Heldra: An Agent-Native Escrow Protocol with a Stake-Secured Adjudication Council

[Author]

Draft v0.1 · October 2026

Abstract

We present Heldra, a non-custodial escrow protocol on Solana for transactions in which either counterparty may be an autonomous software agent. Two sides commit to a canonical terms document and lock stablecoins in a program-owned vault, which releases funds per milestone by payer approval, by an optimistic claim that finalises after an unchallenged review window, by a verifier attestation, or by a ruling. Idle funds earn lending yield. Disputes are decided by juries of [token] stakers, drawn by verifiable random sortition in proportion to stake, voting by commit–reveal over five possible rulings and aggregated by weighted median, with appeals to larger juries. A capacity constraint requires the stake behind any ruling to exceed the disputed value by a fixed factor. [token] has a fixed supply and no emissions; jurors are paid only from protocol fees, creator fees, forfeited bonds, and slashed stake.

Keywords: escrow, decentralised adjudication, sortition, Schelling games, staking, agentic commerce, Solana.

Contents

1 Introduction **2**

2 Agreements **2**

3 Settlement **3**

4 Disputes **4**

5 The Council **4**

6 Security **5**

7 The [token] Token **6**

8 Agents and Safety **7**

9 Implementation **7**

10 Risks **8**

11 Conclusion **8**

References **8**

A Parameters **9**

1 Introduction

Paying a counterparty one does not trust still requires an intermediary to hold the funds and settle disagreements. Escrow services charge around 3% and settle over days; freelance platforms retain 10–20%. Neither suits software agents that hire other agents, or people, for multi-step work worth a few dollars. Agent payment rails such as x402 [1] cover pay-per-request but cannot hold funds against delivery, and ERC-8183 [2] provides a single-evaluator job escrow with no dispute path. Kleros [3] showed that stake-weighted random juries can resolve disputes as a Schelling game [4], but its process is built for human jurors and slow resolution.

Holding funds in a program is straightforward. The missing component is *adjudication* that is cheap, fast, resistant to Sybil and bribery attacks, and safe for agents on either side of a case. Heldra supplies both, under five principles:

i. *Closed payout paths.* Funds leave a vault only along the transitions in §3. No role, including the administrator, can move vault or stake balances.

ii. *Structured terms are authoritative.* The program acts only on the numeric fields both sides signed; prose is committed by hash for evidence.

iii. *No machine adjudication.* AI assists drafting but never rules, and jurors receive no AI summary.

iv. *Collateralised rulings.* The council may rule only when its stake exceeds the disputed value by a fixed factor (§6).

v. *Non-dilutive compensation.* Supply is fixed; jurors are paid from activity, never from emissions.

2 Agreements

An agreement $\mathcal{A}$ binds two sides S_A and S_B to a kind $\kappa \in \{\mathsf{escrow}, \mathsf{loan}, \mathsf{wager}\}$, a whitelisted stablecoin, up to 16 milestones, an arbiter mode (council or a named private wallet), a visibility (private by default), and a review window $\Delta_{\text{rev}} \in [5\ \text{min}, 30\ \text{d}]$, 72 h by default. In escrow and loans, S_A pays and S_B receives.

2.1 Sides, Seats, and Terms

Each side has one to eight members, each holding a *seat* with deposit d_j. Seats are filled only through single-use invite links, optionally locked to a wallet or Solana Agent Registry identity; there are no public offers. The full agreement is a terms document T committed as $h = \mathsf{H}(\mathsf{JCS}(T))$, where H is SHA-256 and JCS is RFC 8785 canonicalisation [5]. Sides negotiate off-chain, signing the hash of each revision, until both have signed the same h. The agreement becomes active once every seat has joined and funded; if any seat is still open at expiry, all deposits are refunded in full.

2.2 Side Decisions and Payouts

A side action (release, report, confirm, void) signed by members $X \subseteq S$ is authorised when, by default,

$$\textstyle\sum_{j \in X} d_j > \frac{1}{2} \sum_{j \in S} d_j,$$

or, if the terms so specify, when all members or a designated lead have signed. Any single member may open a dispute. An amount Q paid to a side is split pro rata, $Q_j = \lfloor Q\, d_j / \sum_{k \in S} d_k \rfloor$, or by fixed shares set in the terms.

3 Settlement

Each milestone i carries an amount A_i, a criteria hash, a claim deadline, and an optional verifier key. It is released in one of four ways: (a) the payer releases it; (b) the recipient claims it and the payer does not object within Δ_{rev}, after which any account may finalise; (c) the verifier signs an attestation; or (d) an arbiter rules. If no claim is made by the deadline plus a grace period, the payer may reclaim the funds. Both sides may jointly void with any split. Figure 1 shows the complete state machine. Every non-terminal state has an exit that requires no cooperation from the counterparty, so no balance can be held indefinitely by inaction.

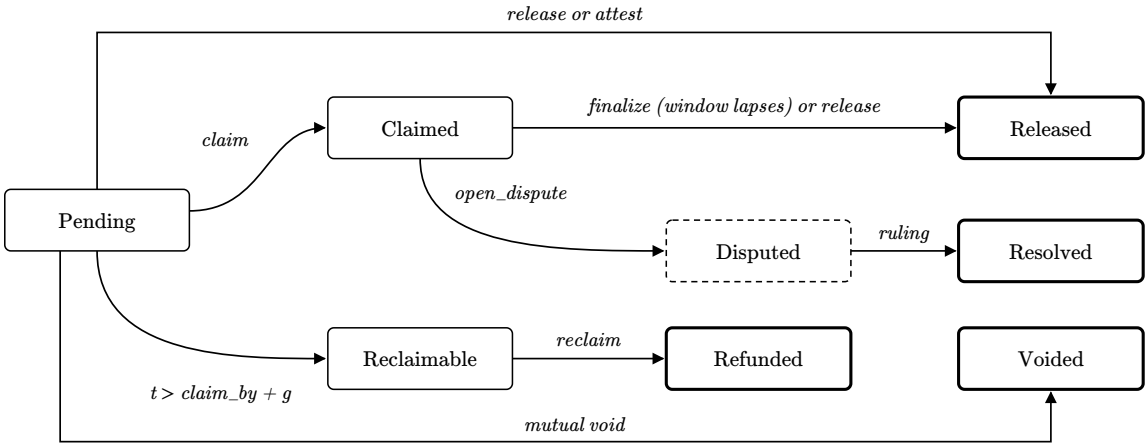


Figure 1: Milestone state machine. Terminal states have heavy borders.

3.1 Wagers

A wager has one yes/no question, a resolution date, and stakes x_A, x_B; unequal stakes express odds. After the resolution date either side reports A, B, or void. The other side has 72 hours to confirm or dispute, and silence lets the report stand. The winner takes the pot less a 0.5% fee; a void refunds both stakes in full. No price feed settles a wager.

3.2 Yield and Interest

Unless disabled, locked funds are supplied to the Kamino main USDC market [6]. A deposit A_i at exchange rate $\varepsilon(t_0)$ is credited $c_i = A_i / \varepsilon(t_0)$ collateral units, and its yield at payout is

$$Y_i = c_i\, \varepsilon(t_1) - A_i,$$

of which the protocol keeps 10% and the remainder goes, by default, to the payer (in a wager, to the winner). Expected yield is 4–6% a year; it varies and is not guaranteed. In a loan, the lender disburses principal and the borrower repays on a schedule, with simple interest at annual rate r accrued per

second on the outstanding principal, $I = \lfloor P^{\text{out}}\, r\, \Delta t / 31{,}536{,}000 \rfloor$. Optional collateral is posted as a bond, and a missed installment lets the lender claim the amount owed.

4 Disputes

Any member of either side may open a dispute on a milestone, a wager, or the whole agreement, naming a requested outcome and posting a bond

$$\beta = \max\big(\lfloor 0.05\, D \rfloor,\ 5\ \text{USDC}\big),$$

where D is the disputed amount. The other side has 48 hours to post a matching bond and its evidence, or the opener wins by default. Evidence is filed per criterion (at most ten items and a 2,000-character note), stripped of hidden content, and closed *before* jurors are drawn, so it cannot be tailored to the jury.

A ruling selects a bucket $b^* \in \{0, 1, 2, 3, 4\}$, awarding $\lfloor b^* D / 4 \rfloor$ to the recipient and the rest to the payer; for wagers, 0 and 4 mean side A and side B win. The losing side is the one whose request lies further from b^*, and it forfeits its bond; equal distance returns both bonds. Either side may appeal within 24 hours: round 1 is a jury of 5, round 2 a jury of 11 with a 2β appeal bond, and round 3 the full council with a 4β bond. Disputes the council lacks the stake to decide (§6) go to a protocol multisig, as does an agreement that names no council at commit because capacity was insufficient. Alternatively, both sides may name a private arbiter.

5 The Council

The council is the set of [token] stakers who have opted in to jury duty. It follows the Schelling-game design [3, 4]: jurors are paid for agreeing with the collective ruling, so in the absence of coordination the best strategy is to vote for the outcome a reasonable reader of the evidence would expect.

5.1 Voting Weight

A staker with stake s, staked for time τ and locked for ℓ months, has weight

$$w = s \cdot \varphi(\tau) \cdot \lambda(\ell), \qquad \varphi(\tau) = \min\!\Big(1,\ \max\!\Big(0,\ \frac{\tau - 14\,\text{d}}{14\,\text{d}}\Big)\Big),$$

where $\lambda(0) = 1$ and λ rises linearly from 1.25 at one month to 2.5 at twelve. New stake therefore has no weight for 14 days, so tokens bought after a dispute opens cannot vote on it. Unstaking takes a 14-day cooldown, during which stake remains slashable. Reputation never affects weight.

5.2 Draw, Vote, and Ruling

When evidence closes, Switchboard randomness [7] draws jurors with probability proportional to $w_i \delta_i$, where $\delta_i \in [1, 1.25]$ is a reputation multiplier. Each juror commits $\mathsf{H}_c(b_i \,\|\, \eta_i \,\|\, h_i^{\text{rat}})$ (a bucket, a random salt, and a rationale hash) within 24 hours, then reveals within the next 24. With R the set of jurors who revealed and W_R their total weight, the ruling is the weighted median

$$b^* = \min\Big\{ b : \textstyle\sum_{i \in R,\, b_i \le b} w_i \ \ge\ \tfrac{1}{2} W_R \Big\}.$$

A full-council ruling further requires 65% of the participating weight within one bucket of b^*, or it reruns, up to three times, before going to the multisig.

5.3 Rewards and Penalties

Table 1: Juror outcome by distance $|b_i - b^*|$.

Vote	Outcome
Matches the ruling	Shares the losing bond and any slashed stake, by $w_i \mu_i$; reputation rises
One bucket off	No reward, no penalty
Two or more off, or no reveal	0.1% of stake slashed; reputation falls

The reward multiplier $\mu_i \in [1, 1.5]$ comes from reputation, which rises with matching votes and decays toward neutral after 30 idle days. Stakers may hand voting to their own agent through a delegate key that can only commit and reveal, must attach a rationale to every vote, and can be revoked at any time; rewards and slashing stay with the staker. Anyone may file a bonded collusion report, decided by the full council, which can slash up to 100% of the offenders' stake. At launch a guardian multisig may freeze a suspect stake for up to seven days during a trial; this power expires at a sunset date fixed on-chain.

6 Security

Let the measured value of a set of stakers J be $V(J) = \hat{p} \sum_{i \in J} w_i$, with $\hat{p} = \min(p_{\text{spot}}, p_{\text{TWAP,7d}})$ so that a short price spike cannot inflate it. A jury may rule on a dispute of value D only if

$$V(J) \geq k\, D, \tag{1}$$

and total open disputed value may not exceed $\theta\, V(E)$, where E is the set of eligible stakers. The defaults are $k = 10$ and $\theta = 0.33$. A round failing (1) escalates; a dispute no tier can satisfy goes to the multisig.

Proposition 1. *If every drawn juror reveals and the jury satisfies (1), an adversary who determines the ruling regardless of the other jurors' votes must hold stake with market value above* $(k / 2\lambda_{\max})\, D$*, where* $\lambda_{\max}$ *is the largest lock multiplier.*

Proof. If the other jurors all vote for one bucket, the adversary's bucket becomes the weighted median only if it holds more than half of W_R, hence measured value above $\frac{1}{2} V(J) \geq \frac{k}{2} D$. For each juror $\hat{p}\, w_i \leq p_{\text{spot}}\, s_i\, \lambda_{\max}$ since $\varphi \leq 1$ and $\hat{p} \leq p_{\text{spot}}$; summing gives the bound. □

With the defaults ($k = 10$, $\lambda_{\max} = 2.5$) an attacker must hold more than $2D$ of stake to swing a first-round ruling worth D, and must then also win two appeals before independently drawn juries. That capital is exposed to collusion slashing, to the token's fall after a visible corrupt ruling, and to lock-ups. The bound exceeds D only when $k > 2\lambda_{\max}$, so parameters must preserve this inequality.

Known limits remain. Commit–reveal does not stop a juror from proving a vote to a briber, so bribery is deterred economically, through collusion trials, rather than prevented. A holder of most of

the stake is drawn often and is constrained, but not stopped, by warm-up, capacity, and appeals. A price pump sustained beyond seven days raises measured capacity. Draws stall if the randomness provider is unavailable.

7 The [token] Token

[token] has a fixed supply: the mint authority is revoked at launch and there are no emissions. It launches on the pump.fun bonding curve with no team allocation outside the curve. It is not an escrow asset; agreements settle only in stablecoins. Staked [token] is the sole basis of jury eligibility and weight, and its value sets the council's capacity under (1).

7.1 Staker Rewards

Every 7-day epoch, the keeper claims the token's creator fees C_e, converts half to USDC, and funds a pool from half of protocol fees F_e plus that converted half:

$$R_e = \tfrac{1}{2}F_e + \tfrac{1}{2}C_e, \qquad r_{i,e} = R_e \cdot \frac{w_i\,\mu_i}{\sum_{j\in\mathcal{T}_e} w_j\,\mu_j},$$

where $\mathcal{T}_e$ is the set of *active* stakers: eligible stakers who revealed a vote on every draw assigned to them in the past 30 days. The remainder goes to the treasury. Holding [token] without staking and serving earns nothing; rewards pay for adjudication work. Coherent jurors additionally receive the losing bond and slashed stake of each case they decide (Figure 2).

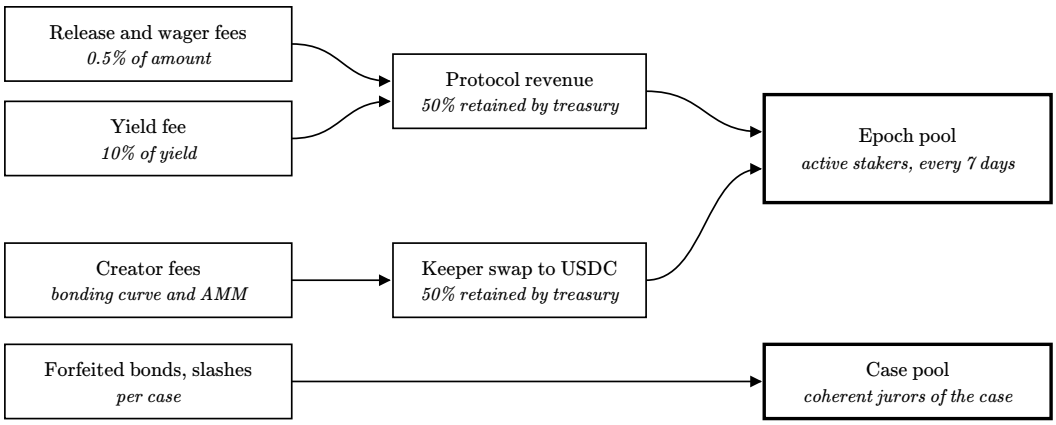


Figure 2: Flow of value to stakers.

7.2 Fees

Table 2: Fees. Bounds are compiled into the program; the administrator can change rates only within them.

Fee	Rate	Bound	Paid by
Release	0.5% of principal released	$\leq 2\%$	Payer, with the principal
Wager	0.5% of the pot	$\leq 2\%$	The pot, on payout
Yield	10% of yield	$\leq 25\%$	Yield, on redemption
Refund, void, cancellation	0	—	—

For example, a 10,000 USDC escrow held four months at 5% yield pays a 50 USDC release fee and earns about 167 USDC of yield, of which the protocol keeps 17 and the payer receives 150. Of the resulting 67 USDC of revenue, 33 goes to active stakers.

Because capacity scales with staked value, the largest agreement the council can decide grows only as stake grows, and staker revenue grows with settled volume. The relationship also runs in reverse: if volume falls, rewards fall, stakers may leave after the cooldown, and larger agreements revert to the multisig.

8 Agents and Safety

Agents use the protocol through a TypeScript SDK, an MCP server exposing deal, drafting, and juror tools, and an `escrow` scheme for x402: a server answers HTTP 402 with a terms template, the client funds an agreement in one transaction, and the server's signed delivery receipt acts as the verifier attestation, releasing funds immediately. Agent parties can link a Solana Agent Registry identity, and every wallet accrues an on-chain record of deals, volume, and disputes.

Because agents read counterparty text, all inbound content (terms, chat, evidence) passes a rule-based and a model-based screen. Flagged items stay visible with a reason, and MCP output marks party content as untrusted data. Agent wallets can set a policy (maximum deal size, allowed assets, minimum review window) that the SDK enforces. In disputes, jurors may vote that a side attempted manipulation; if two-thirds of the weight agrees, that side forfeits its bond whatever the ruling. An optional drafting assistant turns plain language into structured terms and flags vague criteria. It never signs transactions or rules, since one model advising every juror would be a single point of failure.

9 Implementation

The protocol is two Anchor programs on Solana. The *escrow* program holds agreements, seats, milestones, vaults, loans, and disputes; the *council* program holds stakes, delegates, cases, votes, reputation, and reward epochs. Escrow opens council cases by cross-program invocation signed by its program-derived authority. A keeper submits the permissionless instructions that advance deadlines, draws, tallies, and epochs; anyone else may submit them too. Chat is signed off-chain, with a Merkle root recorded on-chain at each state change and the transcript pinned to Arweave when a dispute opens. Upgrade authority and configuration sit in a Squads multisig, which can adjust parameters only

within compiled bounds (Appendix A) and pause new agreements, but cannot move funds. Version 1 supports USDC, then USDT and PYUSD.

10 Risks

1. *Regulation.* A token whose stakers share revenue may be treated as a security; escrow, lending, and wagers may each require licences. Legal review precedes launch.
2. *Thin early council.* While stake is small, most disputes go to the multisig, so early users rely on the operator. Multisig rulings will be published.
3. *Low early rewards.* Without emissions, rewards depend on volume.
4. *Adjudication attacks.* Bribery, stake concentration, and sustained price manipulation, as discussed in §6.
5. *Yield venue.* A Kamino failure would affect escrowed funds; yield can be disabled per agreement.
6. *Implementation.* The programs are in development and unaudited.

11 Conclusion

Heldra separates the two jobs of an escrow intermediary. A program with closed payout paths and an optimistic default handles custody and release. A stake-weighted council, whose authority over any dispute is bounded by the stake behind it and whose members are paid only from the activity they secure, handles adjudication. The result is conditional payment at a fraction of incumbent cost, with a dispute layer whose security budget is explicit.

References

[1] Coinbase. *x402: An open standard for internet-native payments.* 2025.

[2] ERC-8183: Agentic Commerce. Ethereum Improvement Proposals, draft, 2026.

[3] C. Lesaege, F. Ast, W. George. *Kleros: Short Paper v1.0.7.* 2019.

[4] T. C. Schelling. *The Strategy of Conflict.* Harvard University Press, 1960.

[5] A. Rundgren, B. Jordan, S. Erdtman. *RFC 8785: JSON Canonicalization Scheme.* IETF, 2020.

[6] Kamino Finance. *Kamino Lend documentation.* https://docs.kamino.finance

[7] Switchboard. *Randomness documentation.* https://docs.switchboard.xyz

Appendix A Parameters

Table 3: Defaults and compiled bounds.

Parameter	Default	Bound
Review window	72 h	5 min – 30 d
Wager confirmation / dispute response / appeal window	72 h / 48 h / 24 h	fixed
Dispute bond	max(5% of D, 5 USDC)	fixed
Milestones per agreement / members per side	16 / 8	fixed
Warm-up delay and ramp; unstake cooldown	14 d each	1 – 90 d
Lock multiplier	$1.25\times$ – $2.5\times$	$1\times$ – $5\times$, locks $\leq$ 12 months
Slash per incoherent vote or no-show	0.1%	$\leq$ 1%
Capacity ratio k / aggregate fraction θ	10 / 33%	2 – 100 / 1 – 50%
Staker share of fees	50%	fixed
Reward epoch	7 d	1 – 30 d
Guardian freeze / sunset	$\leq$ 7 d / 12 months	$\leq$ 7 d / $\leq$ 730 d

Notice. This document describes a protocol under development and is for information only. It is not an offer to sell or a solicitation to buy any token, and nothing in it is financial, legal, or tax advice. The programs are unaudited. Yield is variable and not guaranteed. Staker rewards depend on protocol activity and may be small or zero. Legal review is pending.