AlphaBench: Benchmarking Large Language Models in Formulaic Alpha Factor Mining

Haochen Luo, Ho Tin Ko, Jiandong Chen, David Sun, Yuan Zhang*, Chen Liu*

香港城市大學 City University of Hong Kong

ICLR International Conference On Learning Representations

Background & Motivation

Formulaic Alpha Factor Mining (FAFM) is central to quantitative investment: interpretable formulas extract predictive signals from historical financial data, enabling systematic portfolio construction via long–short ranking strategies.

TRADITIONAL APPROACHES

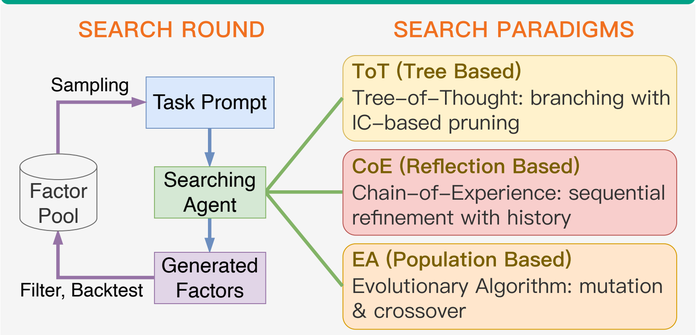
Human experts hand–craft formulas (e.g., Alpha101, Alpha158), limited by prior knowledge. ML methods (RL, genetic programming, symbolic regression) automate search but require heavy engineering.

LLMS AS A NEW PARADIGM

LLMs excel at symbolic reasoning and code generation, making them natural candidates for automated factor design. Recent work shows LLMs can produce interpretable alpha formulas with economic intuition.

AlphaBench: the first benchmark evaluating LLMs in factor generation, evaluation, and searching.

LLM + FAFM Workflow



Benchmark Scale

RESEARCH QUESTIONS

1. Can LLMs reliably generate valid and semantically correct alpha factors from NL?
2. Can LLMs evaluate factor quality without running costly backtests?
3. Which search paradigm (ToT / CoE / EA) best suits LLM-driven factor discovery?
4. How do model scale, reasoning (CoT), and prompting affect each task?

678 Generation Instructions

1170 Evaluation Instructions

3 Main Tasks

Factor Generation

Nature Language → Formula

Text2Alpha | Directional

Easy / Medium / Hard

Factor Evaluation

LLM as Judge (Backtest–free)

Ranking | Scoring | Atomic

Signal + pairwise

Factor Searching

Iterative search. CoE | ToT | EA

Overall Results

Generation

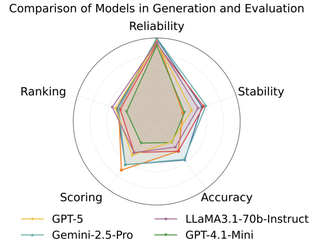
Model	Reliability		Stability		Accuracy		Overall	
	Vanilla	CoT	Vanilla	CoT	Vanilla	CoT	Vanilla	CoT
GPT-5	**1.00**	-	**0.62**	-	0.56	-	**0.72**	-
Gemini-2.5-Pro	0.98	-	0.33	-	0.44	-	0.58	-
Gemini-2.5-Flash	0.99	**0.99**	0.57	0.49	**0.57**	**0.58**	0.71	**0.69**
GPT-4.1-Mini	0.93	0.93	0.59	0.41	0.44	0.44	0.65	0.59
DeepSeek-V3	0.91	0.97	0.35	0.32	0.31	0.32	0.52	0.53
LLaMA3.1-70b-Instruct	0.95	0.94	0.52	0.44	0.38	0.47	0.62	0.61
DeepSeek-R1-Distill-Qwen-32B	0.35	0.58	0.19	0.24	0.14	0.14	0.23	0.32
Qwen2.5-14B-Instruct	0.79	0.58	0.50	**0.51**	0.34	0.46	0.54	0.52
LLaMA3.1-8b-Instruct	0.94	0.84	0.32	0.44	0.18	0.24	0.48	0.51
Gemini-1.5-Flash-8b	0.95	0.94	0.44	0.47	0.30	0.32	0.56	0.58

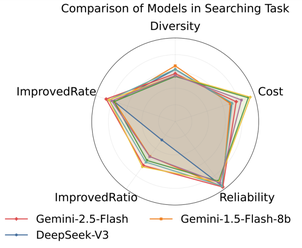
Evaluation

Model	Ranking	Scoring		Overall
	Precision	Signal ACC	MAE	
GPT-5	0.24 / -	0.32 / -	1.67 / -	0.47 / -
Gemini-2.5-Pro	0.24 / -	**0.36** / -	1.66 / -	**0.48** / -
Gemini-2.5-Flash	0.25 / 0.14	0.32 / **0.33**	1.67 / 1.64	0.47 / 0.44
GPT-4.1-Mini	0.23 / 0.24	0.23 / 0.20	1.57 / 1.59	0.44 / 0.43
DeepSeek-V3	0.19 / 0.17	0.16 / 0.17	1.60 / 1.57	0.40 / 0.40
LLaMA3.1-70b-Instruct	**0.28** / 0.26	0.23 / 0.24	1.62 / 1.61	0.45 / 0.45
DeepSeek-R1-Distill-Qwen-32B	0.20 / 0.20	0.24 / 0.23	1.59 / 1.56	0.43 / 0.43
Qwen2.5-14B-Instruct	0.25 / 0.24	0.28 / 0.26	1.65 / 1.62	0.46 / 0.45
Gemini-1.5-Flash-8b	0.26 / **0.27**	0.25 / 0.26	**1.54** / **1.54**	0.46 / **0.47**
LLaMA3.1-8b-Instruct	0.26 / 0.26	0.26 / 0.26	1.67 / 1.58	0.45 / 0.46

Searching

Model	Search Quality	Search Cost
DeepSeek-V3	0.494	0.800
Gemini-1.5-Flash-8b	0.622	0.802
Gemini-2.5-Flash	0.646	0.850
Gemini-2.5-Pro	0.632	0.808
GPT-4.1-Mini	0.608	0.904
GPT-5	0.656	0.940
LLaMA3.1-70b-Instruct	0.624	0.850





Further Analysis

EFFECT OF CHAIN–OF–THOUGHT (COT)

CoT provides only marginal gains in generation and often reduces stability for larger models. In evaluation, CoT shows almost no benefit — models already perform poorly at factor assessment.

FACTOR EVALUATION IS THE WEAKEST LINK

Zero–shot evaluation is near–random (Overall <0.50 for all models). Two root causes:

- No public paired (expression, IC) training data exists
- Plain text omits execution context, lag structure, and regime

SFT on GPT–4.1–Mini boosts Pairwise Selection **0.44→0.86** on CSI300, transferring cross–market. The search process generates rich supervised training data.

Insights

GENERATION

LLMs are reliable generators (>80% validity) but fail semantically on hard instructions. High reliability ≠ high accuracy: GPT–5 drops 0.99→0.10 on hard Text2Alpha.

EVALUATION

Factor evaluation remains near–random. Path forward: structured AST representations, execution metadata, and weak–label SFT from search–generated pairs.

SEARCHING

Algorithm prefers EA over CoE/ToT, better exploration; Medium model such as Gemini–Flash or GPT–Mini is suitable for large scale runs; Backtesting remains essential;