

On the Generalization Gap in Self-Evolving Language Model Reasoning

Cyrus Rashtchian¹

Joint work with

Zhenting Qi^{1,3} (lead author)

Susanna Maria Baby², Stefanie Anna Baby², Kan Yuan²,
Da-Cheng Juan¹, Tu Vu^{1,4}, Andrew Tomkins¹

¹ Google Research

² Google

³ Harvard University

⁴ Virginia Tech

Guiding Research Question

With only a base model and unlabeled prompts, how close can **internally generated supervision** get to **oracle-level, ground-truth post-training**?

Knights and Knaves

Example Problem. A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie.

Knights and Knaves

Example Problem. A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie.

You meet 2 inhabitants: Zoey, and Oliver.

Zoey remarked, "Oliver is not a knight".

Oliver stated, "Oliver is a knight if and only if Zoey is a knave".

So who is a knight and who is a knave?

Knights and Knaves

Example Problem. A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie.

You meet 2 inhabitants: Zoey, and Oliver.

Zoey remarked, "Oliver is not a knight".

Oliver stated, "Oliver is a knight if and only if Zoey is a knave".

So who is a knight and who is a knave?

Solution

- (1) Zoey is a knave
- (2) Oliver is a knight

Knights and Knaves

Example Problem. A very special island is inhabited only by knights and knaves. Knights always tell the truth, and knaves always lie.

You meet 2 inhabitants: Zoey, and Oliver.

Zoey remarked, "Oliver is not a knight".

Oliver stated, "Oliver is a knight if and only if Zoey is a knave".

So who is a knight and who is a knave?

Solution

- (1) Zoey is a knave
- (2) Oliver is a knight

Experimental Setup

Difficulty Levels

- 2–3 People = **Easy**
- 4–5 People = **Medium**
- 6–8 People = **Hard**

Methodology

Train on **2–3**; Test on **all**

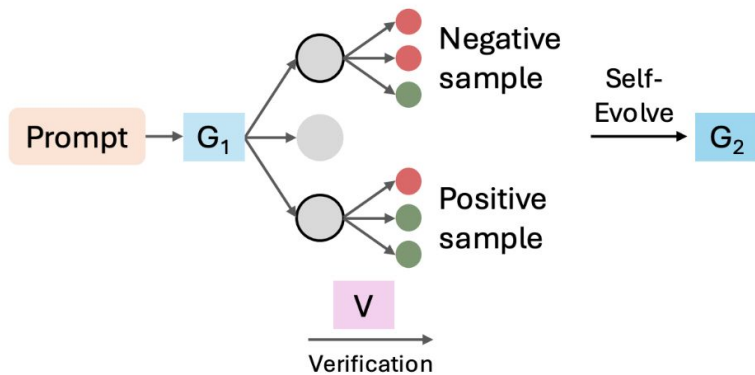
Test Set Details

800 examples, unique solutions
Split by num_ppl in **[2, 3, 4, ..., 8]**

Taxonomy of Closed-Loop Methods

Single-Round Self-Evolution

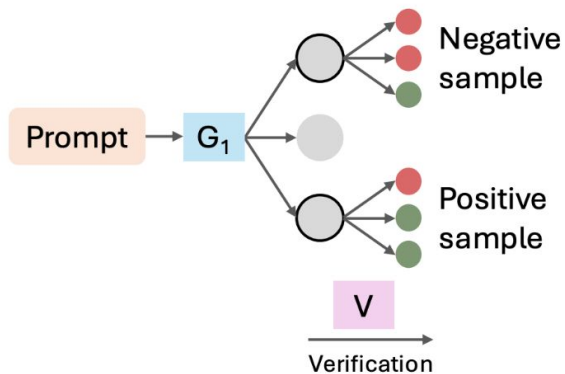
Simple SE



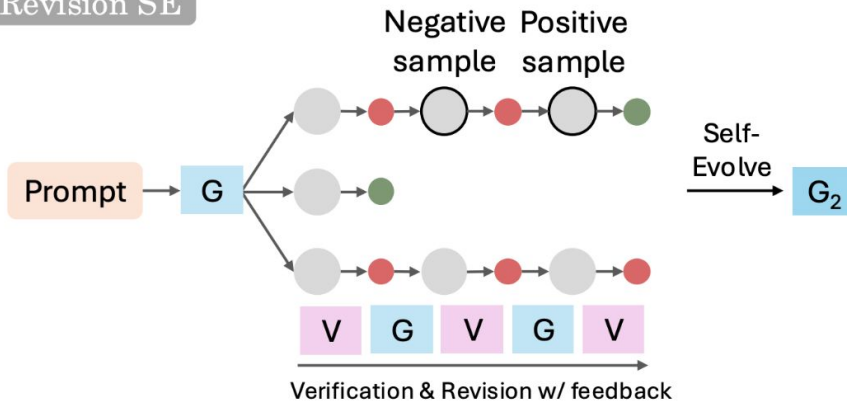
Taxonomy of Closed-Loop Methods

Single-Round Self-Evolution

Simple SE

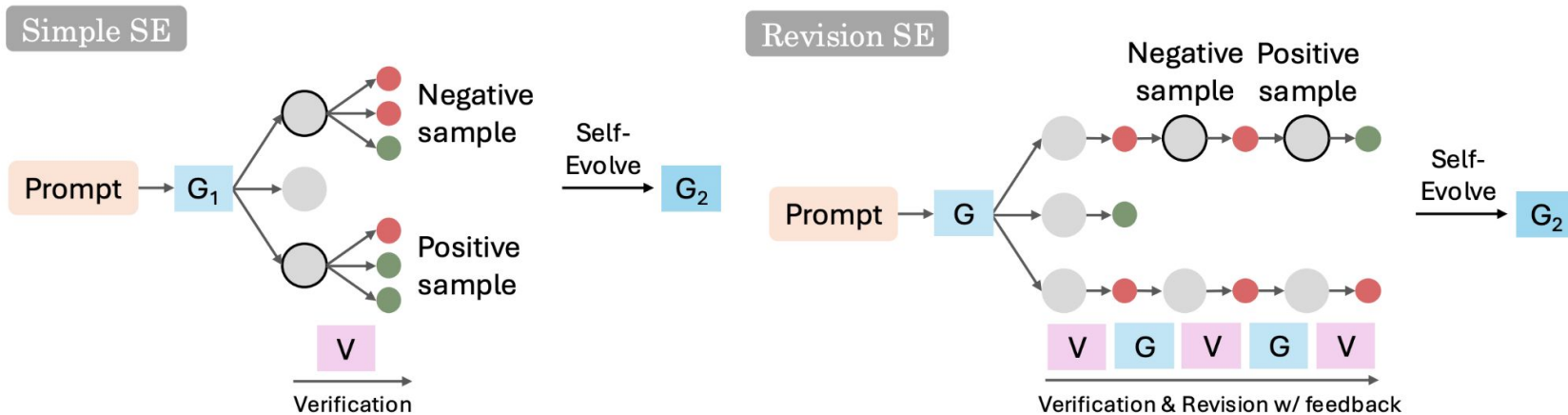


Revision SE

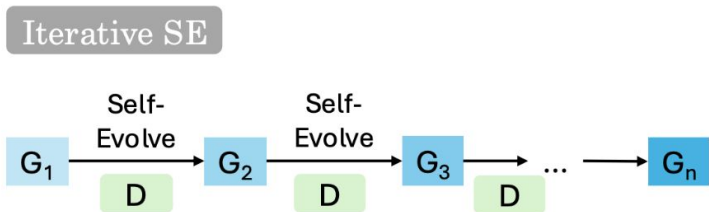


Taxonomy of Closed-Loop Methods

Single-Round Self-Evolution

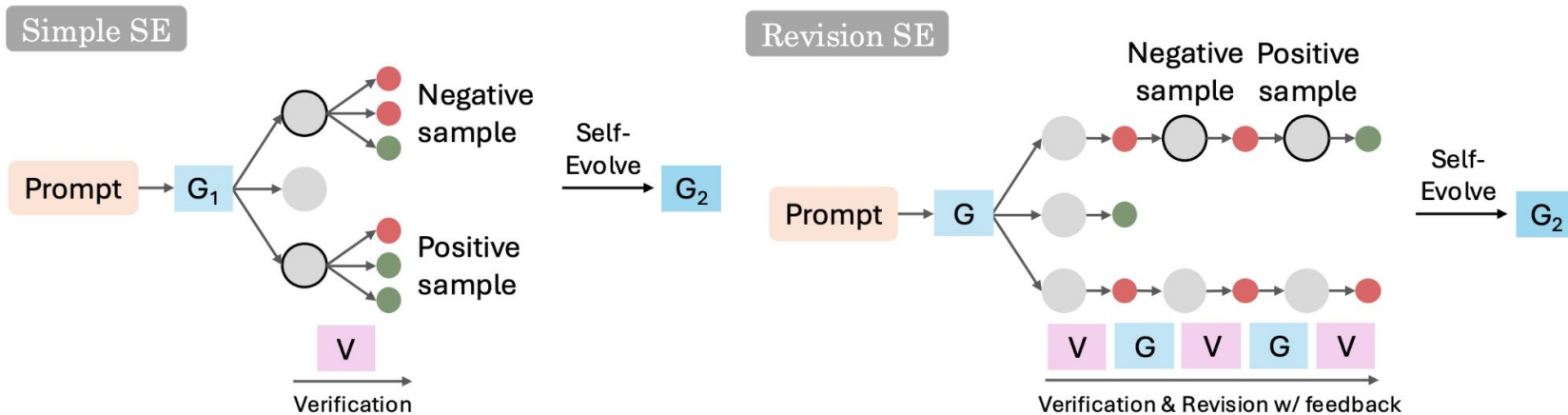


Multi-Round Self-Evolution

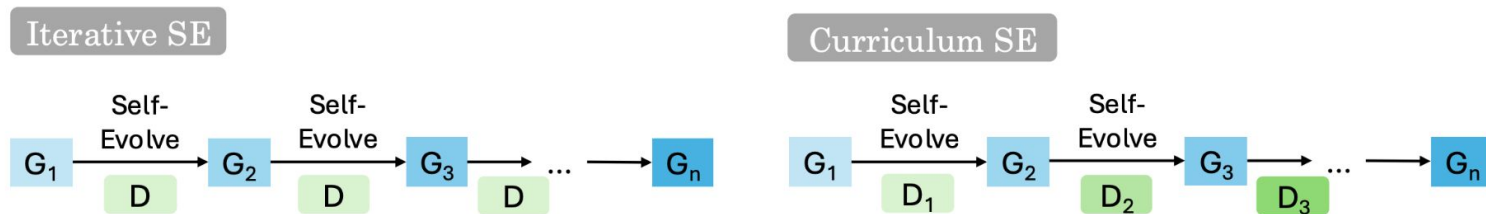


Taxonomy of Closed-Loop Methods

Single-Round Self-Evolution



Multi-Round Self-Evolution



Self-Evolution works...

Experimental Setup

Methodology

Train on **2-3**; Test on **all**

Report best threshold for
positive/negative DPO samples

Results

- SimpleSE improves accuracy
- 12B gets close to matching the 'roofline' of the 27B model

Self-Evolution works...

Experimental Setup

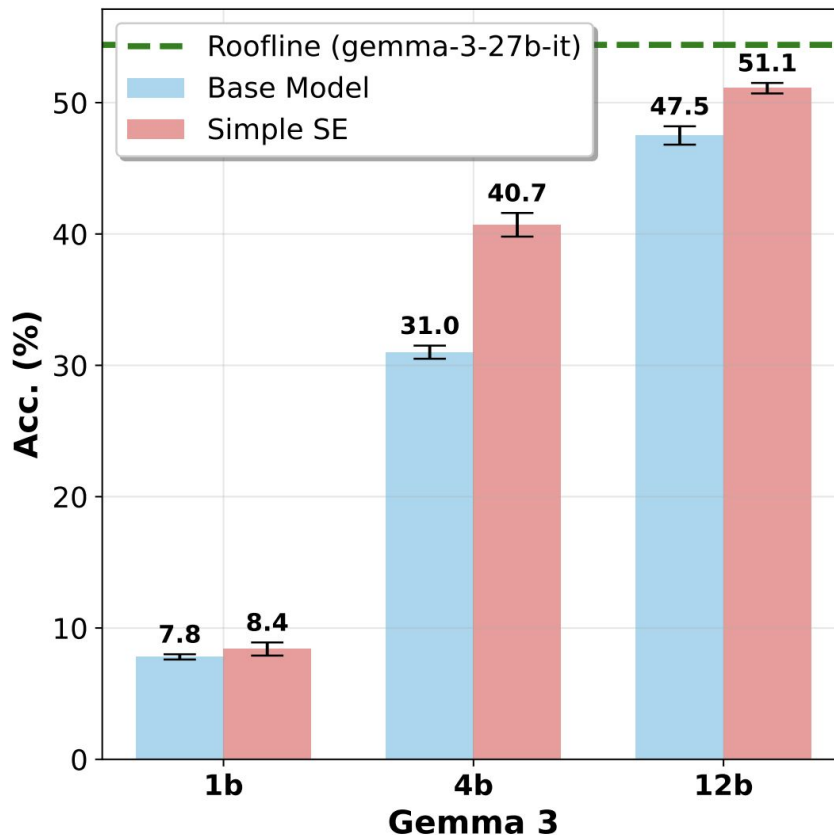
Methodology

Train on **2-3**; Test on **all**

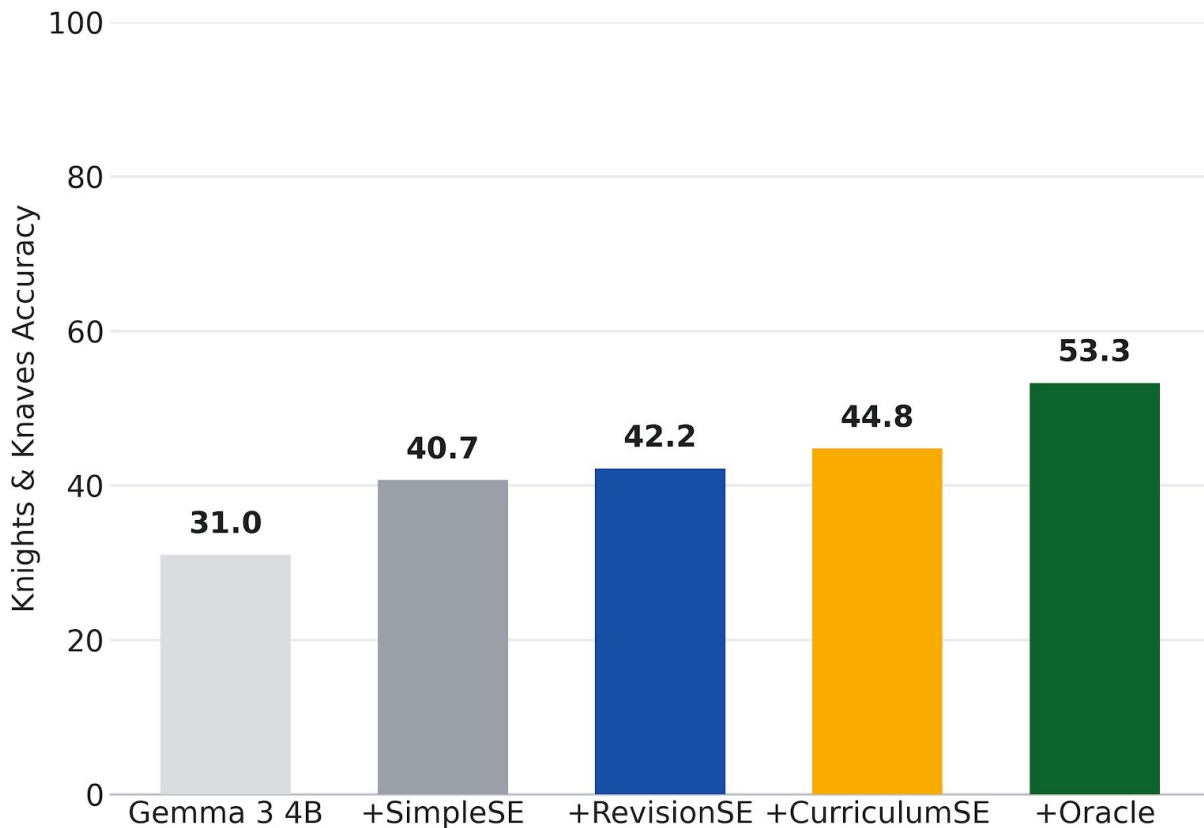
Report best threshold for
positive/negative DPO samples

Results

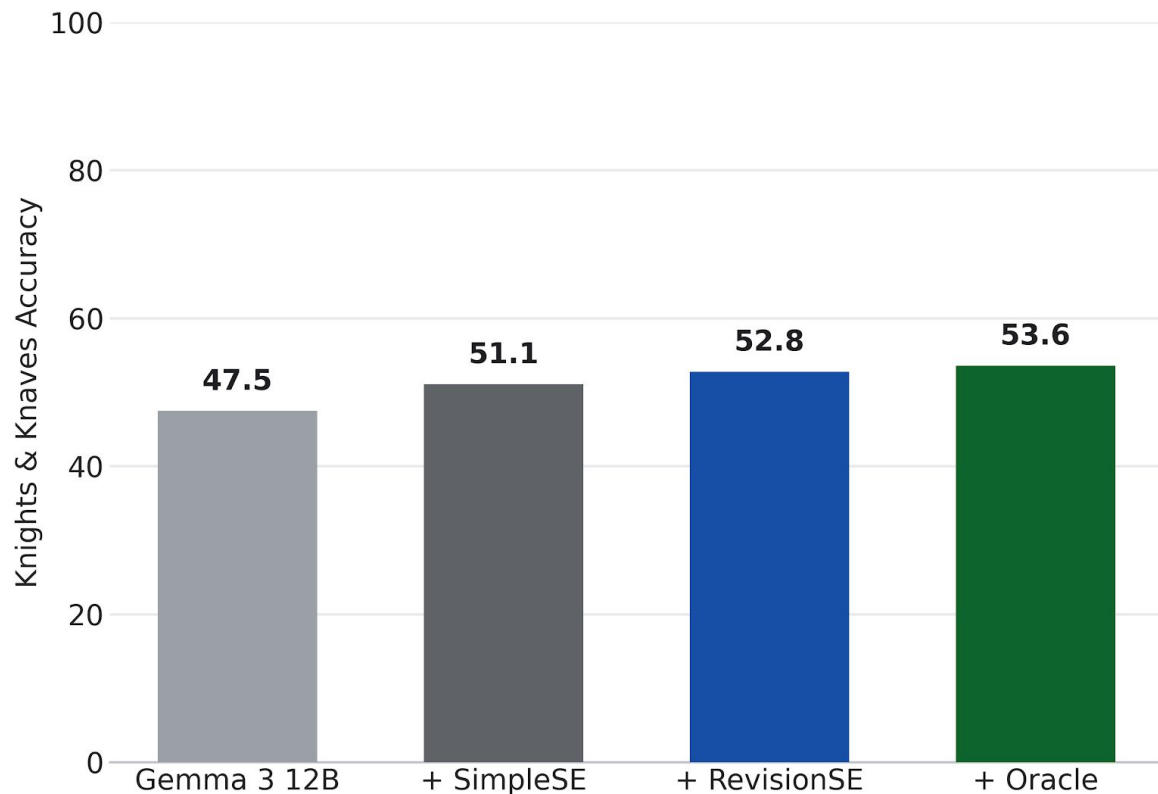
- SimpleSE improves accuracy
- 12B gets close to matching the 'roofline' of the 27B model



... But **Gemma 3 4B** has a **persistent gap** to oracle labels



RevisionSE + Gemma 3 12B *almost* Matches Oracle Labels



Confidence Improves More than Core Abilities

Takeaway: models may not learn new reasoning abilities through self-evolution

Dataset	Model/Algorithm	Pass@1	Pass@32
KK	Gemma 3 4B	31.0	78.0
KK	+ SimpleSE	40.7	77.4
KK	+ SimpleSE (2 iterations)	43.3	78.2
KK	+ SimpleSE (3 iterations)	44.1	78.8

Summary of Main Findings

Systematic Analysis of Closed-Loop SE

Unified comparison of SE methods under strict frameworks on **Knights-and-Knaves** benchmark to enable controlled study.

Internal Supervision Quality Matters

Multi-turn feedback leads to **52.8% acc.** for 12B model, nearly matching oracle based 53.6% (+ high computational cost).

Gains Come from a Confidence Boost

Digging deeper, we find that self-evolution leads to higher accuracy **but it does not improve pass@k or OOD generalization.**

Persistent Gap to Oracle Supervision

Internal feedback is not a substitute for verifiable rewards. SE surfaces latent abilities but does not teach new skills.

Thank you.

Corresponding Authors:

Cyrus Rashtchian

cyroid@google.com

Zhenting Qi

zhentingqi@g.harvard.edu