

DeXposure-Claw: An Agent System for DeFi Risk Supervision

Aijie Shu¹ Bowei Chen² Wenbin Wu³

Cathy Yi-Hsuan Chen² Fengxiang He¹

¹University of Edinburgh ²University of Glasgow

³University of Cambridge

{v1ashu, fhe}@ed.ac.uk

{bowei.chen, CathyYi-Hsuan.Chen}@glasgow.ac.uk

w.wu@jbs.cam.ac.uk

Abstract

Decentralized finance exposes supervisors to fast-moving, networked credit risks that are difficult to inspect directly from raw on-chain records. We present DeXposure-Claw, an end-to-end decision-support system that converts forecasts of future exposure networks into bounded, auditable review tickets. DeXposure-FM produces future graph distributions. Deterministic monitors and stress scenarios organize forecast signals into a typed EvidenceBundle, and an LLM turns this record into ranked tickets with evidence-linked rationales. Data-health and confidence controls restrict intervention-level outputs when the supporting evidence is insufficient, while supervisors retain final authority. DeXposure-Bench evaluates the resulting workflow across forecasting, warning, calibration, stress analysis, ticket construction, and degraded-input behavior. The release schema separates forecast evidence, queue position, and control state for supervisory review. Retained legacy tickets do not contain every designed gate field. Code: <https://github.com/EVIEHub/DeXposure-Claw/tree/finnlp-2026-code-snapshot-v1>.

1 Introduction

Decentralized finance (DeFi) creates networks of token-mediated credit exposure among lending protocols, decentralized exchanges, stablecoins, bridges, and yield aggregators. A shock at one entity can change exposures at other entities before a supervisor can inspect the underlying on-chain records. Historical events including FTX illustrate this networked propagation problem (Vidal-Tomás et al., 2023). A supervisory decision-support system must therefore connect a forward-looking network signal to a limited review queue and preserve the evidence behind each queued item.

Forecasting and supervision operate on different objects. A forecaster estimates future nodes,

edges, weights, and uncertainty. A supervisor must instead choose a bounded set of targets and decide whether the evidence supports escalation. Lower graph error does not establish that the first k queue entries match future high-loss protocols, while a readable ticket does not establish that its target or action is forecast-grounded. We therefore separate the predicted graph, evidence record, ranked queue, and release state.

The decision-time data object is a sequence of weighted directed graphs. DeXposure joins weekly protocol holdings with a token-to-issuer mapping (Wu et al., 2025). A protocol is a node, and an edge $p \rightarrow q$ records that protocol p holds a token issued or managed by protocol q ; its weight records the exposure value. Node descriptors retain protocol size, token composition, category, previous TVL change, and normalized graph degrees. This representation separates protocol size from inter-protocol dependence. Protocols with similar TVL can expose different counterparties, while stable node totals can hide changes in edge direction, weight, or concentration. Figure 2 shows how one observed graph, its frozen evidence, and a recorded bounded queue appear in the replay interface.

Time alignment separates available evidence from later scoring. At decision epoch t , the pipeline consumes only $G_{\leq t}$ and derived fields. The realized graph G_{t+h} is reserved for offline labels after ticket generation. Each evaluation record names the source and target snapshots, horizon, proxy rule, and queue depth, preventing the later outcome from entering the ticket.

An LLM can render numerical evidence as a readable ticket, but fluency does not validate its target or action. Existing systemic-risk evaluations often rank protocols by fractional exposure changes (Bertomeu et al., 2024; Gonon et al., 2025; Li et al., 2025). Such rankings can elevate large percentage changes at small entities without testing whether tickets match future high-loss targets. DeXposure-

Claw instead grounds every ticket in forecast evidence and scores target selection against absolute future loss.

We present *DeXposure-Claw* to connect a graph forecast with the supervisor’s review queue. DeXposure-FM (Shu et al., 2026) predicts future exposure networks. Monitors and stress scenarios add alerts, attribution, uncertainty, and simulated losses. The LLM converts this record into ranked tickets. Release controls then restrict which intervention-level proposals reach human review. The system does not take autonomous regulatory action.

The typed EvidenceBundle is the handoff contract between numerical processing and language generation. It exposes named monitor, attribution, uncertainty, data-health, horizon, and scenario fields rather than raw graph state. Each ticket selects one action level from a fixed playbook, names a target set, and cites EvidenceBundle fields in its rationale. The LLM cannot alter the forecast, monitor functions, stress definitions, gate thresholds, or final review decision. The queue Q_t orders tickets, and k fixes weekly review capacity. The release schema assigns fields for the evidence and control states needed to measure target precision, recall at k , false-intervention rate, stability, and review volume as separate variables. Retained legacy traces omit some designed gate fields, which bounds the traceability claim.

DeXposure-Bench is the evaluation harness for these heterogeneous outputs, not a corpus, forecaster, or composite LLM score. It maps forecast and calibration objects to upstream metrics, warning and stress objects to deterministic evidence checks, tickets to fixed-budget queue metrics, and degraded inputs to robustness tests. A missing output remains not applicable, and an upstream gain cannot support a downstream ticket or release-control claim. Section 5 defines the axes and compatible methods.

Evaluation uses historical replay on held-out weeks. Each comparison names its fixed object and varied component. Queue comparisons hold the weeks, proxy, and top- k rule fixed while changing the evidence path. Control comparisons execute named evidence-function and threshold conditions. Only rule-implied gate outcomes are interpreted deterministically, while other cross-condition differences remain descriptive. We report target coverage, false interventions, review volume, stability, and recorded gate states only for methods that emit

the required object. These offline results do not establish prospective distress detection, realized-loss ground truth, analyst validation, reduced live workload, or safe deployment.

We make three contributions.

- DeXposure-Claw implements an end-to-end path from predicted exposure networks to a bounded supervisory review queue.
- A typed EvidenceBundle and release schema specify the forecast, attribution, stress, rationale, and control fields required for each ticket.
- DeXposure-Bench evaluates forecast, warning, calibration, stress, ticket, and degraded-input objects separately, including workload and false interventions.

Section 2 reviews related work, and Section 3 defines the graph and decision objects. Sections 4–5.3 present the pipeline, evaluation contract, and results.

2 Related Work

Financial networks and DeFi exposure.

Clearing-network models connect bilateral obligations to contagion and systemic loss (Eisenberg and Noe, 2001; Acemoglu et al., 2015; Glasserman and Young, 2016). DeFi changes the observable contract but not the network question. Tokenized claims, shared collateral, and composable protocols create paths through which one entity’s impairment changes the exposures of others (Gogol et al., 2024; Bertomeu et al., 2024). Flash-loan attacks and liquidation spirals show how protocol mechanics can accelerate those paths (Qin et al., 2021; Warmuz et al., 2023). Our setting starts from reconstructed inter-protocol exposure graphs and asks which future changes should enter a bounded supervisory queue. The output is a review object with named evidence, not a clearing solution or an autonomous trading action.

Temporal graph forecasts for supervision.

Dynamic graph models represent how nodes, edges, and weights evolve over time (Pareja et al., 2020; Xu et al., 2020; Rossi et al., 2020). These models normally end at a predicted state or forecast score. A supervisory workflow must still decide which network changes deserve attention, which stress evidence supports them, and which proposals may leave the system. DeXposure-Claw treats the forecaster as the producer of future evidence and evaluates the decisions made after that evidence is available.

Bench positioning and supervisory proxy.

LLM-agent benchmarks including HELM (Liang et al., 2022), SWE-bench (Jimenez et al., 2024), and AgentBench (Liu et al., 2024) score open-ended reasoning, software repair, and generic agent behavior. Temporal-graph benchmarks including TGB (Huang et al., 2023) and OGB (Hu et al., 2020) score structural prediction. Neither family scores the targets and action levels in a supervisory ticket. Prior systemic-risk evaluations (Bertomeu et al., 2024; Gonon et al., 2025; Li et al., 2025) often rank entities by fractional weight change, which can elevate large percentage changes at small entities. DeXposure-Bench instead uses an author-designed absolute-loss supervisory proxy (§3). The proxy supports reproducible offline scoring, but it is not an independently validated regulator label.

LLM agents in finance and DeFi. Fin-GPT (Yang et al., 2023) adapts language models to financial data. DeFi-specific systems address transaction auditing (Yao et al., 2026), intent mining (Mao et al., 2025), smart-contract verification (Hu et al., 2026; Kong et al., 2026), price-manipulation detection (Liu et al., 2025), anomaly explanation (Watson et al., 2025), asset-preference auditing (Wu, 2026), and portfolio construction from graph and LLM encoders (Luo et al., 2025; Jeon and Lee, 2026). DeXposure-Claw instead drafts proxy-scored tickets from structured forecast evidence under separate release authority (§5.3).

Forecasts as evidence for LLMs. Prior work feeds structured numeric inputs into LLMs for PMI nowcasting (de Bondt and Sun, 2025) and macroeconomic forecasting (Carriero et al., 2025). A BIS primer surveys related financial applications (Kwon et al., 2024). Time-series foundation models including Chronos (Ansari et al., 2024), Lag-Llama (Rasul et al., 2024), and TimesFM (Das et al., 2024) make the forecast producer reusable across tasks. Tabular foundation models extend this pattern to heterogeneous structured data (Hollmann et al., 2025; Ereemeev et al., 2025). These systems use language models to make or explain forecasts. In DeXposure-Claw, the language model enters after the forecast and writes a review artifact from named evidence fields. The evaluation then follows that artifact through target selection and release control.

3 Preliminaries

We model decentralized-finance credit exposure as a temporal weighted directed graph. Let $\mathcal{P}$ be the universe of protocols and $\mathcal{X}$ the universe of tokens. At time t , protocol p holds tokens $X_p(t) \subseteq \mathcal{X}$, while protocol q issues or is economically linked to tokens $X_G(q) \subseteq \mathcal{X}$. Protocol p is exposed to q if $X_p(t) \cap X_G(q) \neq \emptyset$. The exposure network is $G_t = (V_t, E_t, W_t)$, where $V_t \subseteq \mathcal{P}$ is the active protocol set, $E_t \subseteq V_t \times V_t$ is the directed exposure set, and $W_t[p, q] = w_{pq,t} \geq 0$ is the corpus exposure weight from p to q . For protocol v , define its incident exposure mass as

$$w_t(v) = \sum_{u \in V_t} w_{uv,t} + \sum_{u \in V_t} w_{vu,t}.$$

The observed history is a sequence of snapshots $\mathcal{G}_{1:T} = \{G_1, \dots, G_T\}$, optionally with covariates X_t .

At decision epoch t , the forecasting problem is to estimate G_{t+h} for horizons $h \in \mathcal{H}$. A probabilistic temporal-graph forecaster outputs

$$P_{t,h} = P(G_{t+h} \mid G_{\leq t}, X_{\leq t}).$$

For each ordered pair (p, q) , let

$$\begin{aligned} \pi_{pq,t+h} &= \Pr(e_{pq,t+h} = 1 \mid G_{\leq t}, X_{\leq t}), \\ \mu_{pq,t+h} &= \mathbb{E}[w_{pq,t+h} \mid e_{pq,t+h} = 1, G_{\leq t}, X_{\leq t}]. \end{aligned}$$

A representative predicted graph $\hat{G}_{t+h}$ is constructed by

$$\begin{aligned} \hat{E}_{t+h} &= \{(p, q) : \pi_{pq,t+h} > \pi_{\min}\}, \\ \hat{W}_{pq,t+h} &= \pi_{pq,t+h} \mu_{pq,t+h}. \end{aligned}$$

Uncertainty is represented by Monte Carlo samples

$$\tilde{G}_{t+h}^{(1)}, \dots, \tilde{G}_{t+h}^{(M)} \sim P_{t,h}.$$

The decision benchmark derives its scoring proxy from future observed graphs, not from the forecast or the LLM output. For protocol v and horizon h , let

$$\Delta_t^h(v) = w_t(v) - w_{t+h}(v).$$

The proxy set S_t^h contains the top fraction π of active protocols with positive $\Delta_t^h(v)$, ranked by absolute decline. We use $\pi = 0.05$ and $h = 4$ weeks. These choices define an author-designed supervisory proxy. They have not been validated as regulator labels or intervention outcomes.

At each weekly decision epoch, the system returns a ranked queue $Q_t = (\tau_{t,1}, \dots)$ of supervisory tickets. Each ticket records an action, a target set, a rationale, its supporting EvidenceBundle, and the release-control states. A supervisor inspects the first k tickets permitted by the available capacity. Changing k does not rerun the forecast or ticket drafting. Let $\hat{S}_t^{\text{int}}$ be the targets assigned RECOMMENDREDUCE or CONTINGENCY. The false-intervention rate is

$$\text{FIR} = \frac{|\hat{S}_t^{\text{int}} \setminus S_t^h|}{|\hat{S}_t^{\text{int}}|},$$

and is undefined when no intervention-level target is emitted. FIR measures disagreement with the proxy. It does not measure the causal harm or benefit of an intervention.

A monitor maps graphs to scalar risk statistics $\Phi_j : \mathcal{G} \rightarrow \mathbb{R}$, such as maximum PageRank, HHI concentration, density, PageRank Gini, or degree Gini. An alert fires when the predicted-graph statistic $\Phi_j(\hat{G}_{t+h})$ deviates from its length- L rolling baseline by more than z standard deviations.

A stress scenario is a graph perturbation $s : \mathcal{G} \rightarrow \mathcal{G}$ that removes exposure mass. We summarize its effect on the Monte Carlo samples by the conditional value-at-risk $\text{CVaR}_\lambda(s, h)$, the mean of the worst $\lceil \lambda M \rceil$ sampled loss fractions.

4 DeXposure-Claw Pipeline

The pipeline converts one observed graph into one bounded review queue while preserving the object produced at every handoff. This section first gives the end-to-end path, then specifies how forecasting, monitoring, ticket drafting, and release control transform the record. The EvidenceBundle separates the stages that compute numerical evidence from the stage that writes language, and the release record separates ticket drafting from action eligibility.

4.1 Overview

Figure 1 follows the record through four stages. Layer 1 forecasts future credit exposure, Layer 2 adds attributed monitor and stress evidence, Layer 3 drafts ranked tickets, and Layer 4 records whether each intervention-level proposal satisfies deterministic release conditions. Raw graph state cannot bypass the EvidenceBundle, and the LLM cannot bypass the release controls.

4.2 Layer 1: Forecasting

At decision epoch t , DeXposure-Claw observes the exposure graph G_t defined in Section 3. For each forecast horizon $h \in \{1, 4, 8, 12\}$ weeks, it queries DeXposure-FM for a predictive distribution over future exposure and constructs an expected-weight predicted graph $\hat{G}_{t+h}$. It also draws Monte Carlo graph samples that carry predictive uncertainty into Layer 2. The edge threshold $\pi_{\min}$, sample count, and horizon set are fixed before evaluation.

4.3 Layer 2: Monitoring

Layer 2 turns the forecast into a typed evidence bundle along two parallel tracks, monitors and stress scenarios. It also scores the snapshot’s *data health* $\text{DH}_t \in [0, 1]$, the mean of the checks available for that snapshot (Section 3). Timestamp validity, explicit feature missingness, and topology are always checked. Raw node and edge validity require source counts, while discontinuity requires a previous snapshot. Data health later modulates both alert confidence and the Layer 4 gates. If $\mathcal{Q}_t$ is the set of checks available at epoch t , the recorded score is

$$\text{DH}_t = \frac{1}{|\mathcal{Q}_t|} \sum_{q \in \mathcal{Q}_t} q.$$

Given $\hat{G}_{t+h}$ and its Monte Carlo samples, five *monitor functionals* summarize concentration, fragility, and instability in the predicted graph, including top-1 PageRank (Brin and Page, 1998), HHI concentration (Rhoades, 1993), network density, PageRank Gini, and degree Gini. An alert fires when a monitor statistic deviates from its rolling baseline by more than two standard deviations. Each alert carries a marginal-contribution top- K edge attribution and an uncertainty-aware confidence score $C_t(a) \in [0, 1]$, which decreases with lower data health or wider horizon-specific Monte Carlo dispersion. For alert a on statistic j , let $\text{CV}_{t,h}^{(j)}$ be the coefficient of variation across its predictive samples, clamped to $[0, 1]$. The confidence record is

$$C_t(a) = \text{DH}_t / \left(1 + \text{CV}_{t,h}^{(j)}\right).$$

The forecast horizon enters through the predictive sampling distribution instead of a second deterministic penalty.

In parallel, a *scenario engine* applies the five fixed shocks in Table 1 to $\hat{G}_{t+h}$ and each Monte Carlo sample. For each scenario s and horizon h , the engine aggregates the worst-half sampled

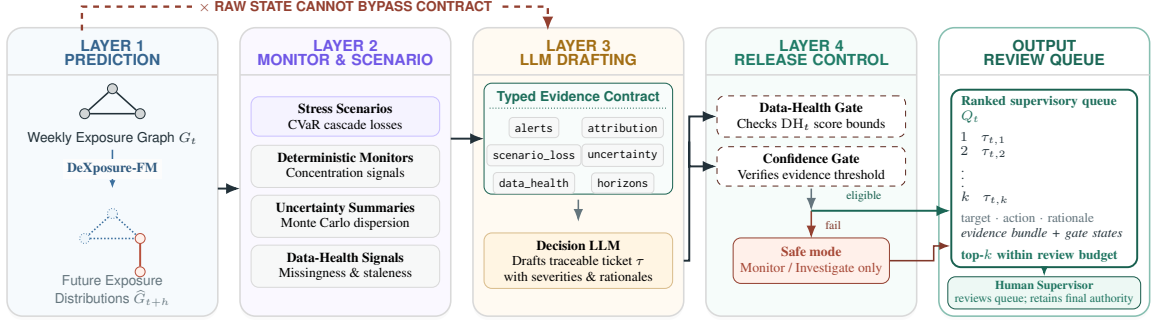


Figure 1: DeXposure-Claw workflow. Forecast and deterministic Layer 2 signals form a typed EvidenceBundle. The LLM drafts ranked tickets, data-health and confidence gates restrict intervention eligibility, and a supervisor reviews the top k while retaining final authority. The release schema requires evidence, rationale, and gate fields.

losses as $\text{CVaR}_{0.5}(s, h)$. These alerts, attribution fields, confidence scores, and scenario-CVaR table form the evidence consumed by the LLM.

Table 1 lists the five monitors and five scenarios side by side. The columns are independent. Monitors summarize the predicted graph, while scenarios store loss and affected targets after a named perturbation. The LLM cannot select or modify either library.

4.4 Layer 3: Ticket Drafting

Layer 3 is the only stage that invokes the LLM, which receives the evidence bundle as a typed JSON payload containing alerts with attribution top- K , a scenario-CVaR table, Monte Carlo dispersion summaries, and the scalar data-health score DH_t . It drafts, but does not release, a ranked list of supervisory tickets. Each ticket contains (a) a severity level from a fixed four-action playbook (Monitor, Investigate, Recommend-Reduce, and Contingency), (b) a target set of protocols, and (c) a rationale that cites specific evidence fields by name. Each decision is drafted three times at temperature 0; the Jaccard overlap of ticket targets and severities across the three drafts is the *target-stability* signal reported in our evaluation.

4.5 Layer 4: Release Controls

Candidate tickets are not released by default. Layer 4 applies data-health and confidence gates to determine whether intervention-level recommendations are feasible. The *data-health gate* thresholds the Layer 2 score. If $\text{DH}_t < \tau_{\text{data}} = 0.7$, the system enters safe mode and emits only Monitor or Investigate tickets. The *confidence gate* blocks intervention-level tickets whenever the mean alert confidence satisfies $\bar{C}_t < \tau_{\text{conf}} = 0.6$, even if the

data-health gate passes. The release schema requires the evidence bundle, the LLM rationale, and the gate states that produced the ticket. The retained legacy m7 records contain forecast and stress evidence plus a post-hoc action flag, but not the data-health and confidence states. They therefore do not validate complete Layer 4 traces. For an intervention-level proposal u , the release condition is

$$\text{Release}_t(u) = \mathbb{I}[\text{DH}_t \geq 0.7 \wedge \bar{C}_t \geq 0.6].$$

Failure of either condition changes the available action to Monitor or Investigate and records the failed state.

The evaluated m7 configuration applies a post-hoc gate to the LLM-reported risk level and score rather than to the data-health and confidence signals above. Its ablation isolates post-hoc action control and does not validate the two design gates.

Figure 2 joins the review queue, observed graph, selected ticket, and release-control record for one historical Terra/Luna replay.

5 Experiments

The evaluation follows the pipeline handoffs. Each DeXposure-Bench axis scores only configurations that emit its required evidence, queue, or control object; missing objects remain unscored.

5.1 Evaluation Questions

The questions separate forecast availability, target selection, and escalation control.

RQ1: What forecast evidence is available for downstream review? **RQ2:** How does the evidence path shape a bounded review queue? **RQ3:** When do deterministic controls restrict escalation?

RQ2 holds the supervisory proxy and queue rule fixed. RQ3 records outcomes under named

Table 1: Fixed EvidenceBundle library. Alerts use a 42-week baseline with $z = 2$; scenarios use the expected graph and $M = 50$ predictive samples.

Monitor	Signal	Definition	Scenario	Shock	Target and magnitude
N1	max PageRank	$\max_v \text{PR}(v)$	S1	protocol failure	top-weight protocol, 100%
N2	HHI	squared out-degree shares	S2	bridge failure	all bridge nodes, 100%
N3	density	$ E /(V (V - 1))$	S3	de-peg	all stablecoin nodes, 50%
N4	PR Gini	Gini of PageRank	S4	lending shock	all lending nodes, 30%
N5	degree Gini	Gini of weighted degree	S5	correlated stress	top ten nodes, 20%

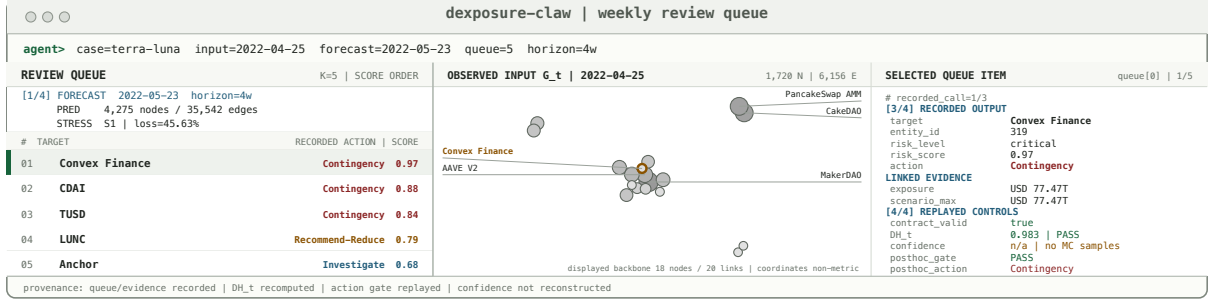


Figure 2: Historical replay interface for the Terra/Luna case at decision epoch 2022-04-25. The left panel shows the recorded $K = 5$ queue. The center shows a filtered backbone of observed G_t , whose complete mapped snapshot contains 1,720 nodes and 6,156 directed edges. The right joins the selected output to frozen evidence, recomputed data health, and a replayed post-hoc action gate. Confidence is unavailable because the archive contains no Monte Carlo samples.

EMITTED OBJECT FAMILY	COMPATIBLE AXES	USE
predicted graph or distribution	b1, b3, b6	RQ1
monitor alerts or scenario outcomes	b2, b4	evidence checks
ranked supervisory ticket queue	b5	RQ2, RQ3

No composite score; absent required objects remain n/a.

Figure 3: Compact object map for DeXposure-Bench. Each axis scores only compatible artifacts, and no composite score is formed.

evidence-function and release-threshold conditions. The three outcome units are not aggregated.

5.2 DeXposure-Bench

DeXposure-Bench is a six-axis evaluation harness rather than a corpus, forecaster, or aggregate leaderboard. It attaches compatible metrics to artifacts emitted at defined pipeline handoffs. Figure 3 groups the axes by artifact family, and Table 2 maps configurations to emitted objects and axes.

Forecast and calibration score the predicted graph and uncertainty; warning and stress score Layer 2 outputs; decision scores target quality, queue-depth coverage, stability, and false interventions; and robustness repeats compatible metrics under controlled degradation. Separate axes prevent one artifact from masking another.

The primary protocol uses the absolute-loss proxy in §3, with $\pi = 0.05$, $h = 4$ weeks, approximately 283 protocols per scored week, and 29 held-out weeks. RQ1 uses b1, b3, b6, RQ2 uses b5, and RQ3 tests the clean rules path and four masking regimes. The b2 and b4 axes validate warning and stress objects.

Data and split. Experiments use DeXposure weekly exposure graphs. The corpus contains 4,300 protocols, 24,300 tokens, 43.7 million exposure entries, and 283 snapshots from 2020-03 through 2025-08 (Wu et al., 2025). All model selection uses pre-2025 data. The 2025 period is the frozen test split. Each reported metric can be traced to a source snapshot, target snapshot, labeling rule, configuration, and scoring function in the evaluation artifact. The benchmark was built by the same authors as the system and has no independent regulator validation.

Reference methods and compatible outputs. Table 2 identifies the input and output of each reference configuration. Predictor-only methods are reported on forecast-compatible axes, and ticket-producing methods are reported on b5. The current-snapshot LLM receives a text summary of G_t and has no forecast or additional tools.

Persistence copies G_t to the forecast horizon, so

Table 2: Configurations, emitted objects, and compatible benchmark axes.

ID	Configuration	Emitted object	Scored axes
h1	weighted-degree monitor on G_t	alert stream	b2
m1	persistence plus rules	predicted graph, ticket set	b1, b4, b5, b6
m2	current snapshot plus LLM	ranked tickets	b5
m3	EvolveGCN (Pareja et al., 2020)	predicted graph	b1, b4, b6
m4	DeXposure-FM only	graph distribution	b1, b3, b4, b6
m5	DeXposure-FM plus rules	graph distribution, ticket set	b1, b3, b4, b5, b6
m6	DeXposure-FM plus LLM	ranked tickets	b5
m7	DeXposure-FM plus LLM and post-hoc gate	ranked tickets	b5

Table 3: Forecast records on 29 test weeks at $h = 4$ (RQ1). PR-MAE is in 10^{-5} . Mean Δ_{rel} is the signed mean across five b6 masks, not a lower-is-better score. § Persistence has fixed zero trend; only the DeXposure-FM forecast reports PI coverage and ECE.

Model	PR-MAE ↓	RankC. ↑	Trend ↑	mean Δ_{rel}	PI-cov/ ECE
Persistence	3.4	.570	.000§	.148	n/a
EvolveGCN	3.5	.551	.324	.040	n/a
DeXposure-FM	4.5	.558	.628	.113	.913/.013

it has zero trend by construction and no learned uncertainty. The m7 post-hoc gate re-labels high-severity outputs from the reported LLM risk level and score. It is evaluated separately from the data-health and alert-confidence controls in the system design.

Statistical protocol. Using the 29 held-out weeks as resampling units, we form percentile confidence intervals from 10,000 bootstrap samples and paired two-sided p -values from 20,000 sampled sign assignments. These are Monte Carlo replicates, not additional observations or runs. Because adjacent graphs share network state, the unblocked intervals and p -values are descriptive rather than causal.

5.3 Empirical Results

Tables 3–5 report the forecast, queue, and control outcomes separately.

RQ1: forecast evidence. Persistence has the strongest point-error and rank-correlation values, while DeXposure-FM alone supplies prediction intervals and samples for later monitors and stress scenarios. EvolveGCN supplies a learned point forecast without those uncertainty fields. The result does not identify a universally superior forecaster.

The two remaining axes check Layer 2 objects. At $K = 5$, the b2 historical event study records (precision, recall $\times 10^3$, lead weeks) of (.60, 1.81,

5) for Terra/Luna, (.60, 1.33, 5) for FTX, and (1.00, 1.81, 4) for SVB-USDC. These observations precede the frozen 2025 split and do not show prospective detection. On b4 at $h = 4$, DeXposure-FM overlap@10 across S1–S5 is (.314, .525, .445, .536, .602), compared with (.314, .521, .445, .536, .602) for persistence and (.034, .189, .269, .292, .330) for EvolveGCN. These values measure scenario-target fidelity rather than ticket quality.

RQ2: bounded review queue. Table 4A reports complete ticket sets, and Panel B applies explicit weekly queue depths. The m1 fields use one consistent rerun; m5 lacks per-target order and receives no matched-budget values. Against the snapshot LLM (m2), paired recall tests for m6 and m7 are unresolved through $k = 5$ ($p \geq .25$). After the snapshot set saturates at $k = 7$, $p = .0001$ for m6 and $p = .0003$ for m7. This reflects additional targets and workload, not better standalone LLM reasoning; FIR remains separate from ticket F1.

RQ3: deterministic controls. Without scenario evidence, precision is .000 versus .600 for full m5; without the confidence gate, FIR is .429. In Panel B, thresholds above the recorded mean health activate safe mode and yield zero interventions. Lower thresholds yield 57, 58, 5, and 25 interventions across the four masking regimes. Because rows are independent condition runs, only the safe-mode implication is deterministic; precision and FIR differences are descriptive.

5.4 Supervisory Interpretation

The separate outcome variables expose missing upstream fields and disabled controls that a system-wide score would conceal.

Increasing k changes available workload and proxy coverage. It neither transfers action authority to the LLM nor establishes that every additional target is correct.

The degraded-input conditions locate the tested

Table 4: Ticket results on 29 test weeks (RQ2). Interv.(wk) counts weeks with interventions; Stab. is three-draft target/action Jaccard. Brackets show 95% CIs. FIR and missing target order are n/a; * marks target saturation.

(A) End-to-end ticket quality

ID	Stack	Interv.(wk)	Prec.	Rec.	F1 [95% CI]	FIR [95% CI]	Stab.
m1	persist.+rules	0/29	.633	.004	.0081 [.0041,.0123]	n/a	.533
m2	snapshot LLM	0/29	.575	.009	.0184 [.014,.022]	n/a	.532
m5	FM+rules	0/29	.600	.010	.0190 [n/a]	n/a	.257
m6	FM+LLM	17/29	.570	.012	.0241 [.020,.028]	.448 [.293,.601]	.488
m7	FM+LLM+post-gate	16/29	.580	.012	.0234 [.019,.027]	.437 [.282,.592]	.435

(B) Matched-budget ($R@k \times 10^3$ / $P@k$, top- k targets/week)

Method	R@1	P@1	R@3	P@3	R@5	P@5	R@7	P@7
m1 persist.+rules	0.71	.24	2.45	.26	4.05	.26	4.05*	.26
m2 snapshot LLM	1.53	.52	5.70	.59	9.08	.58	9.35*	.58
m5 FM+rules			n/a, retained artifact has no per-target order					
m6 FM+LLM	1.64	.55	5.30	.56	9.63	.59	12.14	.57
m7 FM+LLM+post-gate	1.64	.55	5.30	.56	9.13	.57	11.95	.58

Table 5: Condition records on 29 test weeks (RQ3). Safe is the percentage of safe-mode weeks; #int counts intervention tickets. Rows run independently. Only Safe = 100% $\Rightarrow$ #int = 0 is deterministic; other cross-row differences are descriptive, and FIR is n/a when #int is zero.

(A) Clean rules-path ablation

Config	Removed object	Prec.	FIR	Recorded behavior
full m5	none	.600	n/a	reference rules path
A3, no scenario engine	stress evidence	.000	n/a	no correct ticket target
A2, no confidence gate	confidence control	.600	.429	intervention disagreement appears

(B) Degraded-input threshold sensitivity at $\tau_{\text{conf}} = 0$

Masking regime	τ_{data}	$\overline{\text{DH}}$	Safe (%)	#int	FIR
edges + features, 90%	.70	.587	100	0	n/a
	.00	.587	0	57	.510
edges only, 90%	.70	.637	100	0	n/a
	.00	.637	0	58	.441
features only, 90%	.85	.755	100	0	n/a
	.70	.755	0	5	.267
edges + features, 80%	.85	.751	100	0	n/a
	.70	.751	0	25	.400

threshold boundary; other metric differences remain noncausal.

The retained m7 records preserve forecast and stress evidence plus a post-hoc action flag but omit designed data-health and confidence states. Table 6 supports only a ticket-behavior audit.

6 Conclusion

We presented DeXposure-Claw, which connects predicted exposure networks to a bounded supervisory review queue through a fixed EvidenceBundle and release schema. Drafting remains separate from deterministic controls and human authority. The retained records lack some designed gate fields, so complete release-trace validation remains future work.

Table 6: Stored m7 cases. The future proxy is evaluator-only, and action_gate_applied=true is a legacy flag. Missing data-health and confidence fields preclude complete Layer 4 trace validation.

Field	Grounded case, 2025-03-24	Failure case, 2025-04-28
Evidence	S1 loss 18.67%, 84 distressed; top target 2786 at \$9.35B	S1 loss 22.14%, 44 distressed; top target 2269 at \$5.64B
Interventions	2786, 2269, 118	2269, 2786, 1599
Stored con- trol	action_gate_applied=true	action_gate_applied=true
Future proxy	3/3 targets in S_t^4 ; FIR .000	0/3 targets in S_t^4 ; FIR 1.000

7 Limitations

DeXposure-Bench measures offline ticket targeting and control behavior on one DeFi corpus with an author-designed future-loss proxy. It does not measure intervention benefit, autonomous execution,

or supervisor workload. Validation with practicing supervisors must measure ticket choices and review time.

The 29 held-out weekly graphs do not represent shocks that unfold within a day, and adjacent weeks share network state. The reported intervals are descriptive. Historical names may also occur in an LLM’s training data, and the rules and LLM paths receive nonidentical inputs. These constraints preclude causal attribution to the LLM layer.

The retained m7 records store forecast and stress evidence plus a post-hoc action flag, but not the designed data-health and alert-confidence states. The case audit therefore does not validate complete Layer 4 release traces. The degraded-input rows are separate condition runs and support the fail-closed rule implication, not causal estimates for precision or FIR.

8 Responsible Use and Disclosure

DeXposure contains public on-chain aggregates collected through DefiLlama (DefiLlama, 2026). The evaluation uses no wallet-level personal data or off-chain identities. DeXposure-Claw emits research tickets rather than trades or autonomous interventions. A human must review every ticket before any external action. The data-health and alert-confidence controls belong to the system design, while the evaluated m7 result uses a distinct post-hoc action gate.

The authors used AI assistants for code and writing support. The authors remain responsible for every claim, citation, experiment, and sentence.

References

- Daron Acemoglu, Asuman Ozdaglar, and Alireza Tahbaz-Salehi. 2015. Systemic risk and stability in financial networks. *American Economic Review*, 105(2):564–608.
- Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the language of time series. *Preprint*, arXiv:2403.07815.
- Jeremy Bertomeu, Xiumin Martin, and Ibrahima Sall. 2024. Measuring defi risk. *Finance Research Letters*, 63:105321.
- Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117. Proceedings of the Seventh International World Wide Web Conference.
- Andrea Carriero, Davide Pettenuzzo, and Shubhranshu Shekhar. 2025. Macroeconomic forecasting with large language models. *Preprint*, arXiv:2407.00890.
- Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. *Preprint*, arXiv:2310.10688. Accepted at ICML 2024; model commonly referenced as TimesFM.
- Gabe J. de Bondt and Yiqiao Sun. 2025. Enhancing GDP nowcasts with ChatGPT: a novel application of PMI news releases. Working Paper Series 3063, European Central Bank, Frankfurt am Main, Germany. ECB Working Paper Series No. 3063; released 2025-06-30.
- DefiLlama. 2026. DefiLlama: DeFi TVL aggregator and analytics. Data used: 2025; accessed: 2026-02-02.
- Larry Eisenberg and Thomas H. Noe. 2001. Systemic risk in financial systems. *Management Science*, 47(2):236–249.
- Dmitry Ereemeev, Gleb Bazhenov, Oleg Platonov, Artem Babenko, and Liudmila Prokhorenkova. 2025. Turning tabular foundation models into graph foundation models. *Preprint*, arXiv:2508.20906.
- Paul Glasserman and H. Peyton Young. 2016. Contagion in financial networks. *Journal of Economic Literature*, 54(3):779–831.
- Krzysztof Gogol, Christian Killer, Malte Schlosser, Thomas Bocek, Burkhard Stiller, and Claudio Tesone. 2024. Sok: Decentralized finance (defi) - fundamentals, taxonomy and risks. *Preprint*, arXiv:2404.11281.
- Lukas Gonon, Thilo Meyer-Brandis, and Niklas Weber. 2025. Computing systemic risk measures with graph neural networks. *Preprint*, arXiv:2410.07222.
- Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. 2025. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326. Published online 8 Jan 2025.
- Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. 2020. Open graph benchmark: Datasets for machine learning on graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Xiaohui Hu, Wun Yu Chan, Yuejie Shi, Qumeng Sun, Wei-Cheng Wang, Chiachih Wu, Haoyu Wang, and Ningyu He. 2026. An effective and cost-efficient

- agentic framework for ethereum smart contract auditing. *Preprint*, arXiv:2601.17833.
- Shenyang Huang, Farimah Poursafaei, Jacob Danovitch, Matthias Fey, Weihua Hu, Emanuele Rossi, Jure Leskovec, Michael Bronstein, Guillaume Rabusseau, and Reihaneh Rabbany. 2023. Temporal graph benchmark for machine learning on temporal graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Joohyoung Jeon and Hongchul Lee. 2026. Can blindfolded LLMs still trade? An anonymization-first framework for portfolio optimization. *Preprint*, arXiv:2603.17692.
- Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. Swe-bench: Can language models resolve real-world github issues? In *International Conference on Learning Representations (ICLR)*.
- Ziqiao Kong, Wanxu Xia, Chong Wang, Yi Lu, Pan Li, Shaohua Li, Zong Cao, and Yang Liu. 2026. Knowdit: Agentic smart contract vulnerability detection with auditing knowledge summarization. *Preprint*, arXiv:2603.26270.
- Byeungchun Kwon, Taejin Park, Fernando Perez-Cruz, and Phurichai Rungcharoenkitkul. 2024. Large language models: a primer for economists. *BIS Quarterly Review*.
- Xurui Li, Xin Shan, Wenhao Yin, and Haijiao Wang. 2025. Heterogeneous graph pre-training based model for secure and efficient prediction of default risk propagation among bond issuers. *Preprint*, arXiv:2501.03268.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, and 1 others. 2022. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*.
- Lu Liu, Wuqi Zhang, Lili Wei, Hao Guan, Yongqiang Tian, Yepang Liu, and Shing-Chi Cheung. 2025. LLM-powered detection of price manipulation in DeFi. *Preprint*, arXiv:2510.21272.
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, and 3 others. 2024. Agentbench: Evaluating llms as agents. In *International Conference on Learning Representations (ICLR)*.
- Yichen Luo, Yebo Feng, Jiahua Xu, Paolo Tasca, and Yang Liu. 2025. Llm-powered multi-agent system for automated crypto portfolio management. *arXiv preprint arXiv:2501.00826*.
- Qian’ang Mao, Yuxuan Zhang, Jiaman Chen, Wenjun Zhou, and Jiaqi Yan. 2025. Know your intent: An autonomous multi-perspective LLM agent framework for DeFi user transaction intent mining. *Preprint*, arXiv:2511.15456.
- Aldo Pareja, Giacomo Domeniconi, Jie Chen, Tengfei Ma, Toyotaro Suzumura, Hiroki Kanezashi, Tim Kaler, Tao Schardl, and Charles Leiserson. 2020. EvolveGCN: Evolving graph convolutional networks for dynamic graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5363–5370.
- Kaihua Qin, Liyi Zhou, Benjamin Livshits, and Arthur Gervais. 2021. Attacking the defi ecosystem with flash loans for fun and profit. In *Financial Cryptography and Data Security: 25th International Conference, FC 2021, Virtual Event, March 1–5, 2021, Revised Selected Papers, Part I*, page 3–32, Berlin, Heidelberg. Springer-Verlag.
- Kashif Rasul, Arjun Ashok, Andrew Robert Williams, Hena Ghonia, Rishika Bhagwatkar, Arian Khorasani, Mohammad Javad Darvishi Bayazi, George Adamopoulos, Roland Riachi, Nadhir Hassen, Marin Biloš, Sahil Garg, Anderson Schneider, Nicolas Chapados, Alexandre Drouin, Valentina Zantedeschi, Yuriy Nevmyvaka, and Irina Rish. 2024. Lag-llama: Towards foundation models for probabilistic time series forecasting. *Preprint*, arXiv:2310.08278.
- Stephen A. Rhoades. 1993. The herfindahl-hirschman index. *Federal Reserve Bulletin*, 79(3):188–189.
- Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *Preprint*, arXiv:2006.10637.
- Aijie Shu, Wenbin Wu, Gbenga Ibikunle, and Fengxiang He. 2026. DeXposure-FM: A time-series, graph foundation model for credit exposures and stability on decentralized financial networks. *Preprint*, arXiv:2602.03981.
- David Vidal-Tomás, Antonio Briola, and Tomaso Aste. 2023. Ftx’s downfall and binance’s consolidation: The fragility of centralised digital finance. *Physica A: Statistical Mechanics and its Applications*, 625:129044.
- Jakub Warmuz, Amit Chaudhary, and Daniele Pinna. 2023. Toxic liquidation spirals. *Preprint*, arXiv:2212.07306.
- Adriana Watson, Grant Richards, and Daniel Schiff. 2025. Explain first, trust later: LLM-augmented explanations for graph-based crypto anomaly detection. *arXiv preprint arXiv:2506.14933*.
- Wenbin Wu. 2026. Auditing asset-specific preferences in financial large language models: Evidence from Bitcoin representations and portfolio allocation. *Preprint*, arXiv:2606.02528.

- Wenbin Wu, Kejiang Qian, Alexis Lui, Christopher Jack, Yue Wu, Peter McBurney, Fengxiang He, and Bryan Zhang. 2025. Dexposure: A dataset and benchmarks for inter-protocol credit exposure in decentralized financial networks. *Preprint*, arXiv:2511.22314.
- Da Xu, Chuanwei Ruan, Evren K rpeoğlu, Sushant Kumar, and Kannan Achan. 2020. Inductive representation learning on temporal graphs. In *International Conference on Learning Representations (ICLR)*.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. FinGPT: Open-source financial large language models. *Preprint*, arXiv:2306.06031.
- Duanyi Yao, Siddhartha Jagannath, Baltasar Aroso, Vyas Krishnan, and Ding Zhao. 2026. Auditable LLM arbiter for DeFi security: A hybrid graph-of-thoughts approach to intent-transaction alignment. In *NDSS 2026 LAST-X Workshop*.