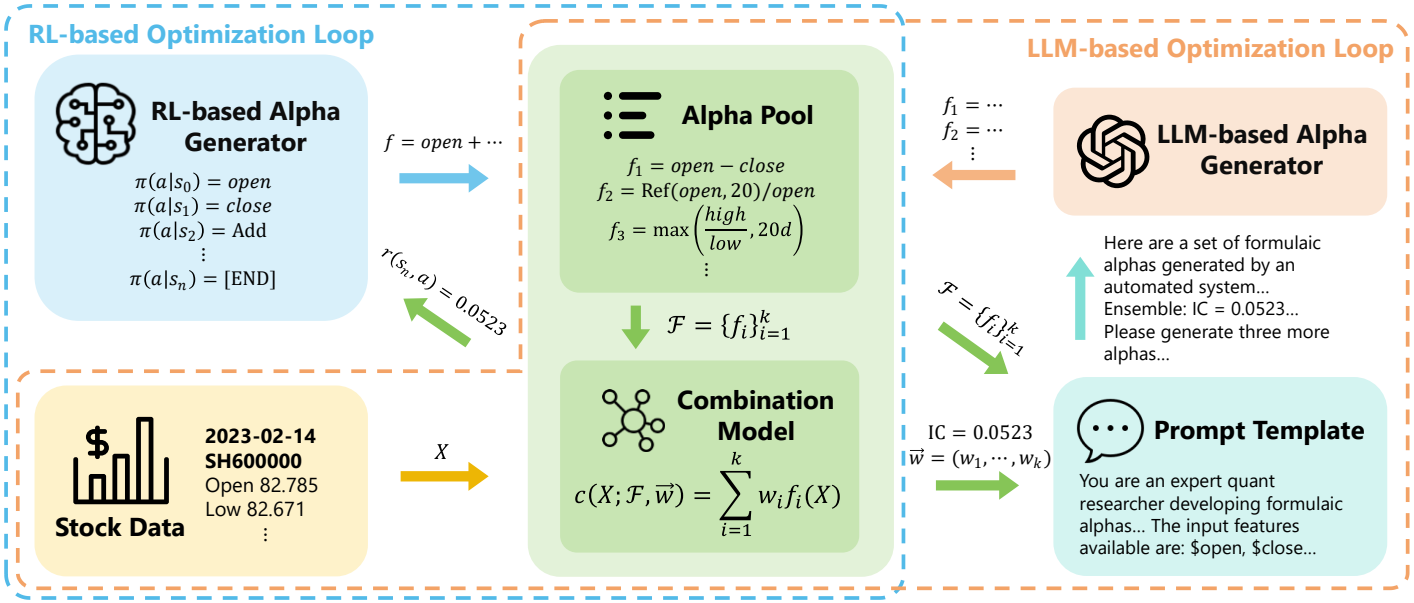
A Hybrid Approach to Formulaic Alpha Discovery with Large Language Model Assistance

Shuo YU, Hongyan XUE, Xiang AO, Qing HE

Frontiers of Computer Science, DOI: 10.1007/s11704-025-41061-5

Problems & Ideas

- Problems of traditional (GP- or RL-based) alpha mining approaches:
  - Alphas sometimes grow too complex, damaging their interpretability.
  - Complex alphas sometimes lead to unexpected overfitting.
- Ideas: Utilize inherent financial knowledge in LLMs to generate interpretable alphas and build a RL-LLM hybrid framework to leverage the advantages of both paradigms.

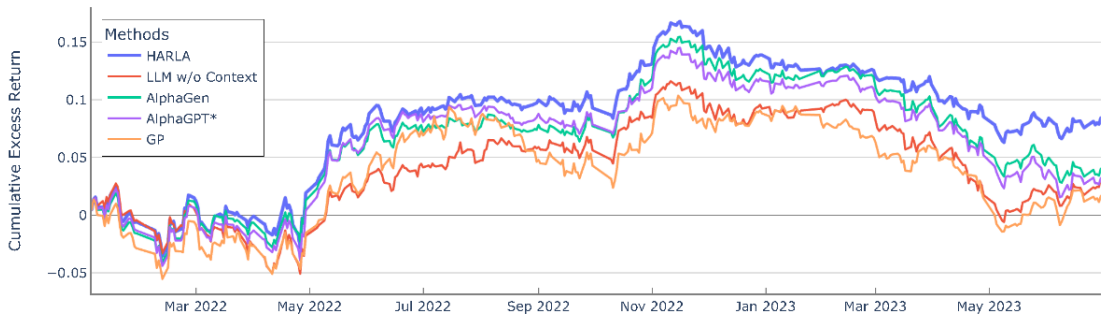


A diagram showing the working process of the RL-LLM hybrid alpha generator HARLA. The RL-based optimization loop and the LLM-based one share the same alpha pool to continuously and alternately improve upon it.

Main Contributions

- Contributions:
  - Utilizing LLMs for formulaic alpha generation, and leveraging their inherent financial knowledge to produce interpretable and diverse alpha expressions.
  - A hybrid optimization framework integrating LLM-generated alphas into an RL-based workflow. This includes scheduling and feedback mechanisms to maximize synergy between the two methods while managing computational overhead.
  - Extensive experiments showing the effectiveness of our approach.

Method	IC	ICIR	Rank IC	Rank ICIR
GP	-0.0048 (0.0169)	-0.0300 (0.1007)	-0.0154 (0.0238)	-0.0762 (0.1250)
AlphaGen	0.0288 (0.0090)	0.1957 (0.0650)	0.0321 (0.0101)	0.2238 (0.0809)
LLM Single Step	0.0182 (0.0185)	0.1224 (0.1358)	0.0283 (0.0211)	0.1933 (0.1594)
LLM Force Replace	0.0248 (0.0195)	0.1895 (0.1435)	0.0424 (0.0287)	0.2904 (0.1993)
LLM w/ Context	0.0301 (0.0181)	0.2165 (0.1247)	0.0423 (0.0230)	0.2963 (0.1572)
LLM w/ Updates	0.0335 (0.0118)	0.2414 (0.0818)	0.0456 (0.0182)	0.3122 (0.1109)
LLM w/o Context	0.0396 (0.0125)	0.2889 (0.0990)	0.0538 (0.0181)	0.3821 (0.1444)
AlphaGPT*	0.0307 (0.0153)	0.2224 (0.1250)	0.0321 (0.0222)	0.2387 (0.1832)
HARLA	**0.0515** (0.0152)	**0.3612** (0.1040)	**0.0572** (0.0161)	**0.4104** (0.1201)



Results of correlation analysis and investment simulations. Our hybrid method achieves an average information coefficient (IC) of 0.0515, a 75% improvement over the baseline RL framework. Backtest experiments reveal that our framework achieves a cumulative excessive return more than twice of the baseline.