LiveOption: Evaluating LLM Agents in Structured Option Trading with Nonlinear Payoffs

Haochen Luo[1], Yifan Li[1], Binh Minh An[1], Xiaolong Luo[1], Zhengzhao Lai[3], Yuan Zhang[2], Chen Liu[1]
[1] City University of Hong Kong [2] Shanghai University of Finance and Economics [3] The Chinese University of Hong Kong (Shenzhen)

Why LiveOption?

Stock-trading benchmarks mainly test price direction. Options also require agents to choose strikes, expiries and multi-leg structures under time decay, volatility and margin constraints. LiveOption benchmarks these structured decisions in a reproducible environment; it is not a stock-price prediction model.

Stock vs. option trading

	Stock trading Direct equity	Option trading Derivatives contract
Payoff	Linear Asset ownership	Nonlinear State-dependent
Decision space	Buy or sell shares	Call / put, strike, expiry; single or multiple legs
Risk factors	Price movement	Price, volatility, time decay, Greeks
Capital	Full share price Higher outlay	Option premium For long options

Options strategy design

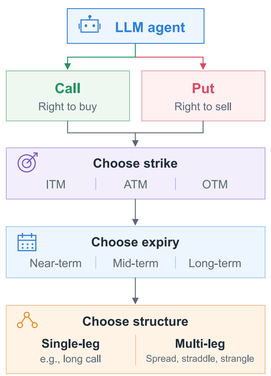


Fig. 1 excerpt · Why directional accuracy is insufficient.

Three Trading Scenarios

Overlay Strategies

Overlay agents protect stocks with puts or earn covered-call premium. Both choices trade potential return against protection costs or capped upside.

Earnings Bet

Agents choose structures around earnings reports, where a correct price view may still lose money as volatility reprices and time value decays.

0DTE Intraday

Agents trade options expiring that day and must size, monitor and close positions as time value rapidly decays and settlement approaches.

Example: A Replayed Trading Day

2025-01-15 | DeepSeek V4 Flash | ret=-3.26% SPY=+0.47% α=-3.73% fills=21



In the 15 January 2025 session, DeepSeek V4 Flash returned −3.26% while SPY gained 0.47%. Figure 18 shows this single trading path, not average performance.

Beyond direction: structure, execution and risk.

How It Works

Agents analyze market context, trade and receive execution feedback.

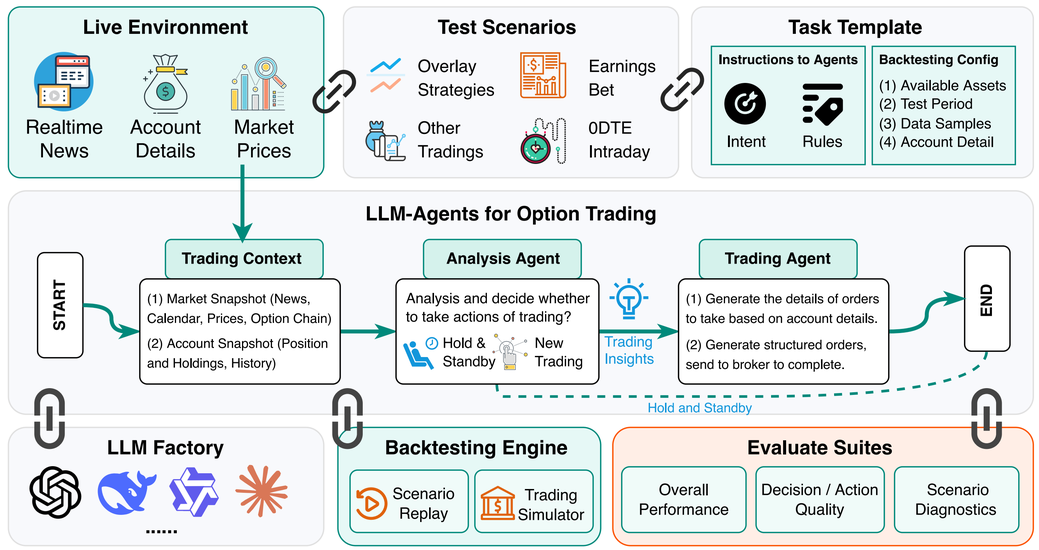


Fig. 2 · Chronological market replay and a brokerage-style simulator.

Scenario replay supplies time-ordered market and account information to an analysis agent, which forms a trading view. A trading agent turns that view into structured orders; backend verification checks contract validity, margin and position limits before the simulator updates positions and returns execution feedback.

To keep the decision space manageable, AVSR summarizes volatility with nine anchors, while QOCR groups contracts into discrete type, strike, expiry and size choices. Greeks describe sensitivity to price, time and volatility, supporting assessment of decision quality, risk and outcomes.

Experimental Results

Six LLMs were evaluated on 2025 market data, yielding 1,235 selected main runs with one rollout per model, configuration and case at temperature 0.5.

Earnings and 0DTE: typical episodes lose

Model	Earnings Bet			0DTE Intraday		
	n	Mean % [95% CI]	Med %	n	Mean % [95% CI]	Med %
DeepSeek-V4-Flash	94	−0.59 [−2.67, 1.54]	−1.28	102	−0.73 [−1.41, 0.02]	−2.12
GPT-OSS-120B	95	−0.20 [−1.91, 1.60]	−0.53	102	0.14 [−1.20, 1.58]	−3.20
Qwen3-235B-A22B	94	−1.23 [−3.57, 1.38]	−1.20	102	−0.36 [−0.83, 0.13]	−0.70
Llama-3.3-70B-Instruct	90	−0.37 [−2.24, 1.66]	−0.78	102	−0.59 [−1.41, 0.31]	−2.18
MiniMax-M3	95	−1.81 [−3.37, −0.26]	−1.61	102	0.14 [−0.36, 0.72]	−0.26
GLM-5.3-Flash	95	−0.48 [−2.32, 1.44]	−0.92	102	0.11 [−0.81, 1.17]	−1.68

Table 2 retains sample count (n), mean return with 95% confidence interval and median (Med). Every median is negative, and no pairwise return difference survives Holm correction.

Trading choices depend on the task

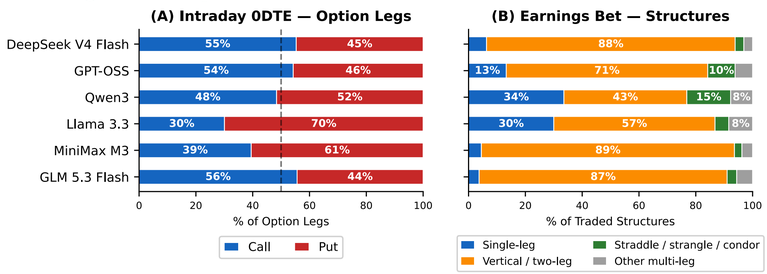


Figure 4 compares 0DTE call/put legs with Earnings structures; the two panels use different denominators.

Portfolio Evidence · Risk reduction rarely adds return

Across five assets per model

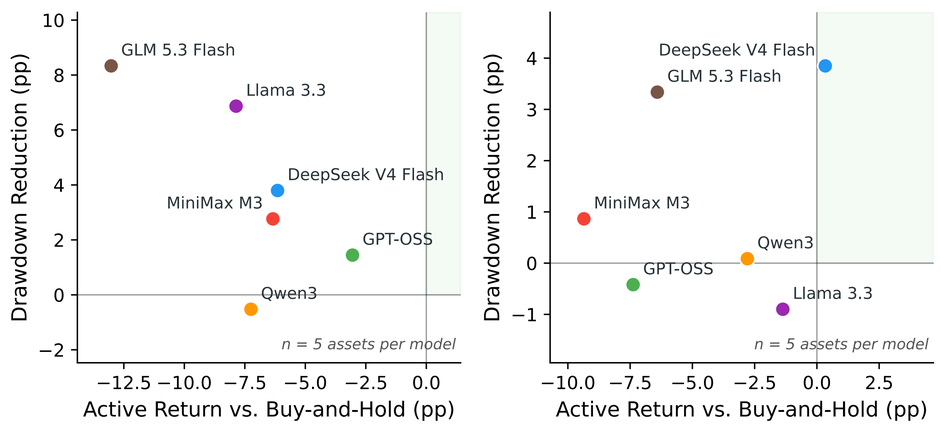


Figure 3 compares hedging and covered calls with buy-and-hold. Both axes use percentage points; lower drawdown means a smaller peak-to-trough loss.

Evaluate risk and information use alongside returns.
Quoted half-spread costs remove all positive 0DTE means (§3.2).

How active returns evolve: TSLA

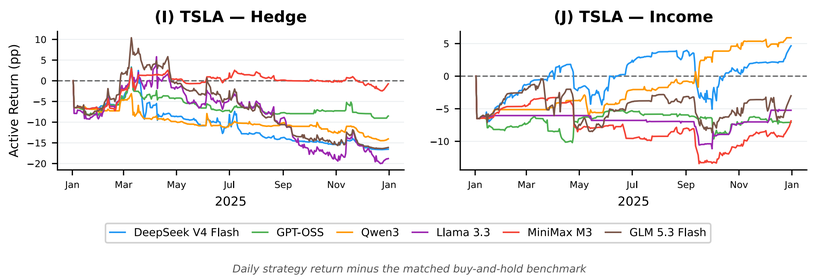


Figure 12 shows the TSLA paths in 2025 with the original model legend. This is one asset; Figure 3 reports the five-asset means.

Context ablations (Table 6 excerpt)

Ablation	Task	ΔRet (pp)	Fills	NT%	MDD (%)	min p
No portfolio state	0DTE	+2.36	29.5 → 16.1	0.0	4.4 → 12.4	0.698
No Greeks / multi-TF	0DTE	+0.25	29.5 → 4.4	47.2	4.4 → 1.0	0.835

Three-model means (24 paired cases/profile/task) show maximum drawdown rising from 4.4% to 12.4% without portfolio state. No return effect survives Holm correction.