Self-evolving AI for causal estimator discovery

Yiqun Chen (yiqunc@jhu.edu)
assistant professor of biostatistics and computer science
AI for Science Workshop, AAAI 2026

A crash course on causal effect estimation

Causal inference is central to scientific and program evaluation:

- Is the policy using AI intervention for smoking cessation effective?
- What's the individual effect of this drug on blood pressure?
- ...

A crash course on causal effect estimation

Causal inference is central to scientific and program evaluation:

- Is the policy using AI intervention for smoking cessation effective?
- What's the individual effect of this drug on blood pressure?
- ...

Unit	T	Y(0)	Y(1)	$\tau = Y(1) - Y(0)$
1	1	?	4.2	?
2	0	2.8	?	?
3	1	?	5.1	?
4	0	3.5	?	?

Unlike prediction problems, we do NOT get to observe the treatment effect ground truth.

A crash course on causal effect estimation

In practice, **all causal estimators** are evaluated on (semi-)synthetic datasets where we simulate and know the ground truth treatment effect $\tau_i = Y_i(1) - Y_i(0)$ for individual i.

Real Data

Unit	T	Y(0)	Y(1)	τ_i
1	1	?	4.2	?
2	0	2.8	?	?
3	1	?	5.1	?
...	...	...	...	...

Simulated Data

Unit	T	Y(0)	Y(1)	τ_i
1	1	3.1	4.2	1.1
2	0	2.8	3.5	0.7
3	1	3.9	5.1	1.2
...	...	...	...	...

A crash course on causal effect estimation

In practice, **all causal estimators** are evaluated on (semi-)synthetic datasets where we simulate and know the ground truth treatment effect $\tau_i = Y_i(1) - Y_i(0)$ for individual i.

Real Data

Unit	T	Y(0)	Y(1)	τ_i
1	1	?	4.2	?
2	0	2.8	?	?
3	1	?	5.1	?
...	...	...	...	...

Simulated Data

Unit	T	Y(0)	Y(1)	τ_i
1	1	3.1	4.2	1.1
2	0	2.8	3.5	0.7
3	1	3.9	5.1	1.2
...	...	...	...	...

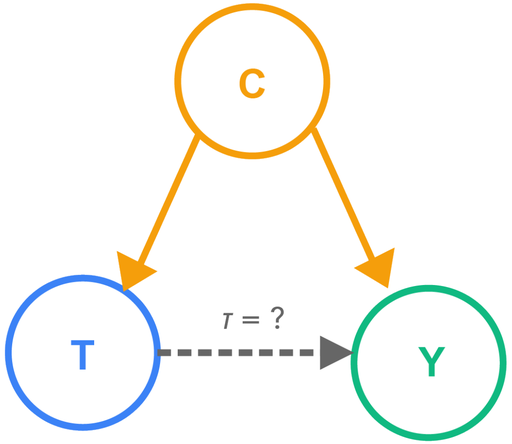
Common metrics: $\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\tau}_i - \tau_i\right)^2$ **(MSE);** $\frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\{\tau_i \in [L_i, U_i]\}$ **(coverage).**

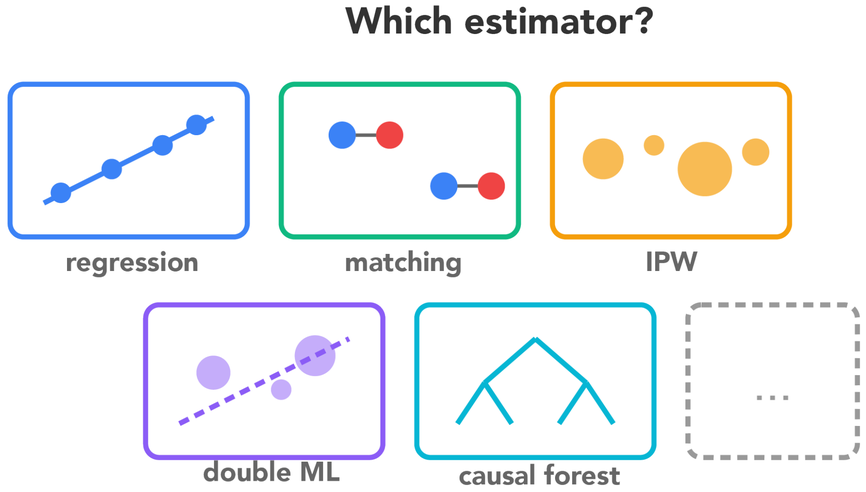
A crash course on causal effect estimation

Over the past decades, many flexible methods have been proposed to tackle causal effect estimation in *observational* data.

Practical challenges:

- What sets of confounders to adjust for?
- How do I select among the ocean of different estimators?



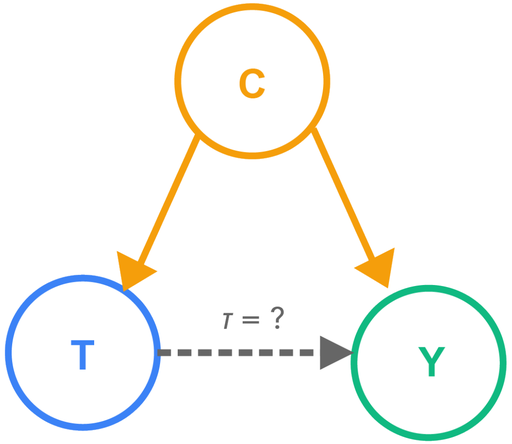


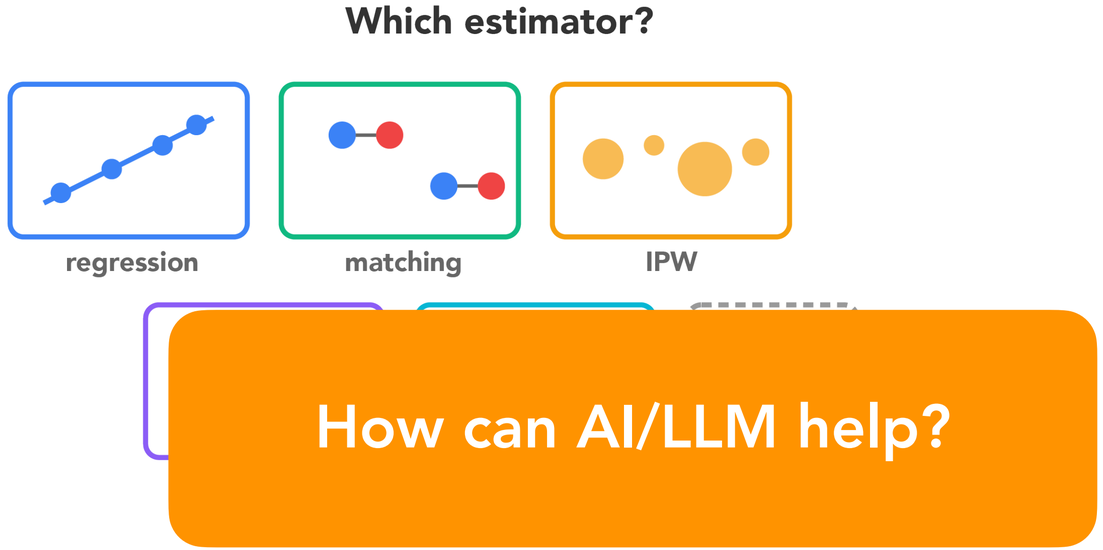
A crash course on causal effect estimation

Over the past decades, many flexible methods have been proposed to tackle causal effect estimation in *observational* data.

Practical challenges:

- What sets of confounders to adjust for?
- How do I select among the ocean of different estimators?





AI agents hold promise for scientific and algorithmic discovery

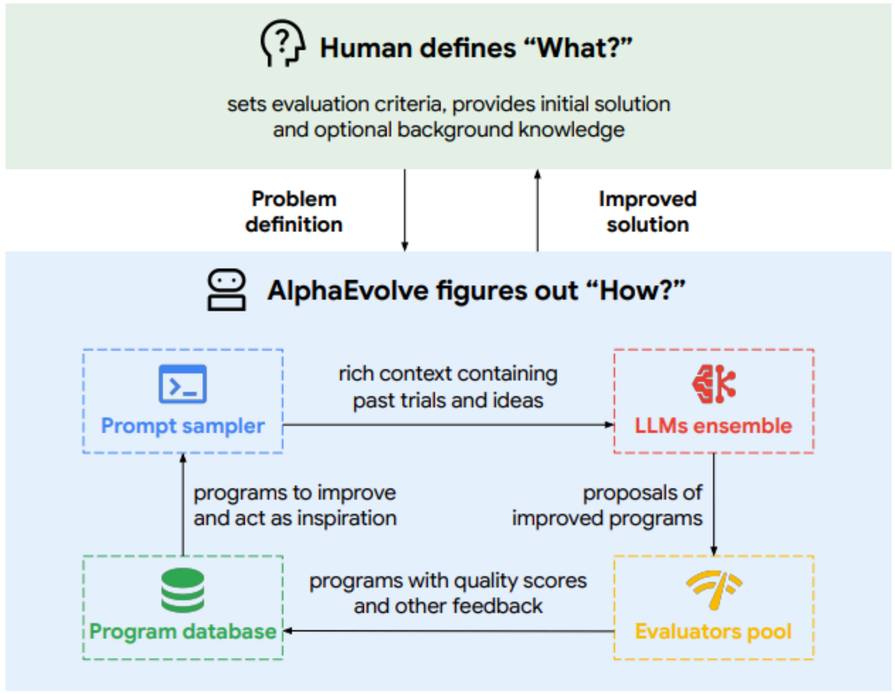
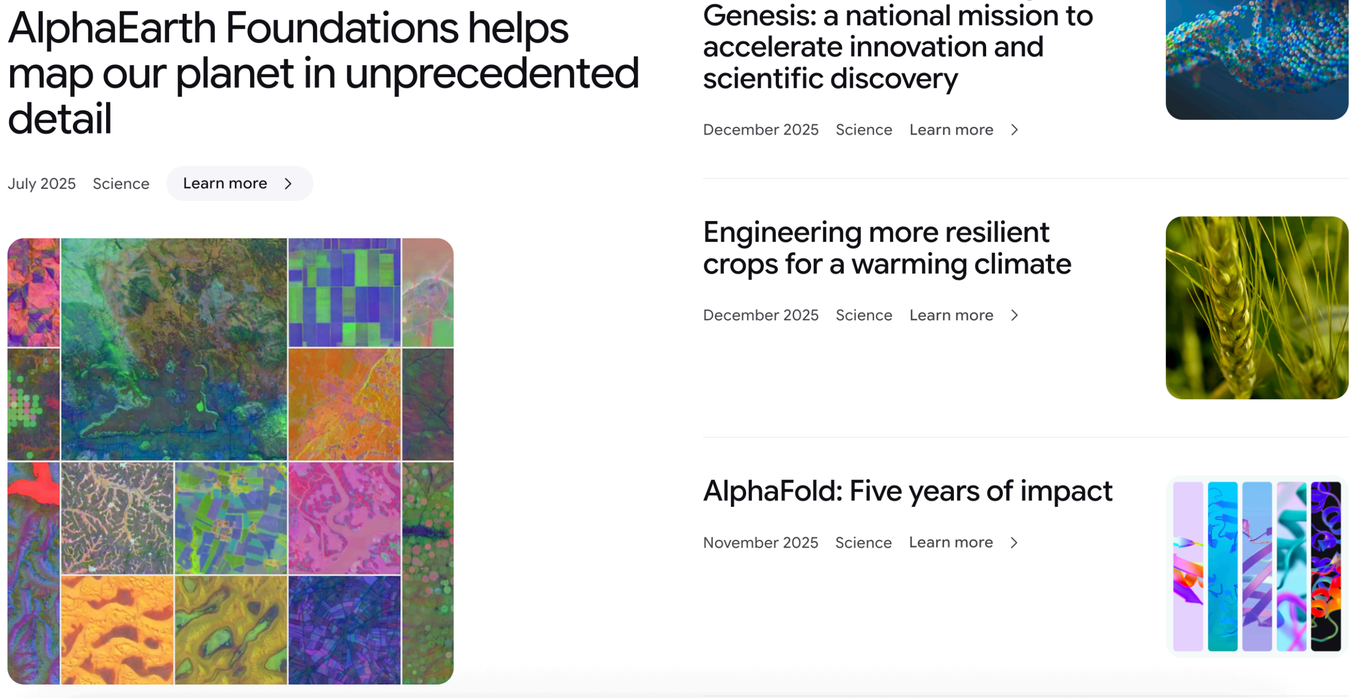
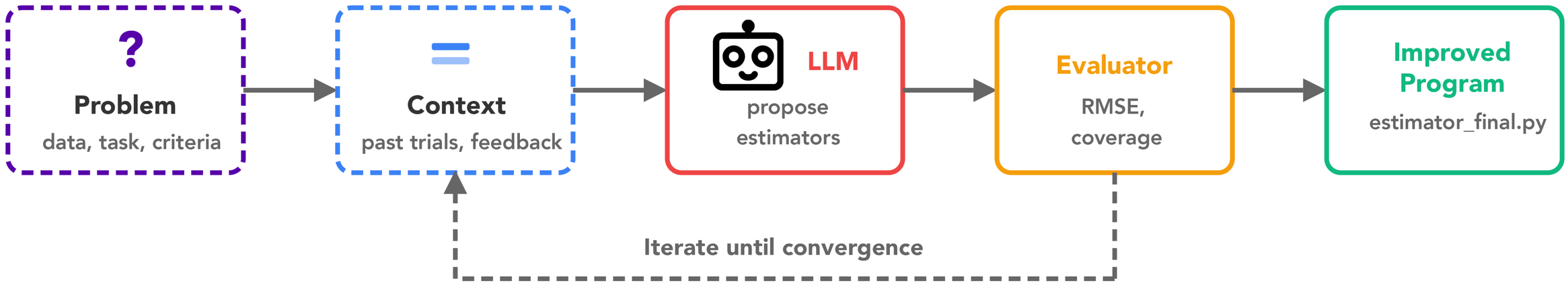


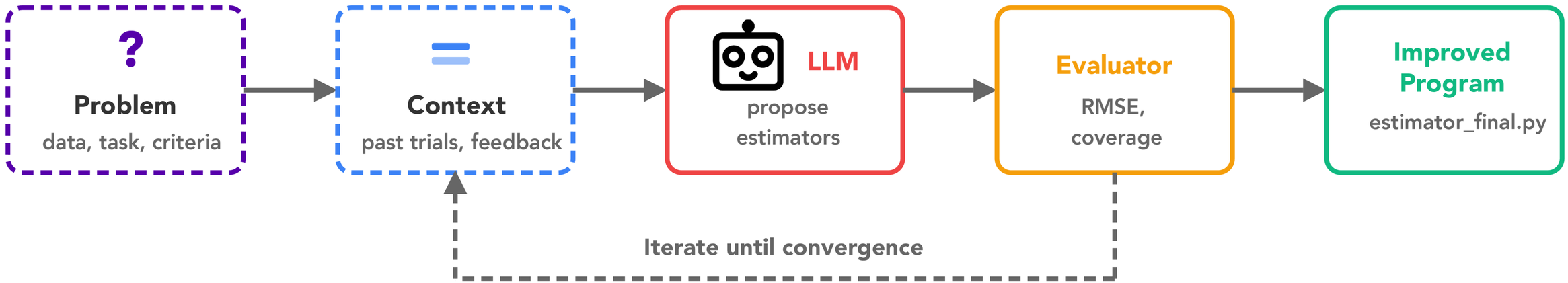
Figure 1 | *AlphaEvolve* high-level overview.



High level summary of current self-evolving agents



High level summary of current self-evolving agents



This framework has shown success in solving math problems, genomics data science, code optimization, …

This talk: apply self-evolving agents to causal effect estimation.

An example zero-shot prompt to seed the agent

ACIC Zero-shot Prompt

```
You are an expert causal inference statistician and Python programmer.

Background:
- We have multisite randomized trial data in long format (practice-year level).
- Each row is a practice in a given year, with aggregated covariates.
- Treatment is assigned at the practice level, randomized, and 'post' marks the
    period
  after treatment rollout.

Data columns:
  - 'id.practice' (int): site (practice) identifier
  - 'year' (int): panel time index
  - 'Y' (float): average monthly spending per patient in that practice-year
  - 'Z' (0/1): treatment indicator for the practice
  - 'post' (0/1): indicator for post-treatment period
  - 'n.patients' (int): number of patients represented in that row (weights)
  - Practice-level covariates: X1..X9
  - Aggregated patient covariates: V1_avg, V2_avg, V3_avg, V4_avg, V5_A_avg,
      V5_B_avg,
  V5_C_avg
```

A natural scoring rule for agent evolution

We used the openevovle framework, which takes a scalar score input:

$$S = \frac{1}{1 + \mathrm{RMSE} + \lambda \left| \mathrm{Coverage} - 90\,\% \right|}.$$

A natural scoring rule for agent evolution

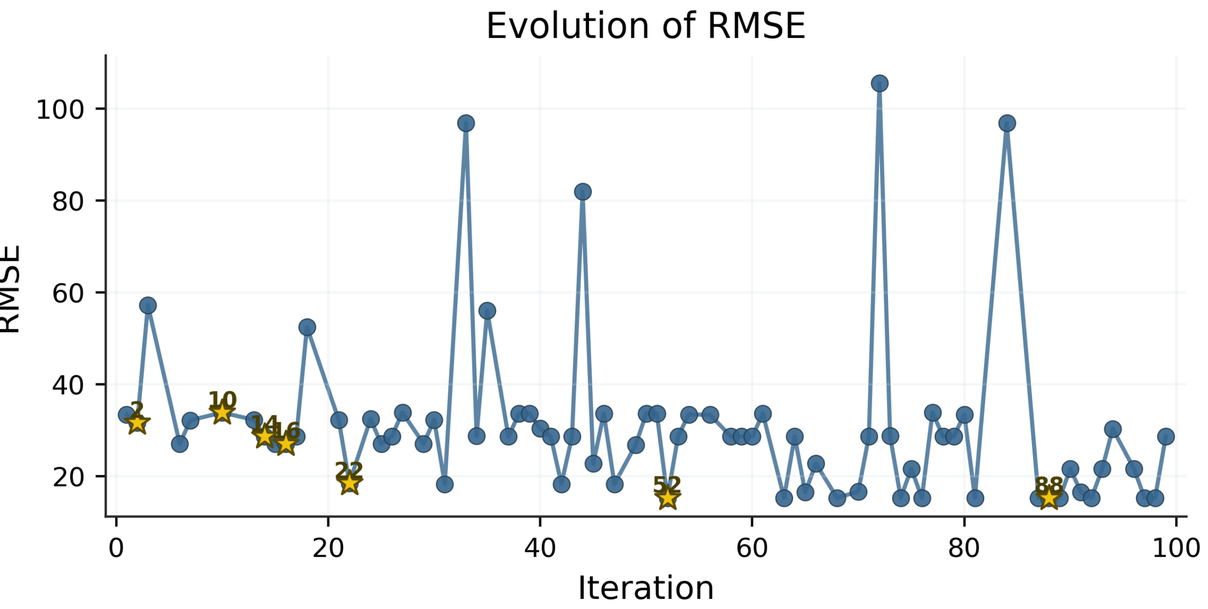
We used the openevovle framework, which takes a scalar score input:

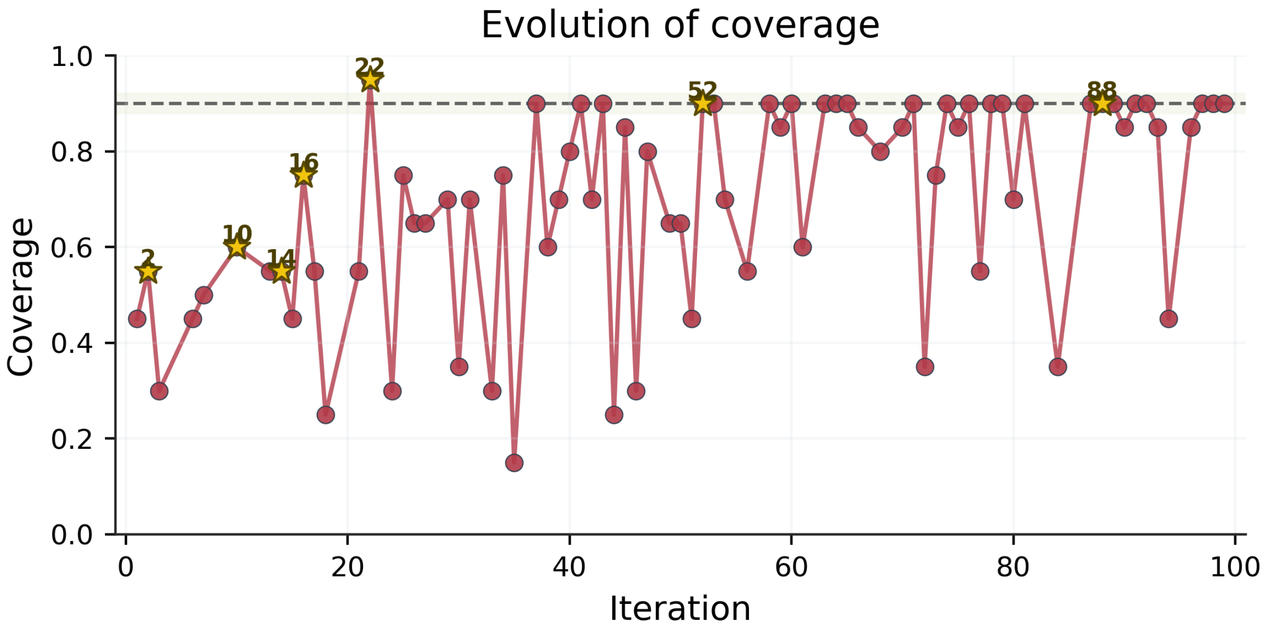
$$S = \frac{1}{1 + \mathrm{RMSE} + \lambda \left| \mathrm{Coverage} - 90\,\% \right|}.$$

At each iteration, we:

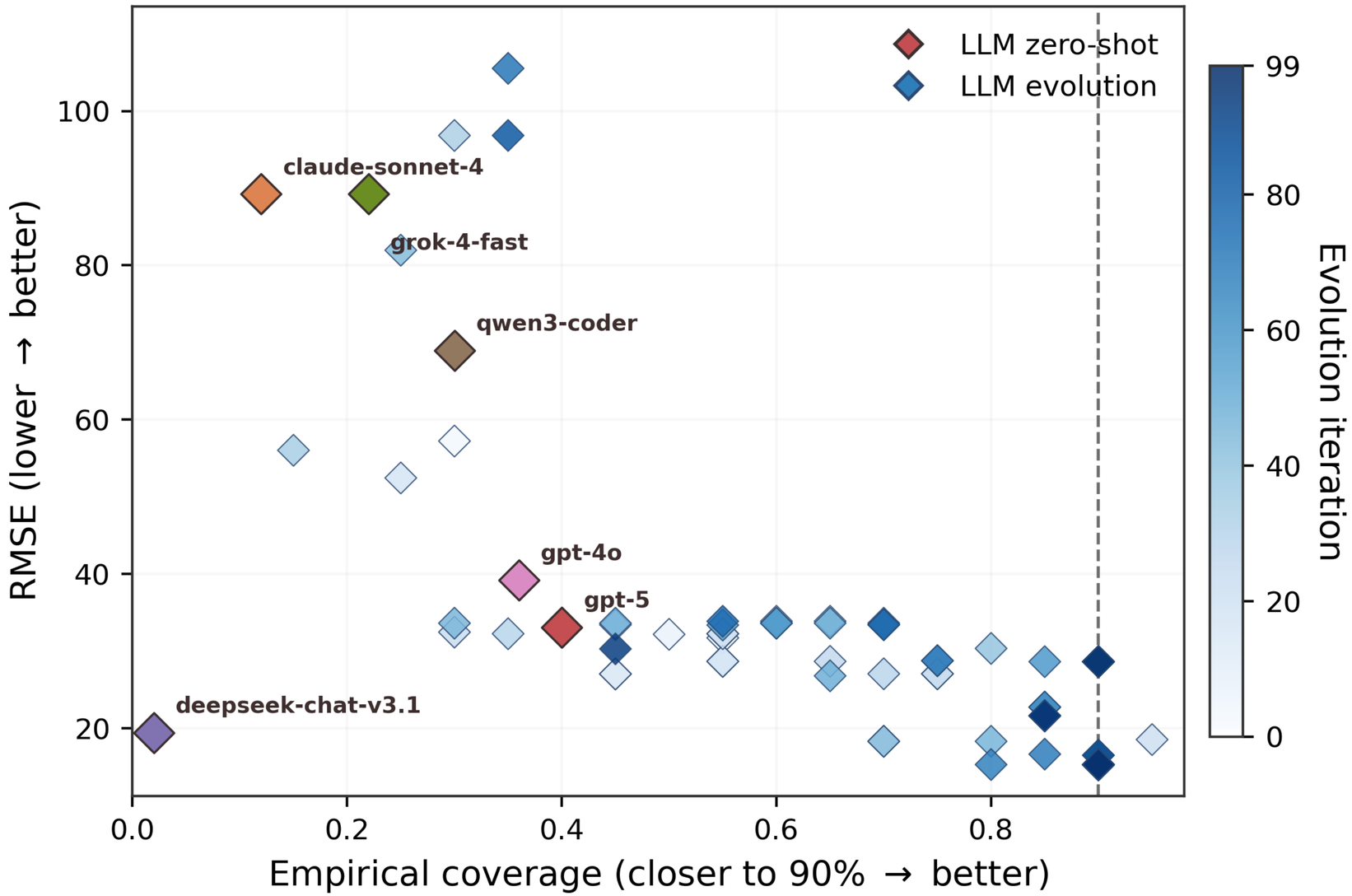
1. Run the estimator and compute RMSE, Coverage, and S on training data;
2. Openevovle then uses a pool of LLMs to refine the estimator (in py code) based on S by testing different code mutations;
3. We move in the direction of **higher** S and iterate until convergence (no changes or worse on validation set);
4. Report scores on **held-out test set**.

Results on the ACIC Benchmark

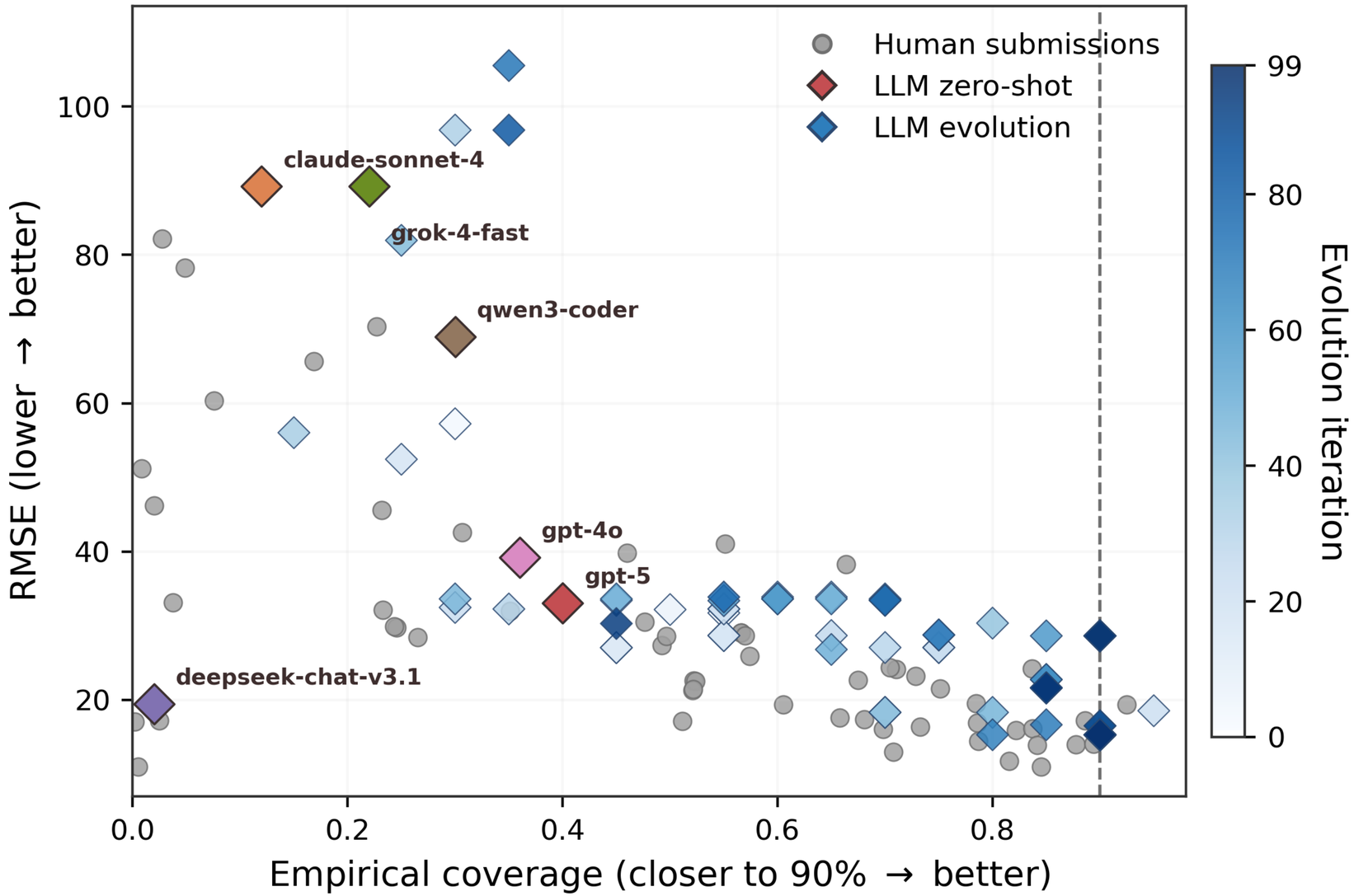




Self-evolution led to better solutions than zero-shot LLMs



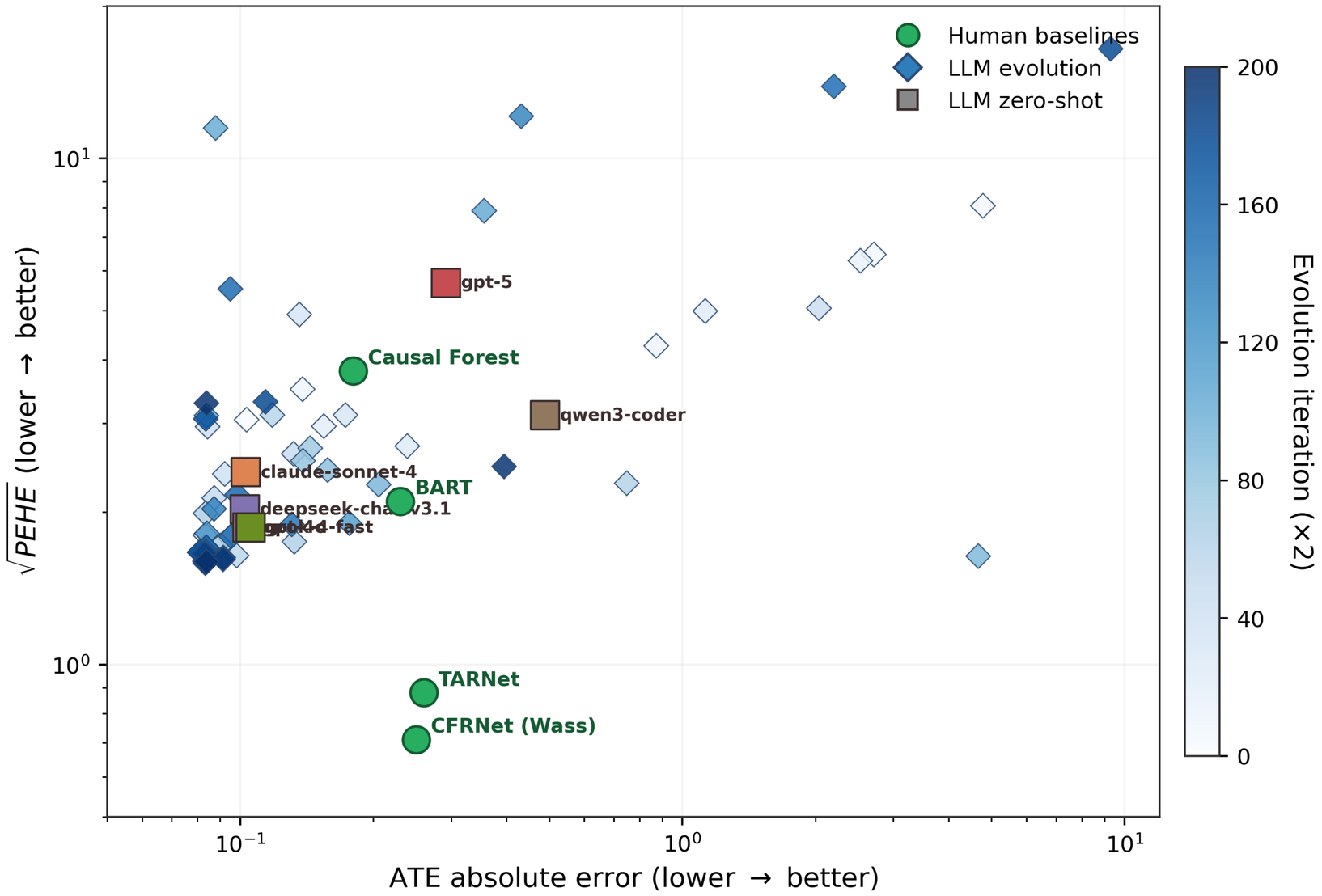
Self-evolution matched performance of human experts



What did self-evolution actually learn?

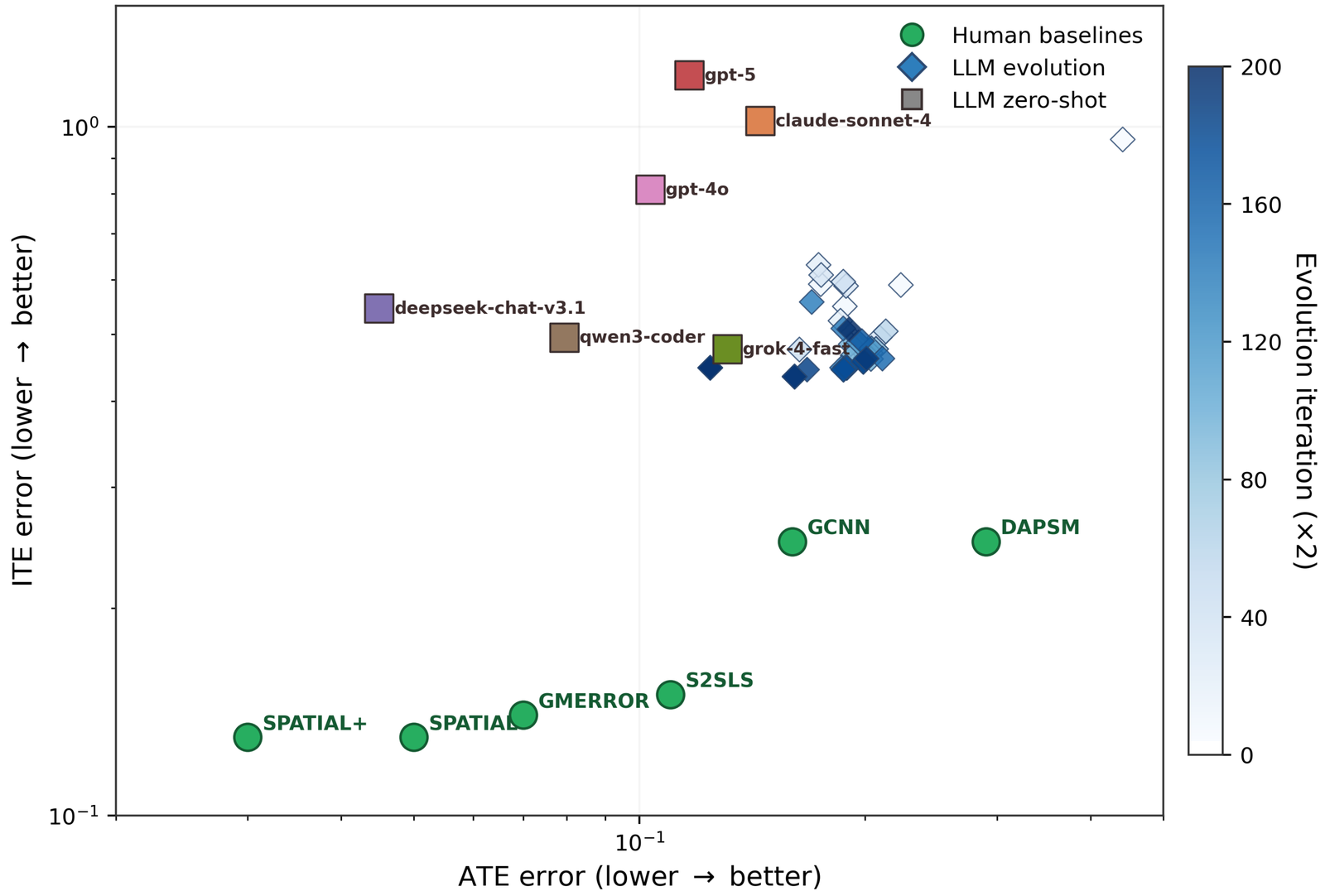
Version	Method highlights	Ridge	Covar. enc.	Pre-mean adj.	Trend adj.	Post-only fit	CI adj.
Baseline	TWFE with post-year interactions and cluster WLS	✗	✗	✗	✗	✗	✗
Iter. 2	Adds ridge regularisation and scaled covariates for stability	✓	✓	✗	✗	✗	✗
Iter. 10	Simplifies design, adds guard-rail fallbacks and CI inflation	✗	✗	✗	✗	✗	✓
Iter. 14	Restores covariate dummies and introduces pre-mean post adjustor	✗	✓	✓	✗	✗	✗
Iter. 16	Covariate-enriched TWFE with post-interacted pre means, cluster-t CI	✗	✓	✓	✗	✗	✓
Iter. 22	Works on post-period differences $Y_{\text{post}} - \bar{Y}_{\text{pre}}$	✗	✓	✓	✗	✓	✓
Iter. 52	Adds practice-level trend slopes and standardised covariates post-only	✗	✓	✓	✓	✓	✓
Iter. 88	Collapses to pooled post-period ATT with slope adjustment	✗	✓	✓	✓	✓	✓

Strong performance generalizes across datasets



On another benchmark data, self-evolution outperformed all human solutions on ATE error, but less so on individual accuracy (PEHE).

A straight-up application doesn't always match human SOTA



On a dataset with spatial confounding, AI solutions perform well on overall metric (ATE) but struggle with individual heterogeneity (ITE).

Zooming out: estimation is only one piece of the puzzle

A causal inference researcher will tell you that effect estimation is only a small part of any rigorous study:

1. **Causal graph & identification** — Can we even estimate the effect from this experimental setup?
2. **Estimation** — Statistical methods, point estimates, uncertainty
3. **Interpretation** — Sensitivity analysis, scientific plausibility, policy implications

Zooming out: estimation is only one piece of the puzzle

A causal inference researcher will tell you that effect estimation is only a small part of any rigorous study:

1. **Causal graph & identification** — Can we even estimate the effect from this experimental setup?
2. **Estimation** — Statistical methods, point estimates, uncertainty
3. **Interpretation** — Sensitivity analysis, scientific plausibility, policy implications

This work is led by my master student collaborator Can Wang; we are looking for collaborations/visiting students — email me at yiqunc@jhu.edu if you wanna chat!