.763±0.17	1.000±0.00	0.988±0.00	1.000±0.00	0.990±0.00	15910	3156	2383	2924	2650	3.925±0.33	5.263±0.06	4.713±0.06	4.240±0.08	4.893±0.71
Easylong	0.988±0.00	0.975±0.00	0.988±0.00	0.988±0.00	1.000±0.00	7063	9413	8649	6898	7404	5.209±0.41	5.888±0.52	6.838±1.10	6.979±1.57	7.695±1.70
Easyshort	1.000±0.00	1.000±0.00	0.975±0.00	1.000±0.00	0.988±0.00	5671	4223	5733	4502	5251	3.975±0.50	4.012±1.03	4.374±0.50	3.950±0.03	4.250±0.24

6. Take-away Message

- Applying Bayesian-IL to LLM's evolution:
  1. Bias in $\boldsymbol{P_0(h)}$ is guaranteed to be amplified if self-boosting too much
  2. Bias can be beneficial or harmful, $\boldsymbol{h}$ can be explicit or implicit
  3. Figure out the bias, understand it, and then design corresponding $\boldsymbol{\mathcal{H}_{\text{eff}}}$
- Iterated learning and $P_0(h)$ in other fields:
  - [1] CogSci: human prefer compositionality → compositional language is achieved after IL
  - [2] EmCom: simple NN prefer compositionality → compositional mapping is achieved after IL
  - [3] Representation Learning: complex NN prefer systematicness → systematical generalization
  - [4] VLM: language prefer compositionality → vision modual also becomes compositional after IL
- In-weights updates (e.g., DPO) amplify the bias in $P_0(h)$ more
  - [5] Analysis of the "squeezing effect" caused by negative gradient part in DPO

[1] Kirby, Simon, et.al. "Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language." PNAS 2008
[2] Ren, Yi, et al. "Compositional languages emerge in a neural iterated learning model." ICLR 2020
[3] Ren, Yi, et al. "Improving compositional generalization using iterated learning and simplicial embeddings." NeurIPS 2023
[4] Zheng, Chenhao, et al. "Iterated learning improves compositionality in large vision-language models." CVPR 2024
[5] Ren, Yi, et. al. "Learning Dynamics of LLM Finetuning", Submitted to ICLR 2025