KonnectAI Trader

Ein systematisches Framework für LLM-gestützten Krypto-Handel
— Implementierungs-Papier

by **Kai Zeh**

v2.1 — May 2026

Inhaltsverzeichnis

0.1 Abstract / Executive Summary

Algorithmischer Krypto-Handel im Retail-Bereich ist mit grossem Abstand ein Verlustgeschäft. Unabhängige Erhebungen beziffern den Anteil verlustträchtiger Retail-Algo-Trader auf rund 92 %.[1] Ein NBER-Working-Paper aus 2024 quantifiziert die strukturelle Hauptursache: Execution-Slippage — die Differenz zwischen dem Preis, den ein Modell sieht, und dem Preis, zu dem eine Order tatsächlich ausgeführt wird — erklärt 34–67 % der realisierten Underperformance gegenüber Backtests.[2] Selten ist das Signal das Problem. Execution, Regime-Blindheit und das Fehlen eines übergreifenden Signal-Kontexts sind es.

KonnectAI Trader ist ein systematisches Framework, das diese drei Schwachstellen direkt adressiert. Anstatt eine Kryptowährung als univariate Preisreihe zu behandeln und ein weiteres neuronales Netzwerk darauf zu trainieren — ein gut dokumentierter Pfad zu Overfitting[3] — behandeln wir jede Trade-Entscheidung als strukturiertes analytisches Problem und delegieren sie an ein Large Language Model (LLM), dessen Kontextfenster gezielt um (i) zeitrahmenübergreifende technische Indikatoren, (ii) eine deterministische Marktregime-Klassifikation, (iii) On-Chain-Flow- und Netzwerk-Health-Metriken und (iv) eine über einen Strategie-Modus konfigurierbare Risikohaltung erweitert wurde. Ein regelbasierter Pre-Filter schränkt das Kandidaten-Universum vor jedem LLM-Call ein — sowohl als Quality-Gate als auch zur Begrenzung der Inferenzkosten. Jede LLM-Ausgabe durchläuft deterministische Risiko-Validatoren — Confidence-Gate und Mindest-Risk/Reward-Verhältnis — bevor sie eine Live- oder Paper-Position öffnen kann.

[1] Kalena Research. *"Regime-Aware Crypto Algorithmic Trading: A 2026 Practitioner Review."* März 2026.
[2] National Bureau of Economic Research. Working Paper 31890. *"Slippage and Performance in Algorithmic Trading."* 2024.
[3] Viprasol Research. *"Algorithmic Crypto Trading in 2026: State of the Art."* Jährliche Review, 2026.

Jede Entscheidung wird mit ihrer vollständigen Begründung und ihrem vollständigen Kapitalereignis-Kontext in einem auditfähigen Speicher persistiert.

Dies ist kein spekulativer Vorschlag. **Das System läuft seit Januar 2026 live mit Kai Zehs eigenem Kapital, und der aktuelle Feature-Stand ist seit April 2026 in kontinuierlichem Produktionsbetrieb.** Die Methodik spiegelt das empirische Muster wider, das das CryptoTrade-Paper der NUS Singapore (arXiv 2407.09546[4]) und der FinRL-DeepSeek-Contest 2025[5] validiert haben: ein LLM mit strukturiertem, multi-source-Kontext und reflektierender Evaluations-Loop schlägt klassische Zeitreihen-Baselines und Buy-and-Hold über mehrere Coins und Marktphasen. Wir erweitern dieses Muster um (i) bidirektionale Long-und-Short-Execution auf Binance Futures, (ii) eine Regime-Change-Exit-Logik, die Positionen schliesst, wenn das zugrundeliegende Regime gegen den Trade kippt, (iii) eine wallet-basierte Netto-PnL-Berechnung, die Ein- und Auszahlungen sauber abbildet, und (iv) eine kapitalsegregierte Multi-Account-Architektur, geeignet für institutionelles Pooling.

Was wir von Investoren erwarten. KonnectAI Trader hat die Build-out-Phase abgeschlossen. Wir suchen kein Kapital, um ein Team aufzubauen; wir öffnen einen kleinen, hart begrenzten Pool von Trading-Kapital-Slots auf dem operativen System. Das Modell ist **vier Investoren-Slots × CHF 100’000 = CHF 400’000 Trading-Kapital total** pro Server-Cluster. Jeder Slot ist in zwei Tranchen strukturiert: **CHF 50’000 direkt bei Vertragsunterzeichnung** und eine **zweite Tranche von CHF 50’000, fällig einen Monat nach Bereitstellung der Produktions-Hardware und funktionalem Live-Trading unter Vertragsbedingungen**. Die zweite Tranche kann nach Wahl des Investors über Treuhand- oder Bankgarantie abgesichert werden. Operative Kosten werden mit **CHF 6’250/Monat pro Investor (CHF 25’000/Monat total, CHF 300’000/Jahr)** geteilt und monatlich automatisch via Binance Pay vom investor-eigenen Binance-Account zum KonnectAI-Betreiber-Wallet transferiert, gesteuert über einen investor-kontrollierten Settings-Toggle (Default ON).

Was Investoren erhalten. KonnectAI Trader ist ausdrücklich **kein Subscription-Produkt, kein Signal-Service, kein Fonds und kein Custody-Service**. Es ist eine **non-custodial Capital-Access-Plattform**. Jeder Investor unterhält sein **eigenes Binance-Account**, jederzeit voll in seinem Eigentum und unter seiner Kontrolle, und stellt KonnectAI einen **gescopten API-Key (Spot-Trading, Futures-Trading, Universal Transfer / Binance Pay für den Auto-Pay-Flow — Withdrawal permanent deaktiviert, IP-Whitelist)** zur Verfügung. Die operative Logik — der LLM-Analyser, der Regime-Klassifizierer, der Pre-Filter, die Validatoren — ist gepoolt (eine Analyse pro Cycle, Fan-out via die vier Investor-API-Keys, Position-Sizing nach individueller Account-Equity); das Kapital selbst verbleibt im direkt im Eigentum stehenden Account jedes Investors. Der Investor behält das rechtliche Eigentum an allen Assets und kann den API-Zugriff jederzeit über das

[4] Li, Z. et al. (National University of Singapore). *“CryptoTrade: A Reflective-LLM-Agent Framework for Cryptocurrency Trading.”* arXiv:2407.09546, 2025.

[5] *“FinRL-DeepSeek Contest: Reinforcement Learning with LLM-Derived Sentiment Features for Financial Decision Tasks.”* Contest-Bericht, 2025.

Binance-UI widerrufen.

Performance-Projektionen. Wir projizieren drei regimeabhängige Daily-ROI-Szenarien auf Basis der MVP-Daten seit Januar 2026. Wir framen diese Zahlen als Projektionen, nicht als Track Record:

Marktregime	Daily-ROI-Projektion	Annualisiert (linear, 365 Tage)
Drawdown	0.35 %	+127 % p.a.
Sideways	0.50 %	+180 % p.a.
Bullish	0.65 %	+237 % p.a.

Diese Werte liegen materiell über dem 15–30 % p.a. Korridor, den die akademische Literatur für LLM-augmentierte Systeme dokumentiert.[6] Der Differenzierer ist operativ, nicht theoretisch: (i) höhere Execution-Frequenz (1'440 LLM-Calls/Tag im Produktions-Cron versus täglich/4h-Zyklen in publizierten Studien), (ii) bidirektionale Long-und-Short-Execution auf Binance Futures (akademische Baselines sind typischerweise Long-only), (iii) regime-aware Execution inklusive Regime-Change-Exit-Branch, der in publizierten Designs nicht vorkommt, (iv) adaptives Tuning von Prompts und Thresholds auf Live-Daten, und (v) wallet-basierte, kapitalereignis-gewahre Buchhaltung, die die realisierte Rendite auf das tatsächlich eingesetzte Kapital erfasst — nicht auf eine Paper-Simulation.

Defensibilität. Unsere Edge ist kein geheimer Indikator-Threshold. Sie ist die Komposition: regelbasiertes Pre-Filtering + Regime-Klassifikation + On-Chain-Integration + disziplinierte LLM-Analyse + bidirektionale Execution + Regime-Change-Exit + deterministische Validatoren + kapitalereignis-gewahres Audit. Jede Komponente ist ersetzbar, keine ist das ganze System, und die Integration selbst ist das geistige Eigentum. Autor Kai Zeh bringt 30+ Jahre Erfahrung in System- und Plattform-Architektur mit, einschliesslich einer früheren CTO/CEO-Rolle, in der er eine von BDO mit über CHF 400 Millionen bewertete Plattform architektiert hat (Abschnitt 14).

Kapital-Denominierung. Trading-Kapital wird in **USDT** auf Binance gehalten. CHF-Beträge in diesem Dokument — CHF 100'000 Slot-Grösse, CHF 6'250/Monat Operating-Cost-Anteil, CHF-denominierte Performance-Projektionen — sind **Referenzwerte für Schweizer Investoren**; tatsächliche Positionsgrössen werden in USDT gegen den prevailing CHF/USDT-Spot-Kurs berechnet. Investoren können einen beliebigen Betrag in ihr Binance-Account einzahlen; die CHF 100'000-Werte sind Beispiel-Deployments, keine vertraglich vorgeschriebenen Mindestguthaben innerhalb des Accounts.

Das restliche Dokument beschreibt die theoretischen Grundlagen (Abschnitt 2), die konzeptuelle Architektur (Abschnitt 3), den Analyser und seine Validatoren (Abschnitt 4), Regime-Detection und den Regime-Change-Exit-Branch (Abschnitt 5), On-Chain-Integration (Abschnitt 6), Strategie-Modi und Position-Sizing (Abschnitt 7), die Rolle

[6] Li, Z. et al. (National University of Singapore). *"CryptoTrade: A Reflective-LLM-Agent Framework for Cryptocurrency Trading."* arXiv:2407.09546, 2025.

0.2 1. Einleitung & Motivation

0.2.1 1.1 Das Retail-Algo-Trading-Problem

Die dominierende Erzählung rund um Retail-Krypto-Handel in den Jahren 2020–2024 lautete, dass algorithmische Tools und Backtesting-Plattformen quantitative Strategien demokratisieren würden. Die empirische Bilanz hat diese Erzählung nicht bestätigt. Eine Felderhebung von Kalena Research im März 2026 schätzt, dass rund 92 % der Retail-Algo-Trader auf einem 12-Monats-Rückblick Geld verlieren.[7] Dies steht in einer Linie mit älteren Industriedaten zum Retail-Aktien-Daytrading (Barber & Odean[8]), ist im Krypto-Bereich aber aus drei Gründen tendenziell sogar schwerwiegender:

1. **Dünnere Order-Books pro Venue**, insbesondere ausserhalb der BTC/ETH-Perpetuals. Was aggregiert wie ein liquider Markt aussieht, ist über Dutzende von Venues fragmentiert; die tatsächliche Queue-Tiefe an einem gegebenen Tick ist häufig eine Grössenordnung kleiner als das angezeigte Volumen vermuten lässt.
2. **24/7-Marktöffnung**, was unaufmerksame diskretionäre Eingriffe bestraft und Systeme belohnt, die entweder durchgängig laufen oder gar nicht handeln.
3. **Schnelle, gewaltsame Regime-Wechsel**, oft in 48–96 Stunden zwischen Trend- und Range-Bedingungen — Horizonte, auf denen klassische Statisch-Parameter-Systeme nicht adaptieren.

Das NBER-Working-Paper w31890[9] aus 2024 quantifiziert die mechanische Hauptursache der Retail-Algo-Underperformance: **Execution-Slippage**. Das Paper zerlegt die Backtest-zu-Live-Performance-Lücke und zeigt, dass 34–67 % davon — je nach Venue und Order-Grösse — auf den Unterschied zwischen Signal-Preis und tatsächlichem Fill-Preis zurückgehen, netto Book-Tiefe, Queue-Priorität und adversem Fill bei News. Die

[7] Kalena Research. *"Regime-Aware Crypto Algorithmic Trading: A 2026 Practitioner Review."* März 2026.

[8] Barber, B. M., und Odean, T. *"Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors."* Journal of Finance, 55(2): 773–806, 2000.

[9] National Bureau of Economic Research. Working Paper 31890. *"Slippage and Performance in Algorithmic Trading."* 2024.

Signal-Logik selbst ist ein sekundärer Faktor.

Diese diagnostische Aussage ist zentral. Sie impliziert, dass ein System, das nur ein bekanntes Signal automatisiert, ohne Execution-Realismus einzubauen, systematisch zu optimistisch über seine Live-Performance ist. Sie impliziert auch, dass Verbesserungen der Signal-Qualität jenseits eines bestimmten Punkts abnehmende Erträge bringen, wenn nicht parallel die Execution adressiert wird — und dass ein System, das **live, mit echtem Kapital, inklusive Gebühren und Slippage gemessen** wurde, einen strukturellen Glaubwürdigkeitsvorteil gegenüber Systemen hat, deren Performance ausschliesslich aus Backtests oder Paper-Trading berichtet wird.

0.2.2 1.2 Warum LLMs die Rechnung verändern

Die 2020–2023er Generation Retail-Trading-Tooling fokussierte auf technische Indikatoren und zunehmend auf ML-Modelle, die direkt auf Preisreihen trainiert wurden. Viprasols State-of-the-Art-Review 2026 kommt zum Schluss: **Direkt auf Preisreihen trainierte neuronale Netze bleiben eine Overfitting-Falle**. Sie memorisieren, sie generalisieren nicht, und sie versagen katastrophal an Regime-Grenzen.[10] Wo ML im Krypto-Bereich erfolgreich war, dann in engeren Aufgaben — Volatilitäts-Spike-Prognose, Cross-Asset-Korrelation unter Stress, Execution-Optimierung — und nicht in direkter Return-Prognose.

Was die Rechnung verändert hat, ist das Aufkommen von General-Purpose-LLMs, die zu *strukturiertem Reasoning über heterogene Inputs* fähig sind. Das CryptoTrade-Paper der NUS Singapore (arXiv 2407.09546, 2025) zeigt empirisch, dass ein LLM-Agent mit (i) On-Chain-Metriken, (ii) Off-Chain-News und (iii) einem reflektierenden Evaluations-Loop Zeitreihen-Baselines und Buy-and-Hold über BTC, ETH, SOL und mehrere Altcoins über mehrere Phasen des 2022–2024er Markts schlägt.[11] Der FinRL-DeepSeek-Contest 2025 generalisiert das Muster, indem er LLM-abgeleitete Sentiment-Features als Eingang für RL-basierte Execution-Agents formalisiert.[12]

Der Insight: Das LLM ist kein "besserer Indikator". Es ist eine andere Klasse von Komponente — ein *Analyst*, der heterogene, teilweise redundante Inputs verarbeitet und eine disziplinierte Entscheidung mit explizitem Reasoning ausgibt. Drei Eigenschaften zählen:

- **Input-Heterogenität.** Ein LLM kann RSI auf einer 4-Stunden-Kerze, Exchange-Outflow in BTC-Einheiten und eine Regime-Klassifikation in einem einzigen Prompt verarbeiten — ohne handgeschriebene Feature-Engineering-Schritte.
- **Reasoning-Artefakt.** Das LLM produziert eine natürlichsprachige Begründung neben seiner strukturierten Ausgabe. Diese Begründung ist auditierbar, vom Operator

[10] Viprasol Research. *"Algorithmic Crypto Trading in 2026: State of the Art."* Jährliche Review, 2026.

[11] Li, Z. et al. (National University of Singapore). *"CryptoTrade: A Reflective-LLM-Agent Framework for Cryptocurrency Trading."* arXiv:2407.09546, 2025.

[12] *"FinRL-DeepSeek Contest: Reinforcement Learning with LLM-Derived Sentiment Features for Financial Decision Tasks."* Contest-Bericht, 2025.

reviewbar und für Post-hoc-Diagnostik unentbehrlich.

- **Strukturierter Output unter Schema-Constraint.** Moderne LLMs liefern unter Schema-Anweisung verlässlich JSON-konforme Ausgaben (action, confidence, entry, stop, take-profit-Levels, reasoning), die direkt in deterministische Risiko-Validatoren und Execution-Logik einfliessen können.

0.2.3 1.3 Was wir anders machen

KonnectAI Trader ist eine Übung in *Integrations-Disziplin*, nicht in neuartigen Algorithmen. Jede Komponente, die wir verwenden — technische Indikatoren, Regime-Klassifikation, On-Chain-Metriken, LLM-basiertes Reasoning, Risiko-Validatoren — ist isoliert gut verstanden. Der Beitrag ist die Komposition, und konkret:

1. **Ein Pre-Filter vor dem LLM.** Wir lassen das LLM nicht jeden Coin in jedem Cycle bewerten. Ein regelbasierter Pre-Filter identifiziert eine kleine Anzahl Kandidaten pro Cycle anhand multi-trigger-Kriterien. Das ist sowohl ein Cost-Gate (LLM-Inferenz ist nicht gratis) als auch ein Quality-Gate (das LLM sieht nur Situationen, die eine Analyse rechtfertigen).
2. **Strukturierter Kontext, kein Free-Form-Prompt.** Das LLM erhält ein deterministisches Prompt-Bundle mit fixem Schema — Symbol-Kontext, zeitrahmenübergreifende Indikatoren, Regime-Klassifikation, On-Chain-Block, Strategie-Modus-Block — und liefert eine deterministische JSON-Antwort mit fixem Schema. Das ist Engineering, nicht Prompting.
3. **Deterministische Validatoren nach dem LLM.** Das LLM kann irren. Confidence-Gate und Risk/Reward-Validator werden auf jede Ausgabe angewandt und können die Entscheidung zurückstufen oder verwerfen.
4. **Bidirektionale Execution von Tag eins.** Das System handelt Long und Short auf Binance Futures, mit nativem Long-oder-Short-Signal-Output aus dem Analyser. Performance-Kontinuität durch Bull-, Bär- und Sideways-Phasen ist strukturell, nicht Wunschdenken.
5. **Regime-Change-Exit.** Wenn der Regime-Klassifizierer gegen einen offenen Trade kippt und der Trade bereits im Profit über dem Round-Trip-Fee-Threshold liegt, wird die Position automatisch geschlossen. Dieser Branch fehlt unseres Wissens in publizierten akademischen Implementationen.
6. **Wallet-basierte, kapitalereignis-gewahre Buchhaltung.** Jede Ein- und Auszahlung sowie jeder interne Spot↔Futures-Transfer wird über geometrisches Period-Linking in die Equity-Curve einbezogen. Netto-PnL wird als Prozent vom *eingesetzten Kapital* gemeldet, nicht als Paper-Simulationswert.
7. **Vollständiger Audit-Trail.** Jeder Kandidat, jede Analyse, jeder Trade, jede Parameteränderung, jedes Kapitalereignis wird mit Zeitstempel und — bei Parameteränderungen — mit Zuschreibung zum auslösenden Nutzer persistiert.

Dies ist keine neuartige Trading-Idee. Es ist ein disziplinierter Engineering-Ansatz in einer

Domäne, in der die meisten Teilnehmer underdiscipliniert agieren.

0.3 2. Theoretische Grundlagen

0.3.1 2.1 Technische Analyse-Indikatoren

Ein rigoroses System muss technische Indikatoren als das behandeln, was sie sind: zusammenfassende Statistiken auf Preis und Volumen. Sie sind isoliert nicht prognostisch; sie sind als Inputs in eine breitere Entscheidung sinnvoll. Wir rekapitulieren kurz die genutzten Indikatoren und die Semantik, die für den LLM-Prompt zählt.

Relative Strength Index (RSI). Welles Wilders Formulierung von 1978[13] misst das Verhältnis durchschnittlicher Gewinne zu durchschnittlichen Verlusten über ein gleitendes Fenster (typisch 14 Perioden):

$$\text{RSI} = 100 - \frac{100}{1 + RS}, \quad RS = \frac{\overline{\text{gain}}_{14}}{\overline{\text{loss}}_{14}}$$

Werte unter einer unteren Schwelle (konventionell 30, in Hochvolatilitäts-Assets häufig tiefer angepasst) gelten als überverkauft; Werte über der oberen Schwelle (konventionell 70) als überkauft. Der häufig ignorierte Caveat: in einem starken Trend kann RSI für längere Zeit in der "überkauften" oder "überverkauften" Zone bleiben, ohne dass eine Umkehr erfolgt. RSI ist allein kein Entry-Signal; es ist ein Kontext-Hinweis.

Moving Average Convergence/Divergence (MACD). Die Differenz aus 12- und 26-Perioden-EMA, geglättet durch einen 9-Perioden-EMA dieser Differenz. Ein "Flip" — MACD-Linie kreuzt ihre Signal-Linie — wird als kurzfristiger Momentum-Wechsel interpretiert. Die Hauptschwäche liegt im Chop, wo es zu hoher Rate von False Flips kommt.

Bollinger Bänder. Ein gleitender Durchschnitt mit Bändern bei ±2 Standardabweichungen. Berührung des oberen Bandes signalisiert gestreckte Aufwärtsbewegung; das untere Band gestreckte Abwärtsbewegung. In einem Trend-Regime kann der Preis das Band "begehen" — d.h. das obere Band über viele Kerzen entlangwandern, ohne Mean-Reversion. Unsere Filter-Logik respektiert das explizit, indem wir nie ein doppelseitiges Bollinger-Signal im selben Cycle auf demselben Symbol auslösen.

Gleitende Durchschnitte (EMA 20 / 50 / 200 und SMA 20 / 50). Hauptzweck als Regime-Filter. Preis über dem 200er-EMA auf Tagesbasis ist ein grober, aber robuster Langfrist-Trend-Indikator. Crossings zwischen Kürzer-Horizont-EMAs (z.B. 20/50) sind Mikro-Trend-Signale.

Average True Range (ATR). Ein Volatilitätsmass in absoluten Termini. Wir verwenden den

[13] Wilder, J. Welles. *New Concepts in Technical Trading Systems.* Trend Research, 1978.

auf den aktuellen Preis normierten ATR als Input für unseren Regime-Klassifizierer und als Referenz für volatilitätsangepasste Stop-Distanzen.

Volumen. Rohes gehandeltes Volumen ist verrauscht; das Verhältnis aus aktuellem Volumen zu einem kurz-horizontigen Trailing-Average ist ein saubereres Signal für ungewöhnliche Aktivität. Wir verwenden eine *Completed-Candle*-Baseline, um Verzerrung durch die laufende Kerze zu vermeiden — eine subtile, aber konsequenzenreiche Quelle von Bias in vielen Open-Source-Regime-Klassifikatoren.

Der kombinierte Einsatz dieser Indikatoren — anstelle der Verlassens auf einen einzelnen — ist, was akademische Meta-Studien zur technischen Analyse konsistent empfehlen.[14] Das LLM ist die Komponente, die sie gegeneinander gewichtet.

0.3.2 2.2 Markt-Regime-Theorie

Ein Marktregime ist ein latenter Zustand, der die bedingte Verteilung der Preisbewegungen formt. Im Kalena-2026-Framework, das wir als Ausgangspunkt übernehmen, werden fünf Regime benannt:

1. **Trend-up, liquide.** Klare Aufwärts-Steigung auf Kürzer-Horizont-EMAs, ATR im Normalbereich, Volumen bestätigend.
2. **Trend-down, liquide.** Symmetrisch: klare Abwärts-Steigung, ATR normal, Volumen bestätigend.
3. **Range.** EMA-Slopes flach, Preis oszilliert in einem engen Kanal, Volumen durchschnittlich oder darunter.
4. **Volatility-Expansion.** ATR scharf erhöht, kein klarer Richtungs-Slope, typisch vor oder nach News-Ereignissen.
5. **Crash.** Heftige Abwärtsbewegung mit erhöhtem ATR und schwächelnder Struktur (z.B. Preis unter 200-EMA, RSI tief negativ, Liquidität dünner werdend).

Der zentrale Insight der Regime-Literatur[15][16]: *Dasselbe Signal hat in verschiedenen Regimes verschiedene Erwartungswerte*. Eine Bollinger-Band-Berührung ist im Range-Regime mean-reverting und im Trend-Regime fortsetzungsfreundlich. Ein Volume-Spike im Vol-Expansion-Regime ist ein Fortsetzungssignal; im Range-Regime häufig ein Fake-out. Ein gut entworfenes System muss daher entweder (i) das Regime klassifizieren und die Signal-Interpretation entsprechend anpassen, oder (ii) ein Modell — ein LLM — verwenden, das zu solchem bedingten Reasoning fähig ist, wenn das Regime explizit im Kontext steht.

Wir wählen Ansatz (ii), aber mit einem deterministischen Regime-Klassifizierer als Input — nicht in der Erwartung, dass das LLM das Regime aus Rohdaten im selben Call ableitet. Das trennt Concerns und macht die Regime-Klassifikation eigenständig auditierbar und

[14] Park, C.-H., und Irwin, S. H. *"What Do We Know About the Profitability of Technical Analysis?"* Journal of Economic Surveys, 21(4): 786–826, 2007.

[15] Kalena Research. *"Regime-Aware Crypto Algorithmic Trading: A 2026 Practitioner Review."* März 2026.

[16] Perfumo, Thomas (Kraken Chief Economist). *"Why This Cycle Isn't Like the Others."* Kraken Research, Februar 2026.

backtestbar. Wir erweitern das Muster um einen **Regime-Change-Exit-Branch** (Abschnitt 5.5): Wenn der Klassifizierer gegen eine offene Position kippt und die Position bereits im Profit über dem Round-Trip-Fee-Threshold ist, wird der Trade unabhängig von Stop und Take-Profit geschlossen.

Kalenas eigene Case-Study-Daten[17] dokumentieren, dass eine Regime-Switching-Strategie (Sharpe > 1.5 im Studienzeitraum) eine regime-agnostische Variante derselben Basis-Strategie auf denselben Daten materiell schlägt. Wir behaupten nicht, dass unsere Implementation jenen exakten Sharpe-Wert reproduziert; wir behaupten, dass Regime-Awareness die strukturelle Verbesserung ist, die die Literatur am konsistentesten identifiziert — und dass Regime-Change-Exit eine natürliche Erweiterung dieses Prinzips in die Execution-Schicht ist.

0.3.3 2.3 On-Chain-Signale

On-Chain-Metriken sind das Krypto-Äquivalent makroökonomischer Indikatoren: langsamer-bewegt, struktureller, in traditionellen Asset-Klassen nicht verfügbar. Es ist die Domäne, in der Glassnode, Nansen und Arkham Intelligence kommerzielle Praxen aufgebaut haben. Die Kategorien, auf die wir uns stützen:

Exchange-Flows. Inflows zu Börsen gehen typisch dem Verkauf voraus; Outflows gehen typisch der Akkumulation voraus oder begleiten sie. Glassnodes publizierte Forschungspraxis behandelt 7-Tage-rollierende Net-Flow-Deltas als Frühindikator für kurz- bis mittelfristige Richtungsbias.[18]

Wallet-Konzentration und Whale-Tracking. Verhalten grosser Halter (Top-100- und Exchange-Cluster-Wallets) ist ein nicht-trivialer Anteil der Richtungsvarianz. Methodisch identifizieren wir Custody-Wallets über gelabelte On-Chain-Adressen (Etherscan-Labels im Free-Tier; Arkham- oder Nansen-Label-Datenbanken im Paid-Tier) und aggregieren per Cluster, mit 24-Stunden- und 7-Tage-Deltas.

Network-Health. Bitcoin-Hashrate, Mempool-Tiefe, Fee-Märkte, aktive Adressen. Langsam-bewegte Strukturindikatoren — nicht nützlich für Intra-Minute-Entry-Timing, aber stark nützlich, um die breitere Marktphase zu rahmen (Akkumulation, Distribution, Capitulation).

Gas-Preis als Regime-Indikator (Ethereum). Ethereum-Gas-Preis ist ein aggregiertes Mass für Blockspace-Nachfrage. Anhaltend tiefes Gas in einer Bull-Phase ist häufig anomalisch und kann nachlassende Nachfrage signalisieren; Spikes um Major-Events sind eine Fortsetzungs-Bestätigung für das betreffende Asset.

---

[17] Kalena Research. *"Regime-Aware Crypto Algorithmic Trading: A 2026 Practitioner Review."* März 2026.
[18] Glassnode. *"Week 16/2026 On-Chain Report."* Glassnode Insights, April 2026.

0.3.4 2.4 Makro und Sentiment

Über Asset-spezifische Technicals und On-Chain hinaus sind drei aggregierte Indikatoren relevant:

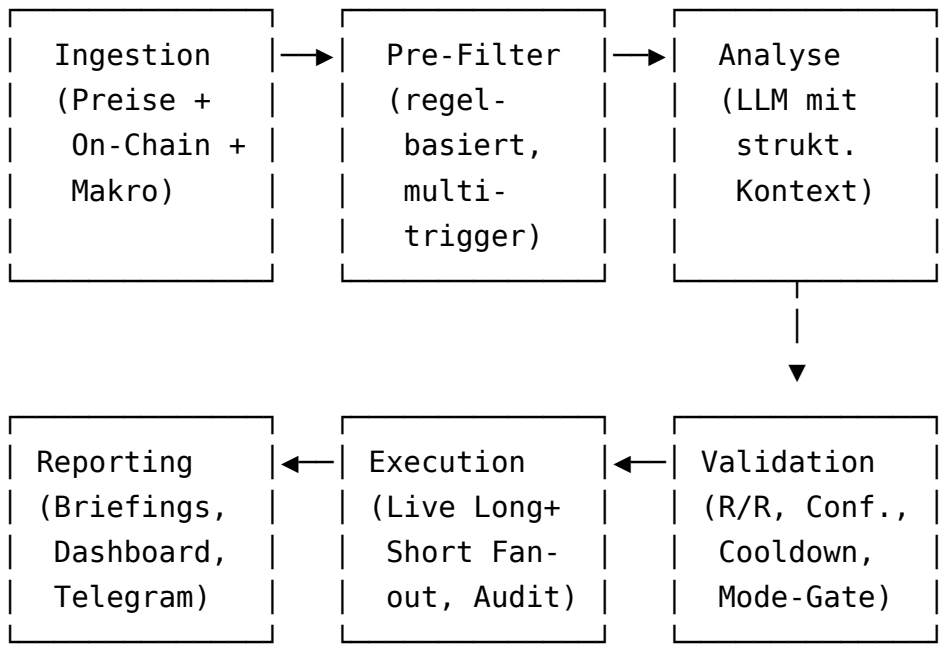
- **DeFi-TVL (Total Value Locked).** Ein Proxy für im Ökosystem gebundenes Kapital über Chains hinweg. Trended-TVL (DefiLlama-Daten) ist ein längerhorizontiges Posture-Signal.
- **Fear & Greed Index (Alternative.me).** Ein Composite-Sentiment-Index 0–100. Extreme Werte (< 10 oder > 90) sind contrarian, nicht confirmatory.
- **BTC-Dominanz.** BTC-Marktkapitalisierung als Anteil an der gesamten Krypto-Marktkapitalisierung. Trended-Dominanz-Bewegungen signalisieren Rotation zwischen BTC und Altcoins.
- **Stablecoin-Supply-Dynamik.** Wachstum impliziert "Dry Powder" im System; Kontraktion impliziert Kapitalabfluss.

Diese Makro-Signale fliessen in unseren Wochen-Makro-Bericht und indirekt über Strategie-Modus und Regime-Klassifizierer in den LLM-Prompt; sie sind keine Per-Trade-Inputs, weil sie sich für Intraday-Entscheidungen zu langsam bewegen.

0.4 3. Systemarchitektur (Konzeptuell)

0.4.1 3.1 Pipeline-Übersicht

Auf konzeptueller Ebene ist die KonnectAI-Trader-Pipeline ein sechsstufiger Fluss in fixer Kadenz:



Jede Stufe ist ein eigenes, testbares Modul. Die Schnittstellen sind streng typisiert: Pre-Filter emittiert einen Kandidaten-Record, Analyser konsumiert Kandidaten und emittiert Analysen, Validator konsumiert Analysen und emittiert Entscheidungen, Executor konsumiert Entscheidungen und emittiert Trades. Der Reporting-Layer ist ein Read-only-Konsument der Persistenzschicht.

Die **Cron-Kadenz ist Konfiguration**, nicht Architektur. Der MVP läuft auf einem 15-Minuten-Core-Loop (96 Cycles/Tag). Die Produktions-Deployment läuft auf einem 1-Minuten-Core-Loop (1'440 Cycles/Tag) mit feinerkörnigen Candle-Daten. Die feinere Granularität hebt die Obergrenze der Signal-Frequenz und ist der primäre Throughput-Differenzierer gegenüber der akademischen Literatur. Die Pipeline-Form bleibt identisch; nur der Schedule ändert sich.

0.4.2 3.2 Daten-Ingestion-Layer

Drei Sub-Streams:

Preis / OHLCV / Volumen. Candle-Daten von Binance' öffentlichen Market-Data-Endpoints, in mehreren Intervallen — 1-Minute, 5-Minuten, 15-Minuten, 1-Stunde, 4-Stunden, 1-Tag. Indikatoren werden je Ingest-Cycle server-seitig berechnet und neben den rohen Kerzen gespeichert. Die Ingest-Kadenz ist auf dem Produktions-Cron auf den 1-Minute-Candle-Close synchronisiert; auf dem MVP-Cron auf den 15-Minute-Candle-Close.

On-Chain. Snapshots der On-Chain-Metriken werden stündlich aus einem kurierten Quellen-Set abgerufen (Abschnitt 6.1). Jeder Snapshot schreibt eine Zeile mit `(source, asset, metric, value, raw_payload)`-Tupel. Raw-Payloads werden für Audit und retrospektive Metrik-Ableitung aufbewahrt. Wir verwenden ein Fail-Safe-Pattern: jede Quelle ist in ihren eigenen Exception-Handler eingehüllt; ein Quellen-Fehler blockiert die anderen nicht.

Makro / Sentiment. Stündliche Snapshots von DeFi-TVL (DefiLlama), Fear & Greed (Alternative.me), BTC-Dominanz (CoinGecko), globaler Marktkapitalisierung. Diese rollen in den Wochen-Makro-Bericht und in die Regime-Klassifikations-Inputs.

Alle drei Streams teilen das gleiche Persistenz-Pattern: append-only mit Captured-at-Zeitstempel, plus optional abgeleitete Spalten für häufig abgefragte Metriken. Das Raw-Payload bleibt für jede Zeile erhalten.

0.4.3 3.3 Pre-Filter-Layer

Auf jedem Cycle bewertet der Pre-Filter jeden Coin auf der aktiven Watchlist gegen ein Set multi-trigger Regeln. Trigger umfassen RSI-Schwellenüberschreitungen, Bollinger-Band-Nähe, EMA-Crossings (20/50), MACD-Signal-Linien-Flips und Volume-Spikes. Ein Kandidat wird emittiert, sobald mindestens N Trigger im selben Cycle auf demselben Symbol feuern (N ist ein Strategie-Modus-Parameter).

Der Pre-Filter dient zwei Zwecken:

1. **Kosten-Effizienz.** LLM-Inferenz ist der grösste variable Kostenposten in unserer Pipeline. Ein Pre-Filter, der die Kandidatenzahl von der Watchlist-Grösse auf eine kleine einstellige Zahl pro Cycle reduziert, spart eine Grössenordnung an Inferenz-Spend.
2. **Quality-Gate.** Selbst wenn Inferenz gratis wäre, wollten wir das LLM nicht jeden Coin in jedem Cycle analysieren lassen. Der Mehrwert des LLM kommt aus der Analyse *interessanter* Situationen. "Interessant" wird auf der Pre-Filter-Schicht regelhaft definiert.

Eine bewusste Engineering-Entscheidung: der Pre-Filter ist vollständig deterministisch und regelbasiert. Er enthält kein ML. Das macht ihn reproduzierbar, auditierbar und schnell iterierbar. Das LLM tritt erst stromabwärts in die Pipeline ein.

0.4.4 3.4 Analyse-Layer (LLM-basiert)

Der Analyse-Layer ist das Herz des Systems. Für jeden Kandidaten konstruiert der Analyser einen strukturierten Prompt (Abschnitt 4.2), der enthält:

- Symbol-Kontext (aktueller Preis, 24h-Veränderung, Marktkapitalisierungs-Rang).
- Indikator-Snapshot über mehrere Zeitrahmen.
- Recent-Candle-Historie.
- Marktregime-Klassifikation (Abschnitt 5).
- On-Chain-Kontext-Block (Abschnitt 6) — asset-spezifisch wo verfügbar, sonst aggregiert.
- Strategie-Modus-Block (Abschnitt 7) — die aktuell vom Operator gewählte Posture.
- Output-Schema-Anweisung — ein deterministisches JSON-Schema, das das LLM zu befüllen hat.

Das LLM liefert eine strukturierte Antwort: `action ∈ {BUY, SELL, HOLD, AVOID}`, ein `confidence`-Score, ein `entry`-Level, ein `stop`-Level, ein oder mehrere Take-Profit-Levels, eine `reasoning`-Narrative, ein kategoriales `risk_level` und ein kategoriales `timeframe`. Dasselbe Schema wird für Long- und Short-Richtung verwendet; eine SELL-Action auf Binance Futures öffnet eine Short-Position, schliesst nicht eine bestehende Long-Position.

Warum ein LLM? Drei Gründe, jenseits der akademischen Validierung in Abschnitt 1:

- *Kompositionales Reasoning.* Ein klassischer Scorer bräuchte explizite Feature-Interaktions-Logik für jede Signalkombination. Ein LLM produziert begründete Outputs aus heterogenen Inputs in einem einzigen Prompt.
- *Natürlichsprachige Begründung.* Jeder Trade kommt mit einer englischen Begründung, die auditiert, reviewt und für diagnostische Muster gemined werden kann. Klassische Systeme produzieren Scores; LLMs produzieren Argumente.
- *Adaptierbarkeit.* Strategie-Modus-Wechsel oder ein neues Signal hinzufügen ist ein

Prompt-Edit und ein Kontext-Bundle-Update — kein Modell-Retraining.

Warum Opus-Klasse-Reasoning konkret. Die Cost/Quality-Frontier hat sich 2026 entschieden in Richtung Reasoning-optimierter Modelle bewegt. Kleinere/günstigere Modelle sind kosten-attraktiv, performen aber bei multi-signal Finanz-Entscheidungs-Tasks in unseren internen Vergleichen materiell schlechter.[19] Wir nutzen aktuell die Anthropic-Claude-Opus-4-7-Familie; das Framework ist modell-agnostisch und Substitution ist eine Schema-Adaption.

Pooling: eine Analyse, viele Executions. Ein einzelner LLM-Call pro Cycle pro Kandidat liefert ein einzelnes strukturiertes Signal. Das Signal wird dann via die vier Investor-API-Keys auf die vier investor-eigenen Binance-Accounts gefächert, mit Position-Sizing gegen die jeweils verfügbare Equity des einzelnen Accounts. Das ist die ökonomische Kerneigenschaft der gepoolten Architektur (Abschnitt 11.4): Inferenz-Kosten skalieren mit Cycle-Frequenz, nicht mit der Anzahl Investoren.

0.4.5 3.5 Risiko-Management-Layer

Drei deterministische Filter werden auf jeden LLM-Output vor Trade-Eröffnung angewandt:

1. **Confidence-Gate.** Eine modus-abhängige Mindest-Confidence. Outputs unter Threshold werden auf HOLD heruntergestuft.
2. **Risk/Reward-Validator.** Das Verhältnis `(take-profit − entry) / (entry − stop)` für Long-Trades und seine symmetrische Form für Short-Trades muss eine modus-abhängige Mindest-Schwelle überschreiten. Outputs unter Schwelle werden auf HOLD heruntergestuft mit Confidence-Strafe. Dieser Validator fängt einen wichtigen LLM-Failure-Mode: das Modell schlägt gelegentlich technisch plausible Entries mit ungenügendem Reward pro Risikoeinheit vor.
3. **Cooldown.** Nach einem Stop-out auf einem Symbol verhindert eine Cooldown-Periode den unmittelbaren Re-Entry desselben Symbols. Schutz gegen das "Revenge-Trade"-Pattern.

Über diese Per-Trade-Gates hinaus enthält die Risk-Schicht: **Partial-Take-Profit mit Runner** (Abschnitt 8.2), **TP1-aware Position-Sizing** (Risk gegen den *originalen* Stop, nicht den TP1-trailed Stop), **Original-SL-Preservation** (TP1-Fill überschreibt nie den ursprünglich gewählten Risk-Boundary im Audit-Record), **Regime-Change-Exit** (Abschnitt 5.5), **Timeout-Closure** (Trades, die in definierter Zeit weder Target noch Stop erreichen) und **Daily-Cap-Alerts** gegen Signal-Alert-Fatigue.

0.4.6 3.6 Execution-Layer

Sämtliche Execution fliesst durch eine einzelne Execution-Schicht, mit **Live-Trading auf Binance Spot und Binance Futures** als Produktions-Posture. Das System läuft seit Januar

[19] Li, Z. et al. (National University of Singapore). *"CryptoTrade: A Reflective-LLM-Agent Framework for Cryptocurrency Trading."* arXiv:2407.09546, 2025.

2026 live auf Kais eigenem Kapital und gegen dieselbe Execution-Schicht, die die vier investor-eigenen Binance-Accounts in der Produktions-Deployment via deren jeweilige gescopte API-Keys bedienen wird.

Die Execution-Schicht adressiert beide Modi:

- **Spot (Long-only).** Verwendet für den Long-Bucket, wo Kapitaleffizienz und Einfachheit voller Kollateral-Positionen bevorzugt werden.
- **Futures (Long und Short).** Verwendet für Short-Seite und für kapitaleffizientes Long-Exposure. Hebel, Margin und Funding werden pro Trade getrackt. Stop- und Take-Profit-Levels werden als Exchange-seitige Conditional-Orders platziert.

Produktionsreife Execution-Mechanik:

- **Intrabar-Fills werden vom Exchange honoriert.** Der Exchange — nicht unser Paper-Simulator — ist die Autorität über Fills. Stops und Take-Profits sind Exchange-seitig als Conditional-Orders platziert; unsere Reconciliation-Schicht zeichnet Fills auf, sobald sie über den User-Data-Stream eintreffen.
- **Partial-Exits werden in-place geführt, nicht geklont.** Bei TP1-Fill wird `remaining_pct` reduziert, der Stop auf TP1 angehoben, und der Trade läuft im selben Record auf TP2 weiter. Der *originale* Stop-Level bleibt als separate Spalte für korrekte Risiko-Attribution erhalten. Dieses Vorgehen vermeidet die Buchhaltungs-Pathologie paralleler Klon-Trades.
- **Manual-Close ist eine First-Class-Action.** Ein Operator kann jeden offenen Trade vom Dashboard aus schliessen. Die Action wird mit Closing-Preis und Operator-Identität geloggt.
- **Regime-Change-Exit** (Abschnitt 5.5) ist mechanisch in den Cron-Loop eingebunden und operiert mit derselben Audit-Disziplin wie automatische Stop- und Take-Profit-Fills.
- **Capital-Event-Reconciliation.** Ein- und Auszahlungen sowie Spot↔Futures-Internal-Transfers werden je Cycle aus Binance gezogen und mit der Equity-Curve abgeglichen, mit Phantom-Dip-Prevention via Internal-Transfer-Bridge-Logik (Abschnitt 9.4).

Die Phase-5-Build-out-Erzählung der vorherigen Whitepaper-Version ist vollständig: Live-Trading ist die operative Realität, kein Zukunftsziel. Die Frontier ab hier ist *Skalierung*, nicht *Ermöglichung*.

0.4.7 3.7 Reporting und Audit

Vier Reporting-Surfaces:

- **Dashboard.** Web-UI in eine Public-Surface (sanitarisiert, vorzeigbar) und eine Private, authentifizierte Surface (vollständige Trade-Details, Equity-Curve, Capital-Events, Settings) gespalten. Beschrieben in Abschnitt 10.
- **Telegram-Briefings.** Zweimal täglich, morgens und abends, zur Operator-Lokalzeit.

Inhalt: offene Trades, jüngste Signale, Marktkontext, On-Chain-Anomalien.

- **Wochen-Makro-Bericht.** LLM-generierte Synthese der Makro- und On-Chain-Daten der vergangenen Woche, an festem Schedule ausgeliefert.
- **Audit-Speicher.** Jeder Kandidat, jede Analyse, jeder Trade, jede Parameteränderung, jedes Alert, jedes Capital-Event und jeder Internal-Transfer wird mit vollem Zeitstempel-Payload persistiert. Das ist die Ground Truth für Post-hoc-Evaluation und für jegliche zukünftige Compliance-Review.

0.5 4. Der Analyser: LLM als disziplinierter Analyst

0.5.1 4.1 Prompt-Engineering-Prinzipien

Der Unterschied zwischen einem gut und einem schlecht eingesetzten LLM in einer Finanz-Entscheidungs-Aufgabe liegt fast vollständig in der Prompt-Architektur. Drei Prinzipien leiten unser Design:

Prinzip 1 — Strukturierter Kontext, kein Free-Form-Prosa. Der Prompt ist ein deterministisches Template, befüllt mit deterministischen Daten. Es gibt keine natürlichsprachige Anfrage wie "bitte analysiere BTC für mich". Jede Sektions-Überschrift und jeder Datenblock steht in fester Reihenfolge und festem Format. Das ist weniger expressiv, aber weit reproduzierbarer; es macht auch Prompt-Versions-Diffing aussagekräftig.

Prinzip 2 — Expliziter Regime- und Posture-Kontext. Wir liefern Regime-Klassifikation und Strategie-Modus als explizite, gelabelte Inputs — nicht als Inferenzen, die das LLM ableiten muss. Das spart Context-Window und entfernt eine Quelle von Varianz im Reasoning.

Prinzip 3 — Deterministisches Output-Schema. Das LLM ist angewiesen, ein JSON-Objekt mit einem spezifischen Schema zurückzugeben. Moderne LLMs sind unter Schema-Anweisung verlässlich, besonders kombiniert mit einem Beispiel-Payload im System-Prompt. Das nachgelagerte System konsumiert nur die schema-validen Antworten; nicht-konforme Antworten werden einmal retried und dann verworfen.

0.5.2 4.2 Prompt-Struktur (generalisiert)

Folgendes ist ein *generalisiertes* Template. Unser Produktions-Prompt enthält zusätzliches Tuning, das wir hier nicht publizieren. Die Struktur ist jedoch repräsentativ.

```
SYSTEM:
Du bist ein disziplinierter Krypto-Markt-Analyst.
Dein Output muss dem unten spezifizierten JSON-Schema entsprechen.
Füge keinen Kommentar ausserhalb des JSON hinzu.
Respektiere die in der STRATEGY-MODE-Sektion spezifizierte Risk-Posture.
```

```
Du darfst BUY (Long), SELL (Short), HOLD oder AVOID empfehlen.

USER:
ANALYSIS REQUEST

SYMBOL CONTEXT
  Symbol:             [SYMBOL]
  Current price:      [PRICE]
  24-hour change:     [PCT]
  Market-cap rank:    [RANK]

INDICATOR SNAPSHOT (1m / 15m / 1h / 4h / 1d)
  1m:  RSI=..  EMA20=..  EMA50=..  MACD=..  ATR%=..
  15m: (same fields)
  1h:  (same fields, plus EMA200, BB)
  4h:  (same fields)
  1d:  (same fields)

RECENT CANDLES
  1m  (last 30): ...
  15m (last 10): ...
  1h  (last 10): ...
  4h  (last 10): ...

MARKET REGIME
  Regime:            [one of: trend_up_liquid, trend_down_liquid, range,
                              vol_expansion, crash]
  Confidence:          [0..1]
  Key metrics:         EMA20-slope=..  vol-ratio=..  price-vs-EMA200=..
                       RSI-1h-mean=..  ATR-normalised=..

ON-CHAIN CONTEXT
  Asset-specific (if available):
    24h exchange net-flow (native):  ..
    24h exchange net-flow (USD):     ..
    Tracked-wallet delta (24h):      ..
  Aggregate:
    BTC hash rate (7d avg, TH/s):    ..
    Mempool depth (MB):              ..
    ETH gas (gwei):                  ..
    DeFi TVL (USD, 24h delta %):     ..
```

```
STRATEGY MODE
  Posture:              [conservative | moderate | aggressive]
  Min confidence:       [value]
  Min R/R:              [value]
  Direction allowed:    [long_only | short_only | both]
  Notes:                [posture-specific guidance]

OUTPUT SCHEMA
{
  "action":      "BUY|SELL|HOLD|AVOID",
  "confidence":  integer 0..100,
  "entry":       float,
  "stop":        float,
  "tp1":         float,
  "tp2":         float,
  "reasoning":   string (max 500 chars),
  "risk_level":  "low|medium|high",
  "timeframe":   "short|medium|long"
}
```

Drei Eigenschaften verdienen Hervorhebung:

- Die Regime-Klassifikation wird *geliefert*, nicht inferiert. Bewusste Architektur-Entscheidung (Abschnitt 5).
- Der On-Chain-Block ist in asset-spezifische und aggregierte Sub-Sektionen strukturiert. Wo asset-spezifische Daten fehlen (z.B. kleinere Altcoins ausserhalb der gängigen Exchange-Label-Sets), degradiert der Block sauber auf “nur aggregiert” mit Hinweis.
- Der Strategie-Modus-Block enthält *Werte* (Min-Confidence, Min-R/R, erlaubte Richtung), nicht nur ein Label, sodass das LLM darüber reasonen kann, warum die Posture relevant ist, ohne dass der Operator den System-Prompt bei Parameter-Änderungen modifizieren muss.

0.5.3 4.3 Validatoren

Das LLM ist nicht der finale Entscheider. Zwei deterministische Validatoren sitzen stromabwärts:

Risk/Reward-Validator. Das Verhältnis `(tp1 − entry) / (entry − stop)` für Long-Trades und `(entry − tp1) / (stop − entry)` für Short-Trades muss eine modus-abhängige Schwelle überschreiten. Unter Schwelle wird der Trade *nicht stillschweigend verworfen* — die Action wird auf HOLD heruntergestuft und die Confidence um eine fixe

Strafe reduziert. Das erhält den Audit-Record: wir wissen, was das LLM tun wollte und warum der Validator interveniert hat.

Confidence-Gate. Modus-abhängige Mindest-Confidence. Actions unter Threshold werden HOLD; HOLD-Actions propagieren in den Analyse-Speicher, generieren aber keinen Trade und kein Alert.

Diese Validatoren erfüllen drei Funktionen:

1. **Sicherheitsnetz gegen LLM-Überkonfidenz.** Gelegentlich generiert das LLM ein technisch plausibles Trade-Setup mit ungenügendem R/R. Der Validator ist das mechanische Backstop.
2. **Operator-Kontrolle.** Über den Strategie-Modus (Abschnitt 7) kann der Operator die Gesamt-Posture verschärfen oder lockern, ohne den Prompt zu ändern.
3. **A/B-Analyse.** Trades werden mit dem zur Erstellungszeit aktiven Strategie-Modus getaggt — erlaubt Post-hoc-Vergleich modus-bedingter Erwartungswerte.

0.6 5. Marktregime-Detection

0.6.1 5.1 Motivation

Regime-Awareness ist aus unserer Sicht die strukturell konsequenzenreichste Verbesserung gegenüber einem naiven algorithmischen Krypto-System. Akademischer und Praktiker-Konsens konvergieren:

- Kalena Research (2026) berichtet materielle Performance-Differentiale — in der Grössenordnung Dutzender Basispunkte pro Woche im Studienzeitraum — zwischen Regime-Switching und regime-agnostischen Varianten derselben Basis-Strategie.[20]
- Perfumo (Kraken, 2026) argumentiert, dass speziell der 2025–2026er Markt durch ausgedehnte Range-Phasen, unterbrochen durch kurze Volatilitäts-Expansionen, geprägt ist — eine Konfiguration, die Trend-Following-Systeme ohne Regime-Filter bestraft.[21]
- Viprasol (2026) identifiziert Regime-Detection-vor-Position-Sizing als das definierende Merkmal post-2025er professionellen algorithmischen Tradings.[22]

Die methodische Herausforderung: *wahre* Regimes sind latent und unbeobachtbar; jeder Klassifizierer ist eine Approximation. Die Designwahl steht zwischen statistischen Ansätzen (Hidden-Markov-Modelle, Change-Point-Detection) und regelbasierten Decision-Trees. Wir wählen regelbasiert, aus drei Gründen: (i) Transparenz für Audit und Operator-Review, (ii) schnelle Iteration und deterministisches Testing, (iii) die Inputs (EMA-Slope, ATR,

[20] Kalena Research. *"Regime-Aware Crypto Algorithmic Trading: A 2026 Practitioner Review."* März 2026.

[21] Perfumo, Thomas (Kraken Chief Economist). *"Why This Cycle Isn't Like the Others."* Kraken Research, Februar 2026.

[22] Viprasol Research. *"Algorithmic Crypto Trading in 2026: State of the Art."* Jährliche Review, 2026.

Volume-Ratio) sind ohnehin Schätzer latenter Variablen — ein statistisches Modell darauf aufzulegen fügt Parametrisierung ohne grossen Informationsgewinn hinzu.

0.6.2 5.2 Klassifizierer-Inputs

Unser Klassifizierer operiert auf dem 4-Stunden-Candle-Stream und nutzt folgende per Symbol berechneten Inputs:

- **EMA20- und EMA50-Slopes.** Prozentuale Veränderung des EMA über einen 5-Kerzen-Lookback. Konsistent positiv signalisiert Aufwärtstrend; konsistent negativ Abwärtstrend; nahe null Range.
- **ATR normalisiert.** 14-Perioden-ATR als Prozent des aktuellen Preises. Volatilitätsmass unabhängig vom Preis-Niveau.
- **Volume-Ratio.** Volumen der laufenden Kerze geteilt durch den Durchschnitt der vorhergehenden 19 *abgeschlossenen* Kerzen. Das "abgeschlossen" ist wichtig: Wenn die laufende Kerze als Zähler gegen einen Completed-Candle-Nenner verglichen wird, entsteht systematischer Undershoot — einer der häufigeren Bugs in Open-Source-Regime-Klassifikatoren.
- **Preis-vs-EMA200.** Binärer Indikator, ob Preis über oder unter dem Tages-EMA200 liegt.
- **Bollinger-Band-Position.** Relative Position des Preises in der Band-Range.
- **RSI 1-Stunden-rollierender-Mean.** Geglätteter RSI gegen Single-Candle-Rauschen.

0.6.3 5.3 Entscheidungs-Methodik

Der Klassifizierer ist ein kurzer Decision-Tree, in fixer Prioritätsreihenfolge ausgewertet:

Die Reihenfolge zählt: Wir prüfen Crash zuerst (sodass eine Vol-Expansion, die *gleichzeitig* ein Crash ist, als Crash klassifiziert wird), dann Vol-Expansion, dann Range, dann Trend. Die Thresholds sind strategie-relevant und werden hier nicht publiziert; die Methodik zur Setzung: (i) Backtests über 2023–2025-Daten, (ii) Regime-Forward-Validierung auf 2026-Daten, (iii) laufendes statistisches Monitoring der Klassifizierer-Stabilität (Regime-Change-Frequenz als Diagnostik).

Konservatismus-Bias. Unsere Thresholds sind so gesetzt, dass — bei fehlendem klaren Beweis — die Klassifikation auf *Range* defaultet. Ranges sind das modale Regime in 2025–2026 (Perfumo 2026), und range-passende Strategien (Mean-Reversion, Fade-zur-Band) degradieren in schwachen Trends gracefully. Trend-passende Strategien (Breakout-Continuation) degradieren in Ranges schlecht. Dieser Konservatismus-Bias ist eine bewusste asymmetrische-Verlust-Wahl.

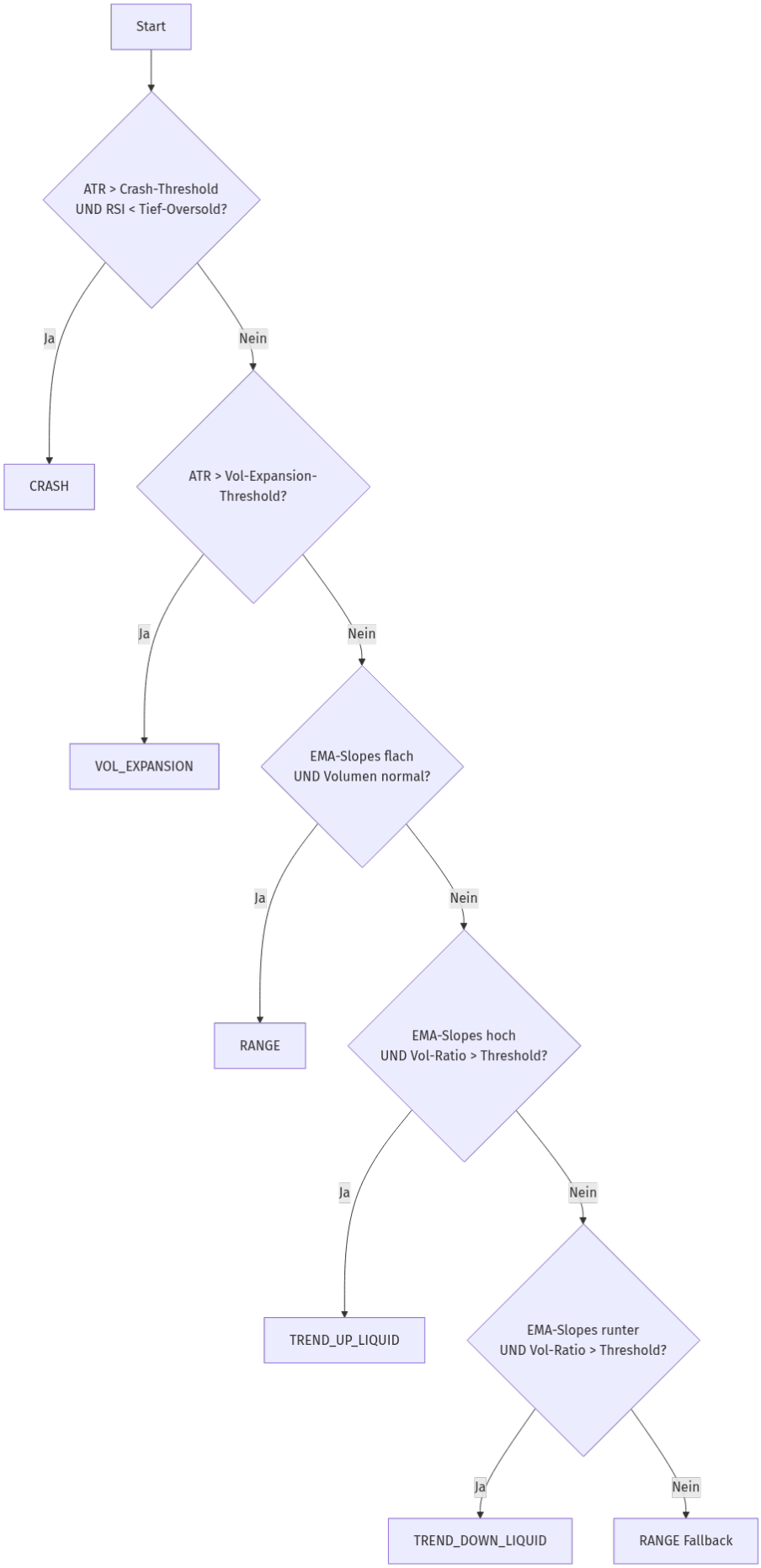


Abbildung 1: Diagram

0.6.4 5.4 Wie Regime den Analyser beeinflusst

Die Regime-Klassifikation tritt als gelabelter Input in den LLM-Prompt ein (Abschnitt 4.2) und ändert, wie das LLM über den Kandidaten reasonen soll. Der Prompt *instruiert* das LLM nicht zu einer bestimmten Strategie; er liefert Regime als Kontext und vertraut darauf, dass ein gut-aligntes Modell regime-passende Outputs produziert. In Produktion beobachten wir das: das Reasoning-Text des LLM referenziert das Regime explizit ("Range-Regime begünstigt Fade Richtung unteres Band", "Trend-up-liquid unterstützt Momentum-Continuation"), bei BUY-, SELL- und AVOID-Calls.

0.6.5 5.5 Regime-Change-Exit

Die Regime-Klassifikation wird auch in der **Execution**-Schicht konsumiert, in Form eines Regime-Change-Exit-Branch. Die Logik:

```
Für jeden offenen Trade:
    direction := bull | bear | sideways  (abgeleitet vom Regime-Label)
    trade_side := long | short

    falls direction unbekannt oder ambiguous:
        skip (keine Action)

    sonst falls (trade_side == long  UND direction == bear)
       ODER (trade_side == short UND direction == bull):
        # Das Regime steht jetzt gegen den Trade.

        hypothetical_pnl_pct := pnl_at_current_price / notional
        round_trip_fee_pct   := 2 × taker_fee_pct  (Entry + Exit)

        falls hypothetical_pnl_pct > round_trip_fee_pct:
            Trade zum aktuellen Preis schliessen
            exit_reason = 'regime_change'
        sonst:
            # Trade ist im Profit, aber der Regime-Wechsel
            # würde Round-Trip-Fees nicht decken. Ursprüngliche
            # Stop-/TP-Levels weiter gelten lassen.
            keine Action.
```

Die Fee/Funding-Berechnung erfolgt auf **Notional** (`qty × Preis`), das ist hebel-agnostisch. Der Regime-Exit-Branch feuert nie auf einem verlierenden Trade — er bewirkt nie, dass eine ausgestoppte Position früher geschlossen wird als an ihrem Stop-Level. Er triggert nur, wenn eine profitable Position einem frisch feindlichen Regime gegenübersteht, in dem die erwartete Fortsetzung des ursprünglichen Profits jetzt negative-EV gegen einen Trade in

der regime-bevorzugten Richtung ist.

Dieser Branch fehlt unseres Wissens in publizierten akademischen Implementationen LLM-augmentierten Krypto-Tradings. Es ist ein kleines Stück Code mit deutlicher Wirkung in Regime-Übergangsphasen, in denen in einem Regime etablierte Positionen einer gewaltsam veränderten Erwartungs-Distribution ausgesetzt sind.

0.6.6 5.6 Diagnostik: Regime-Verteilung

Eine der nützlichsten Diagnose-Outputs des Regime-Klassifizierers ist die empirische Verteilung der Regimes über die Watchlist. In einem ausgeglichenen Markt erwarten wir die meisten Coins meiste Zeit in Range oder mildem Trend; Regime-Homogenität über die Watchlist (z.B. alle Coins simultan in Crash) ist ein Cross-Validation-Signal markt-weiten Stresses. Regime-Persistenz-Statistiken (wie lange dauert eine Klassifikation im Schnitt?) sind eine Diagnostik der Klassifizierer-Stabilität — ein Klassifizierer, der jedes Cycle das Regime flippt, ist nutzlos, unabhängig von In-Sample-Genauigkeit.

0.7 6. On-Chain-Integration

0.7.1 6.1 Datenquellen-Landschaft

Krypto-On-Chain-Daten haben eine gestaffelte Kostenstruktur. Die für eine institutionelle Implementation relevante Landschaft:

Free-Tier (Public APIs).

- **Blockchain.com Charts-Endpoints** — netzwerk-level BTC-Metriken, Hashrate, aktive Adressen, Transaktionen.
- **Mempool.space API** — Echtzeit-BTC-Mempool-Tiefe, Fee-Märkte, jüngste Blöcke.
- **Bitnodes API** — erreichbare Node-Counts, Netzwerk-Topologie-Grobsignale.
- **DefiLlama API** — DeFi-TVL, Chain-Breakdowns, Stablecoin-Supply pro Chain.
- **Coin Metrics Community API** — keylos, rate-limited. Liefert Exchange-Inflow/Outflow für BTC und ETH und ein Subset Netzwerk-Metriken.
- **Etherscan Public Endpoints (V2)** — Free-Tier nach Key-Registrierung, 5 Requests/Sekunde, 100 k/Tag. Ethereum-spezifisch: Gas, Supply, Balance-Lookups gegen gelabelte Adressen.

Mid-Tier (kostenpflichtig, monatlich).

- **Glassnode** — Advanced (~USD 49/Monat) für die meisten On-Chain-Indikatoren; Pro (~USD 999/Monat) für proprietäre Metriken (True Market Mean, Short-Term-Holder-Cost-Basis, Coin-Days-Destroyed, realisierte Gewinn-Dekomposition).
- **Nansen** — Smart-Money-Wallet-Tracking, gelabelte-Entitäts-Analyse.

Top-Tier (kostenpflichtig, höher monatlich).

- **Arkham Intelligence** — umfassendste Label-Datenbank; löst das Coinbase/Kraken-Smart-Contract-Custody-Tracking-Problem, das mit Etherscan-Labels allein unlösbar ist.
- **Chainalysis / TRM Labs** — primär compliance-orientiert; weniger direkt nützlich für Signal-Generierung.

Trade-offs. Free-Tier reicht für grobe Richtungs-Signale und für Bitcoin-spezifische Analyse. Die Hauptlücke ist akkurates Exchange-Custody-Tracking auf Ethereum, wo moderne Börsen zunehmend Smart-Contract-Custody-Architekturen verwenden, die nicht traditionelle Etherscan-Labels tragen. Unser aktuelles Free-Tier-Tracking erfasst rund 2.66 Millionen ETH über 13 gelabelte Wallets; die echte Custody-Höhe (per Glassnode-/Arkham-Reconciliation) ist materiell höher. *Relative* Bewegungen bleiben aussagekräftig; *absolute* Niveaus sollten bis zur Paid-Tier-Anbindung nicht als autoritativ behandelt werden. Migration zur Arkham-Label-Datenbank steht auf dem Post-MVP-Enhancement-Track (Abschnitt 13).

0.7.2 6.2 Exchange-Flow als Frühindikator

Der Interpretations-Rahmen — konsistent über Glassnodes Forschung und unsere eigenen Beobachtungen — lautet:

- Anhaltender **Net-Outflow** von Börsen (Coins von custodial Wallets in Self-Custody) zeigt **Akkumulation** an. Halter reduzieren die Distanz zu einem Verkaufs-Venue. Bullish auf 1–4 Wochen.
- Anhaltender **Net-Inflow** zu Börsen zeigt **Distribution** an. Halter positionieren für Verkauf. Bearish auf demselben Horizont.
- **Kurz-horizontige Spikes** (Single-Day-Outliers) sind verrauscht und sollten isoliert nicht als Signal behandelt werden. Das brauchbare Fenster ist 7-Tage-rollierender Net-Flow.

Wir exponieren das dem Analyser als 24-Stunden-Net-Flow, 7-Tage-rollierender Net-Flow und Perzentil des aktuellen Werts gegen eine Trailing-Distribution. Das LLM gewichtet das gegen Preis-Action und Regime.

0.7.3 6.3 Whale-Wallet-Tracking

Methodisch:

1. **Wallet-Identifikation.** Für jede grosse zentrale Börse identifizieren wir ein Set gelabelter On-Chain-Adressen. Free-Tier: Etherscan-Labels plus öffentliche Disclosures. Paid-Tier: Arkham-/Nansen-Cluster-Labels, die den Smart-Contract-Custody-Gap lösen.
2. **Aggregation.** Adressen werden nach Exchange-Cluster gruppiert. Balance und Net-Flow werden pro Cluster summiert.

3. **Delta-Tracking.** Per-Cluster-24h- und 7d-Balance-Deltas sind das Signal. Cluster-interne Transfers sind keine Net-Flows und müssen ausgeschlossen werden — im Free-Tier nicht-trivial, ein Schlüssel-Vorteil von Paid-Tier-Label-Datenbanken.
4. **Threshold-basierte Alerts.** Cluster-Bewegungen über Trailing-Distribution-Perzentil triggern Alerts. Den genauen Perzentil-Threshold publizieren wir nicht.

0.7.4 6.4 Network-Health

Die langsam-bewegten Strukturindikatoren:

- **Bitcoin-Hashrate.** Trended-Hashrate ist ein Langfrist-Confidence-Proxy (Miner committen Kapital). Plötzliche Hashrate-Drops korrelieren mit Capitulation-Phasen.
- **Mempool-Tiefe und Fees.** Hoher Mempool mit erhöhten Fees signalisiert Blockspace-Nachfrage — typisch begleitend zu Marktstress oder erhöhter spekulativer Aktivität.
- **Aktive Adressen.** Wöchentlicher Durchschnitt aktiver Adressen, getrendet, ist ein Adoptions-Proxy.

Keiner davon ist ein Entry-Signal. Alle sind Kontext-Signale, die das Framing des Analysers beeinflussen — im LLM-Prompt erscheinen sie in der "aggregate"-Sub-Sektion des On-Chain-Blocks.

0.8 7. Strategie-Modi und Position-Sizing

0.8.1 7.1 Drei-Posture-Switch

KonnectAI Trader exponiert ein einziges operator-kontrolliertes Setting — den **Strategie-Modus** — mit drei diskreten Werten: **conservative**, **moderate**, **aggressive**. Das Setting ist global; es gilt für alle Coins der Watchlist.

Der Modus kontrolliert mindestens:

- Mindest-Confidence-Threshold für Live- oder Paper-Trades.
- Mindest-R/R-Ratio, durchgesetzt vom Validator.
- Mindest-Trigger-Count, durchgesetzt vom Pre-Filter.
- RSI-Oversold-/Overbought-Thresholds des Pre-Filters.
- Alert-Confidence-Threshold für Telegram-Notifications.
- Richtungs-Envelope (long-only, short-only, both) — ein Strategie-Layer-Overlay über das Per-Trade-Signal.

Wir publizieren die spezifischen numerischen Werte pro Modus nicht. Die *Methodik* zur Setzung: (i) Erbe akademische Defaults oder Practitioner-Heuristiken (z.B. RSI 30/70) wo sinnvoll, (ii) modus-anpassen, sodass eine messbare Posture-Differenz entsteht, (iii) Validierung via In-MVP-A/B-Vergleich, dass jeder Modus ein distinktes Risk/Return-Profil hat, (iv) Live-Trade-Records mit dem aktiven Modus taggen, sodass laufende Performance

A/B-vergleichbar bleibt.

0.8.2 7.2 Warum global, nicht pro Symbol

Ein per-Symbol-Strategie-Setting ist operativ attraktiv (verschiedene Coins wollen verschiedene Postures), erzeugt aber rasche Parameter-Explosion: 14 Coins × 3 Modi × 5 Parameter = 210 operative Stellschrauben. Wir wählten bewusst den globalen Switch als einfacheres System, mit der expliziten Zukunfts-Absicht, regime-bedingte Per-Symbol-Anpassungen einzuführen, sobald genug Daten für per-Symbol-statistische Inferenz vorliegen.

0.8.3 7.3 Position-Sizing

Drei Position-Sizing-Optionen sind implementiert und operator-wählbar:

- **Fester Anteil der Equity.** Z.B. 1.5 % der verfügbaren Equity pro Trade. Einfach, robust, verlust-insensitiv.
- **ATR-normalisiertes Sizing.** Position-Grösse so gewählt, dass die Distanz zum Stop in absoluten Termini einem festen Prozent der Equity entspricht. Robuster gegen Volatilitäts-Regime.
- **Volatility-Budget-Sizing.** Portfolio-Heat ist gedeckelt: simultanes offenes Risiko über alle Positionen ist auf einen Portfolio-Prozent-Target begrenzt.

Zwei Verfeinerungen in der Produktions-Sizing-Logik sind erwähnenswert:

- **TP1-aware Risk-Berechnung.** Risiko pro Trade wird gegen den *originalen* Stop berechnet (Entry → Original-SL-Distanz), nicht gegen den nach TP1 trailed Stop. Nach TP1 ist das Runner-Risiko effektiv klein oder negativ geworden, aber das *initiale* Sizing des Trades basierte korrekt auf der vollen Original-SL-Distanz, nicht auf der günstigeren Trailed-Distanz.
- **Capital-Event-aware Basis.** Die "verfügbare Equity" im Sizing ist die jüngst rekonzilierte Equity aus dem Wallet-Tracking-Pipeline (Abschnitt 9.4), nicht ein veralteter Snapshot. Ein- und Auszahlungen propagieren ohne manuellen Eingriff in den nächsten Sizing-Cycle.

Der professionelle Praxis-Konsens (Kalena 2026, Viprasol 2026) lautet, dass eine Form volatilitäts-normalisierten Sizings festes-Prozent-Sizing materiell schlägt, besonders über Regime-Übergänge.

0.8.4 7.4 A/B-Analyse und Tagging

Jeder Trade wird mit einem `strategy_mode`-Tag und einem `regime_at_entry`-Tag persistiert. Nicht kosmetisch: ermöglicht rigorose Post-hoc-Analyse modus-bedingter Erwartungswerte. Nach 30+ Trades pro Modus-und-Regime-Zelle berechnen wir:

- Modus-spezifische Win-Rate.
- Modus-spezifischer Erwartungswert (durchschnittlicher Profit pro Trade, positiv oder negativ).
- Modus-spezifischer Profit-Faktor (Brutto-Profit / Brutto-Verlust).
- Modus-spezifische Drawdown-Reihe.

Unter dieser Stichprobengrösse ist keine Inferenz statistisch sinnvoll. Das stellen wir öffentlich klar.

0.9 8. Paper-Trading-Methodik

0.9.1 8.1 Die Rolle des Paper-Tradings: MVP-only Sandbox

In Produktion ist alles Trading live. Es gibt keinen Paper-Trading-Layer in der Produktions-Deployment für die vier Investor-Accounts. Jedes Signal wird gegen Binance Spot oder Binance Futures executiert, auf dem eigenen Account jedes Investors via dessen gescoptem API-Key, mit echten Fees, echter Slippage, echtem Funding.

Paper-Trading existiert im **MVP** als R&D-Bench: eine sandbox-isolierte Strategie-Validierungs-Umgebung, in der Prompt-Revisionen, Threshold-Anpassungen, Regime-Klassifizierer-Tuning und neue Signal-Quellen end-to-end exerciert werden, ohne Kapital zu riskieren. Der MVP-Paper-Trader ist die Staging-Surface; der Produktions-Trading-Pool ist die Live-Surface; Übernahmen aus einem in den anderen werden auf der Bench reviewed, bevor sie in Produktion appliziert werden.

Diese Positionierung ist eine bewusste Änderung gegenüber früheren Framings. Paper-Trading ist nicht die Validierung eines zukünftigen Live-Systems; das Live-System existiert bereits und produziert die einzigen Performance-Daten, über die wir berichten. Paper-Trading ist die Entwicklungs-Bench, die kontinuierliche Verbesserung des Live-Systems unterstützt.

0.9.2 8.2 Partial-Take-Profit mit Runner

Das Execution-Pattern, angewandt auf Live- wie Paper-Trades:

```
Trade öffnet bei `entry`, mit Stop bei `original_SL`, tp1 bei `tp1`, tp2 bei `tp2`.

Phase 1:  remaining_pct = 100
          stop_active = original_SL
       - Falls Preis stop_active erreicht: alles bei Stop schliessen.  Reason: stop.
          - Falls Preis tp1 erreicht:          50 % bei tp1 schliessen.
                                               remaining_pct := 50
                                       stop_active := tp1   (Stop auf TP1
```

```
                                                    anheben, TP1-Gewinn
                                                    auf Runner sichern)
                                                original_SL bleibt unverändert
                                                     weiter (nächster Cycle).

Phase 2:  remaining_pct = 50
      - Falls Preis stop_active erreicht:  alles bei stop_active schliessen.  Reason
      - Falls Preis tp2 erreicht:       alles bei tp2 schliessen.       Reason: tp2.
```

Original-SL-Preservation. Eine subtile, aber wichtige Eigenschaft: der *originale* Stop-Level bleibt als Spalte am Trade-Record erhalten, auch nachdem TP1 den aktiven Stop nachgezogen hat. Zwei Gründe: (i) Risk-Attribution zur Sizing-Zeit basierte auf dem originalen Stop, (ii) Post-hoc-Analyse von "wie weit ist der Trade gegen die ursprüngliche Risk-Prämisse maximal in den Drawdown gelaufen?" funktioniert nur, wenn der originale Stop weiterhin abrufbar ist.

Begründung Partial-TP-Runner. Ein Voll-Exit bei TP1 sichert moderate Gewinne und gibt Upside in Trend-Regimes auf. Ein No-Scaling-Ansatz mit einzelnem TP2-Target triggert zu selten: in unseren beobachteten Live-Daten wird TP2 nur in einer kleinen Minderheit der Fälle ohne vorhergehendes TP1 erreicht, und der Trade stoppt häufig auf dem Weg zu TP2 aus, nachdem TP1 bereits durchquert wurde. Das Partial-Exit-mit-Runner-Pattern erhält den Hauptanteil des Upside, sichert eine Teil-Profit ein und entfernt das Downside-Risiko auf dem Runner-Bein.

Timeout. Trades, die in einem festen Horizont weder Stop noch Take-Profit erreichen, werden zwangsweise zum dann-aktuellen Preis geschlossen. Verhindert primär unbestimmtes Driften der Trades.

Manual-Close. Operator kann jeden offenen Trade vom Dashboard schliessen. Close wird mit Operator-Identität und Closing-Preis aufgezeichnet. First-Class-Action, keine Hintertür.

Regime-Change-Close. Unabhängig von Stop, Take-Profit oder Timeout kann ein Trade unter dem Regime-Change-Exit-Branch (Abschnitt 5.5) geschlossen werden.

0.9.3 8.3 Slippage in Paper vs Live

Paper-Trading simuliert keine Slippage. Bewusste Scope-Limitation, explizit geflaggt: Paper-Trading-PnL-Zahlen sind Best-Case-Schranken, keine Live-Erwartungs-Forecasts. Das Live-System dagegen *misst* Slippage Trade für Trade — jeder Exchange-Fill wird gegen den Signal-Zeit-Preis abgeglichen, und die realisierte Slippage-Kosten geht in die Per-Trade-PnL und die wallet-basierte Netto-Rendite ein.

Das ist ein struktureller Grund, in jeglicher investor-bezogener Performance-Diskussion Live-Zahlen über Paper-Zahlen zu stellen: die Live-Zahlen enthalten bereits den grössten

in der Praxis gemessenen Drag (NBER 2024[23] berichtet 34–67 % der Backtest-zu-Live-Performance-Lücke als Slippage), Paper-Zahlen tun das nicht.

0.9.4 8.4 Sample-Size-Ehrlichkeit

Statistisch sinnvolle Inferenz braucht Stichprobengrössen. Unsere Disziplin:

- Unter 30 Trades pro (Modus, Regime)-Zelle: keine Inferenz.
- 30–100 Trades: grobe Erwartungswert-Schätzung, breite Konfidenzintervalle.
- 100+ Trades pro Zelle: enger werdende Konfidenzintervalle, sinnvolles A/B.
- 300+ Trades pro Zelle, über Regimes verteilt: defensible Erwartungswerte.

Investoren sollen erwarten, dass Performance-Reporting diese Disziplin reflektiert. Wir extrapolieren nicht aus kleinen Stichproben.

0.10 9. KPIs und Evaluations-Methodik

0.10.1 9.1 Per-Trade- und aggregierte KPIs

Die KPIs, die wir tracken, sind in der systematischen Trading-Literatur Standard. Ihre Standardisierung ist ein Feature — Investor-Due-Diligence-Berater können unsere Werte gegen vergleichbare Systeme benchmarken, ohne Metriken neu zu engineeren.

- **Win-Rate** — Anteil geschlossener Trades mit positivem PnL. Neben Average-Win und Average-Loss zu lesen.
- **Erwartungswert** — `(win_rate × avg_win) − ((1 − win_rate) × avg_loss)`. Single-Number-Zusammenfassung erwarteten PnLs pro Trade.
- **Profit-Faktor** — Brutto-Gewinn / Brutto-Verlust. Werte über 1.25 sind professionell akzeptabel; über 2.0 stark; unter 1.0 verlustträchtig.
- **Sharpe / Sortino-Ratio** — Return-zu-Volatilität und Return-zu-Downside-Volatilität, annualisiert. Konzeptuell für ein System mit regulärer Trade-Kadenz wohldefiniert; in Krypto vorsichtige Interpretation wegen Fat-Tails.
- **Maximaler Drawdown** — Peak-to-Trough-Rückgang auf der Equity-Curve.
- **Per-Regime-Performance.** Erwartungswert bedingt auf Regime-Klassifikation bei Trade-Entry. Regime-bedingter Erwartungswert ist die Diagnostik, die uns sagt, ob der Regime-Klassifizierer Mehrwert liefert.
- **Per-Strategie-Modus-Performance.** Modus-getaggte A/B-Analyse wie in Abschnitt 7.4.
- **Per-Direction-Performance.** Long-vs-Short-Erwartungswert, regime-bedingt, ist eine strukturelle Diagnostik der bidirektionalen Architektur.

[23] National Bureau of Economic Research. Working Paper 31890. *"Slippage and Performance in Algorithmic Trading."* 2024.

0.10.2 9.2 Cost-KPIs

Die operativen Kosten des Systems sind selbst ein First-Class-KPI:

- **Cost pro Analyse** (USD). Total LLM-Inferenz-Spend / durchgeführte Analysen.
- **Cost pro Winning-Trade** (USD). Total LLM-Inferenz-Spend / Anzahl gewinnender geschlossener Trades. Holistische Metrik — ein System, das viele Analysen, aber wenige Winning-Trades macht, ist teuer, ungeachtet seiner Genauigkeit.
- **Total Daily Cost** (USD). Operatives Budget. Auf dem Produktions-1-Minuten-Cron, mit der Vier-Investor-Pooled-Execution-Architektur, skaliert das mit Cycle-Frequenz, nicht mit Investor-Count.

0.10.3 9.3 Wallet-basierte Netto-Rendite

Die Headline-Performance-KPI ist **wallet-basierter Netto-PnL %**. Berechnung:

1. **Capital-Events** (Ein-, Auszahlungen) werden aus Binance gezogen und mit Occurred-at-Zeitstempel und USD-Equivalent persistiert.
2. **Interne Spot↔Futures-Transfers** werden gezogen und in einer separaten Audit-only-Tabelle persistiert — sie zählen *nicht* als Capital-Events. (Sie als Ein-/Auszahlungen zu behandeln, würde die Equity-Curve korrumpieren.)
3. Die Equity-Curve wird durch Capital-Event-Grenzen in **Perioden** partitioniert: eine Periode ist das Intervall zwischen zwei aufeinanderfolgenden Ein-/Auszahlungen.
4. Innerhalb jeder Periode wird die Per-Period-Rendite `r_i` als `(equity_end − equity_start − net_capital_event) / equity_start` berechnet.
5. Die Perioden-Renditen werden **geometrisch verkettet**: `total_return = (1 + r_1) × (1 + r_2) × ... × (1 + r_n) − 1`.

Diese Methodik produziert eine Renditezahl, die **Capital-Events übersteht**: das Einzahlen von CHF 10’000 mitten in einer Periode erscheint nicht als 10 % “Gewinn” auf der Equity-Curve, und das Auszahlen von CHF 10’000 erscheint nicht als 10 % “Verlust”. Die berichtete wallet-basierte Rendite ist die tatsächliche Rendite auf eingesetztes Kapital.

Bridge-Fix für interne Transfers. Interne Spot↔Futures-Transfers laufen asynchron über zwei Wallets. Naive Aggregat-Equity-Berechnung produziert Phantom-Dips (das Transfer-Volumen ist kurz aus beiden Wallet-Snapshot-Fenstern abwesend), die Intra-Day-Equity-Curves kontaminieren. Die Bridge-Logik rekonziliert diese Transfers als Zero-Net-Events auf der Aggregat-Curve und eliminiert die Phantom-Dips. Für die Produktions-Deployment nach Beobachtung im MVP gezielt entwickelt.

0.10.4 9.4 Audit-Trail

Capital-Events, Internal-Transfers, Live-Trades, Futures-Trades, Paper-Trades (MVP only), Analysen, Kandidaten und Regime-Klassifikationen werden in dedizierten Tabellen persistiert. Jede Zeile hat:

- Zeitstempel (Erstellung und, wo zutreffend, Update).
- Volles Payload mit Inputs, Outputs und Reasoning.
- Eine `external_id` (wo es aus einem Binance-Event stammt) markiert UNIQUE für idempotente Re-Sync.
- Zuschreibung zum Operator (manuelle Aktionen) oder Service-Identität (automatisierte Aktionen).
- Versions-Tag des aktiven Pipeline-Stands.

Audit-Speicher ist konventionsgemäss append-only. Schema-Evolution erfolgt über additive Migrationen; destruktive Änderungen erfordern dokumentiertes ADR.

0.10.5 9.5 Performance-Projektionen — explizite Methodik

Jede numerische Performance-Projektion in einem investor-bezogenen Dokument ist eine Behauptung, die defendierbar sein muss. Wir trennen zwei Kategorien:

- **Live-Beobachtungen aus dem MVP seit Januar 2026** — deskriptiv, intern als Ground Truth genutzt, die die Projektions-Methodik motiviert.
- **Forward-Projektionen** — explizit als Projektionen geframed, sowohl in MVP-Daten als auch in einem regime-bedingten Modell verankert.

Wir projizieren Performance in drei Regime-Szenarien:

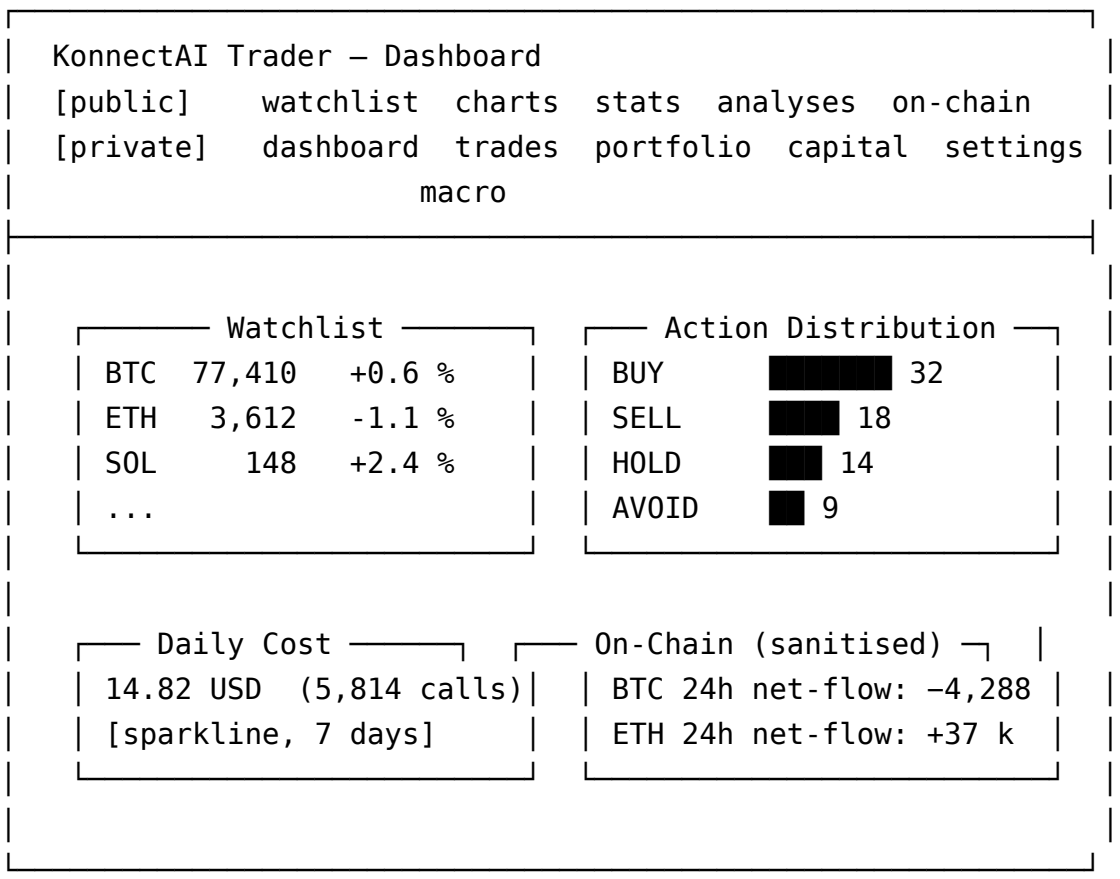
Marktregime	Daily-ROI-Projektion	Annualisiert (linear, 365 Tage)
Drawdown	0.35 %	+127 % p.a.
Sideways	0.50 %	+180 % p.a.
Bullish	0.65 %	+237 % p.a.

Das sind **Projektionen auf MVP-Daten-Basis**, keine historischen Track-Records. Der Lift gegenüber dem akademischen 15–30 % p.a. Korridor wird durch fünf operative Faktoren erklärt: 1’440-Cycle-Cron-Granularität in Produktion, bidirektionale Long-und-Short-Execution, regime-aware Execution inklusive Regime-Change-Exit, adaptives Prompt- und Threshold-Tuning, kapitalereignis-gewahre Buchhaltung (Abschnitt 1.3).

Die investor-bezogenen Capital-Deployment-Szenarien in Abschnitt 14 verwenden diese drei Regime-Szenarien als Evaluations-Grid. Investoren sollten besonders das **Drawdown-Szenario** als Worst-Case-Argument betrachten: unter der pessimistischsten Annahme erreicht selbst das einfachste Szenario in der Tabelle innerhalb von 24 Monaten Breakeven.

0.11 10. Dashboard und Operator-UX

Die Operator-Experience ist nach Authentifizierungs-Grenze gespalten. Ein screenshot-freies Wireframe des aktuellen Designs:



0.11.1 10.1 Public-Surface

Die Public-Surface ist die demonstrierbare, no-Authentifizierungs-Sicht. Zweck: zeigen, dass das System lebt und funktioniert, ohne Trade-Level-Daten freizugeben, die investor-relevante Informations-Asymmetrie darstellen würden.

- **Watchlist** — aktuelle Preise, 24h-%-Änderung, Marktkapitalisierungs-Rang.
- **Charts** — Per-Symbol-Candle-Chart mit Indikator-Overlays (TradingView lightweight-charts).
- **Stats** — Analysis-Action-Verteilung (BUY/SELL/HOLD/AVOID-Counts über rollierende Fenster), aggregierte Throughput-Werte.
- **Recent Analyses (sanitarisiert)** — jede jüngste Analyse mit Action und redigiertem Reasoning-Auszug. Konkrete Entry-/Stop-/TP-Levels nicht öffentlich.
- **Daily-Cost-Telemetrie** — rollierender Spend-Tracker als Transparenz- und Vertrauenssignal.
- **On-Chain-Summary** — aggregierte Network-Health, Exchange-Flows in 24h-Granularität, DeFi-TVL.

0.11.2 10.2 Private / authentifizierte Surface

Hinter der Authentifizierung:

- **Volle Analysen** — jede Analyse mit Entry, Stop, Take-Profit, Reasoning-Text, Risk-Level, Timeframe, Cost.
- **Offene und geschlossene Trades** — volle Live- und Futures-Trade-Records mit Entry, Exit, PnL, Exit-Reason (stop / tp1 / tp2 / regime_change / timeout / manual), Modus-Tag, Regime-at-Entry-Tag, Original-SL-Spalte, optionale Notes.
- **Equity-Curves** — wallet-basierter Netto-PnL mit Capital-Event-Markern, Asset-Breakdown-Stacked-Area-Chart, bridge-fixed Spot-/Futures-Aggregat-Curve.
- **Portfolio-Snapshots** — periodische Snapshots der Holdings auf dem Exchange-Account des Operators, mit Dust-Threshold-Filter.
- **Capital-Events** — chronologische Sicht auf Ein-, Auszahlungen und Internal-Transfers mit ihren rekonzilierten USD-Equivalent-Werten.
- **Settings** — Strategie-Modus-Switch, Watchlist-Editor, Threshold-Tuning (Superuser-only).
- **Macro-Reports-Archiv** — die wochenweise LLM-generierte Makro-Synthese, nach Datum browsbar.

0.11.3 10.3 Authentifizierung

Authentifizierung erfolgt über **TOTP** (Time-based One-Time Password, RFC 6238). Wir nutzen Passwörter nicht als Primär-Faktor.

- Passwörter sind der dominante Breach-Vektor in Small-Team-Web-Anwendungen. Sie zu eliminieren entfernt den grössten Teil der Angriffsfläche.
- TOTP-Secrets werden server-seitig gespeichert, gehasht wo möglich, nach Initial-Provisionierung nie übertragen.
- Replay-Schutz über das 30-Sekunden-TOTP-Fenster; ein kurzlebiges Session-Token (JWT) wird nach erfolgreichem TOTP ausgestellt und über separaten Flow refreshed.
- Rate-Limits am Login-Endpoint verhindern TOTP-Brute-Force; standardmässige Account-Lockout-Logik gilt.

Dies ist passend für ein kleines autorisiertes Operator-Set (Founder plus bis zu vier Investoren plus deren autorisierte Principals). Per-Investor-Account-Sichten werden über die User-zu-Investor-Account-Bindung im Auth-Speicher gegated, die auf den API-Key-Vault-Eintrag referenziert statt direkt Credentials zu halten.

0.12 11. Operations und Observability

0.12.1 11.1 Scheduling

Die Pipeline ist auf mehreren Kadenzen geschedulet. Die **Core-Loop**-Kadenz ist Konfiguration:

- **MVP-Cron: 15-Minuten-Core-Loop** (96 Cycles/Tag). Die Kadenz, unter der das

System seit Januar 2026 live läuft.
- **Production-Cron: 1-Minuten-Core-Loop** (1'440 Cycles/Tag). Die Ziel-Kadenz für die Vier-Investor-Produktions-Deployment, mit feinerer Candle-Daten und Throughput-Vorteil gegenüber akademischen Baselines.

Andere Kadenzen:

- **Stündlicher On-Chain-Snapshot.** Niedrigere Kadenz akzeptabel, weil On-Chain-Metriken sich langsam bewegen.
- **Stündlicher Macro-Snapshot.** Gleiche Begründung.
- **Stündlicher Portfolio-Snapshot.** Read-only Binance-API-Call.
- **Capital-Events-Sync.** Auf jedem Cycle wird der Ein-/Auszahlungs-/Internal-Transfer-Log mit Binance abgeglichen.
- **Zwei-Mal-Tägliche Briefings.** Morgens und abends Telegram-Briefings zur Operator-Lokalzeit.
- **Wochen-Macro-Report.** LLM-generierte Synthese auf festem Schedule.

Die MVP-Kadenz wurde als Balance zwischen Daten-Aktualität und Kosten in der Build-out- und Validierungs-Phase gewählt. Die Produktions-Kadenz ist für das Hochfrequenz-Operating-Envelope der Vier-Investor-Deployment ausgelegt, in dem die LLM-Inferenz-Kosten über die vier Investor-Accounts amortisiert werden (Abschnitt 11.4).

0.12.2 11.2 Monitoring

Jeder geschedule Job wird auf Completion und Runtime überwacht. Verpasste Executions oder übermässige Runtimes triggern Operator-Alerts. Health-Schlüsselsignale:

- Jüngster Ingest-Zeitstempel pro Datenquelle (Staleness-Detection).
- Analyser-Call-Rate und Cost-Envelope (Budget-Alerting auf Tages- und Monats-Schwellen).
- Open-Trade-Count pro Investor-Account (Sanity-Check gegen Erwartungen).
- Regime-Verteilung über die Watchlist (Cross-Validation gegen Marktkontext).
- Pre-Filter-Emissions-Rate (geht sie auf null, ist ein stillschweigender Upstream-Failure wahrscheinlich).
- Capital-Events-Reconciliation-Drift (wenn unsere berichtete Equity über Schwelle vom Binance-Wallet-Snapshot abweicht: Alert).
- Internal-Transfer-Bridge-Integrität (keine Phantom-Dips auf der Aggregat-Equity-Curve).

Logs sind strukturiert (Zeitstempel, Level, Modul, Event, Correlation-ID) und an einen für die Skalierung passenden Log-Aggregations-Tier verschifft. Für die aktuelle Deployment ist das journald-plus-Datei-Modell ausreichend; auf Multi-Cluster-Skalierung ist ein Managed-Log-Store (Datadog, Grafana Cloud oder gleichwertig) das natürliche Upgrade.

0.12.3 11.3 Audit

Jedes Entscheidungs-Artefakt — Kandidat, Analyse, Trade, Alert, Setting-Änderung, Capital-Event, Internal-Transfer — wird als Zeile in einem strukturierten Speicher persistiert mit Zeitstempel, vollem Payload, Zuschreibung und Pipeline-Version. Audit-Speicher ist konventionsgemäss append-only und policy-gemäss append-only-mit-additiven-Migrationen. Ein Regulator oder investor-bestellter Auditor kann das Verhalten des Systems zu jedem Zeitpunkt rekonstruieren.

0.12.4 11.4 Gepoolte Multi-Investor-Architektur

Das Vier-Investor-Operating-Modell wird von einer Architektur unterstützt, die **operative Logik poolt und Kapital in investor-eigenen, non-custodial Accounts hält**. Wir verwenden bewusst **keine** Binance-Sub-Accounts: dieser Mechanismus würde verlangen, dass KonnectAI ein Binance-VIP-Master-Account betreibt, und würde Investor-Funds unter unsere Kontrolle bringen — genau die Custody-Posture, die wir ablehnen. Stattdessen:

- **Investor-eigene Binance-Accounts.** Jeder Investor eröffnet (oder verwendet ein bestehendes) Standard-Retail-Binance-Account auf seinen Namen, durchläuft Binance' KYC/AML direkt mit der Exchange und zahlt CHF/USD via Schweizer Fiat-On-Ramp ein (z.B. SEBA, Sygnum, Bitstamp, Kraken Pro oder einen anderen regulierten Weg seiner Wahl), die er innerhalb seines Accounts in USDT konvertiert.
- **Gescopte API-Keys, Withdrawal permanent deaktiviert.** Der Investor stellt KonnectAI einen Binance-API-Key mit präzise gescoptem Permission-Set zur Verfügung: Spot-&-Margin-Trading, USDS-M-Futures-Trading und `Universal Transfer` / Binance Pay (erforderlich für den monatlichen Operating-Cost-Auto-Deduct). **Withdrawal ist jederzeit deaktiviert.** Der Key ist auf die KonnectAI-Execution-Server-IP ge-whitelisted. Der Investor kann den Key in Sekunden über das Binance-UI widerrufen, was sofort jegliches KonnectAI-Trading auf seinem Account stoppt.
- **Ein LLM-Call pro Cycle, Fan-out via die vier Investor-API-Keys.** Der Analyser läuft einmal pro Cycle pro Kandidat. Das resultierende Signal wird dann parallel via die vier Investor-API-Keys auf die vier Investor-Accounts executiert, mit Position-Grösse pro Account gegen die jeweils verfügbare Equity berechnet. Die Kosten der Analyse werden vier-fach geteilt: LLM-Kosten pro Investor = TotalLLM / 4.
- **Kapital-segregiert by construction.** Jede Equity-Curve, jede Ein-/Auszahlung und jede PnL eines Investors sind Eigenschaften seines eigenen Binance-Accounts und der Audit-Records, die auf seinen API-Vault-Slot verschlüsselt sind. Es gibt keinen gemeinsamen Kapital-Pool; Segregierung ist eine Tatsache über die zugrundeliegenden Binance-Accounts, keine buchhalterische Konvention auf unserer Seite.
- **Operator-Wallet für Cost-Routing via Binance Pay.** Operative Kosten (CHF 6'250-equivalent USDT pro Investor pro Monat, Abschnitt 14) werden monatlich automatisch

via den Binance-Pay-Endpoint (`/sapi/v1/pay/transactions`) vom Investor-Account zu einem dedizierten KonnectAI-Operator-Wallet überführt. Der Transfer erfolgt am 1. jedes Monats um 09:00 Europe/Zurich. Ein **investor-kontrollierter Settings-Toggle** steuert den Auto-Deduct (Default ON); schaltet der Investor ihn aus, pausiert das System das Trading auf diesem Account, bis die operativen Kosten manuell beglichen sind. Es gibt keine Vorauszahlung und keine Rückerstattung: Pay-as-you-go.

- **Hard-Cap von vier Investoren pro Cluster.** Operative, nicht willkürliche Grenze. Oberhalb von vier Investor-Accounts wird der Per-Account-Kapital-Anteil pro Signal zu klein, um effizient gegen die von Binance auf bestimmten Symbolen erzwungenen Round-Lot-Grössen zu sein, und die Varianz der Fan-out-Fill-Outcomes steigt. Der Vier-Cap stellt sicher, dass jeder Investor substantiell die gleiche Execution-Qualität sieht.
- **Skalierungs-Pfad.** Nachfrage über vier Investoren wird durch Deployment eines **zweiten Clusters** mit derselben Architektur abgedeckt: separate Hardware, separates Operator-Wallet, separater Vier-Investor-Pool von investor-eigenen Accounts. Jedes Cluster hat eigene LLM-Analysen und eigene Execution-Loop. Das ist Parallel-Scale-out, kein vertikales Fan-out über den Cap hinaus.

Die Pooled-Architektur-Ökonomik ist der Schlüssel zum Kostenmodell: per-Investor-operative-Kosten sind beschränkt, weil die grösste variable Kosten (LLM-Inferenz) innerhalb eines Clusters nicht mit Investor-Count skaliert. Die non-custodial Posture ist der Schlüssel zum regulatorischen Modell: Investor-Funds gelangen zu keinem Zeitpunkt in ein Wallet unter KonnectAI-Kontrolle, und KonnectAI kann zu keinem Zeitpunkt Funds von der Exchange wegbewegen (Abschnitt 12, Abschnitt 14).

0.13 12. Security und Risk-Posture

0.13.1 12.1 Secrets-Management

Production-Secrets — Investor-Exchange-API-Keys, LLM-Provider-Keys, DB-Credentials, Webhook-Tokens — werden in einem **verschlüsselten API-Key-Vault** abgelegt, einem von der Anwendungs-Code getrennten Credentials-Tier. Auf der aktuellen Skalierung ist das ein gehärteter, OS-permission-gesicherter verschlüsselter Store auf dem Execution-Server; auf Multi-Cluster-Skalierung sind AWS Secrets Manager oder HashiCorp Vault die natürlichen Upgrades. Investoren liefern ihren API-Key und Secret über einen verschlüsselten Kanal (PGP, Signal oder äquivalent) an KonnectAI, und der Wert wird direkt in den Vault aufgenommen.

Drei Invarianten:

- **Kein Secret tritt in einen Log ein.** Logging-Schichten scrubben bekannte Secret-

Pattern vor Emission.

- **Least-Privilege pro Secret.** Read-only-Market-Daten-Keys sind getrennt von Trade-Execution-Keys. **Jeder vom Investor bereitgestellte Trade-Execution-Key hat Withdrawal permanent deaktiviert und ist auf den KonnectAI-Execution-Server IP-ge-whitelisted**, gescoped nur auf Spot-Trading, Futures-Trading und die Binance-Pay / `Universal Transfer` Permission, die für den monatlichen Operating-Cost-Auto-Deduct nötig ist (Abschnitt 11.4, Abschnitt 14).
- **Investor-Widerrufbarkeit.** Der Investor behält jederzeit die Fähigkeit, seinen API-Key über das Binance-UI zu widerrufen; der Widerruf ist sofortig und stoppt umgehend jegliches KonnectAI-Trading auf seinem Account.

0.13.2 12.2 Authentifizierung und Authorisierung

Abschnitt 10.3 deckt TOTP für das Dashboard. Für programmatische Zugriffe:

- LLM-Provider-Zugriff über Provider-Keys, auf Schedule rotiert.
- Exchange-Read-only-Zugriff (sofern intern für Marktdaten genutzt) über IP-restricted API-Keys.
- Der Exchange-Trade-Zugriff jedes Investors erfolgt über einen vom Investor ausgestellten, IP-restricted Key mit Withdrawal permanent deaktiviert, gescoped nur auf Spot, Futures und `Universal Transfer` / Binance Pay.
- Investor-API-Keys sind im Vault siloed; ein Compromise eines Investor-Keys propagiert nicht auf die anderen, und im Worst-Case ist der Schaden durch den No-Withdrawal-Scope begrenzt: ein bösartiger Akteur könnte keine Funds exfiltrieren, sondern nur Trades auf dem betroffenen Account ausführen — auf welchem der Investor den Key in Sekunden widerrufen kann.

0.13.3 12.3 Rate-Limits

Inbound-API-Endpoints am Dashboard sind am Reverse-Proxy-Layer rate-limited. Outbound-API-Nutzung respektiert Provider-Limits (Etherscan 5/s, Coin Metrics Community lightweight, Anthropic Tier-passend, Binance Request-Weight-Scheduling). Backoff-with-Jitter für transiente Upstream-Failures.

0.13.4 12.4 Live-Trading-Safety-Rails

Der Live-Trading-Modus operiert mit den folgenden mechanischen Guards, die heute in Produktion sind:

- **Confirmation-Buttons in Telegram für jeden Trade über konfigurierbarer Grösse.** Operator muss vor Order-Submission auf dem betroffenen Investor-Account bestätigen.
- **Max-Daily-Loss-Circuit-Breaker.** Wenn kumulierter realisierter Verlust an einem Kalendertag eine Schwelle überschreitet, wird automatisches Trading bis zur

Operator-Reaktivierung deaktiviert.

- **Position-Sizing-Guardrail.** Grösse pro Trade durch Account-Equity-Prozent-Decke begrenzt, ungeachtet des LLM-Vorschlags.
- **Withdrawal deaktiviert auf jedem Investor-API-Key.** Kein KonnectAI-gehaltenes Credential, auf keinem Investor-Account, kann Funds vom Exchange bewegen. Das wird auf der Binance-Permission-Ebene durchgesetzt, nicht in unserer Application-Logik, und kann nur durch eine explizite Änderung des Investors am eigenen Key im Binance-UI re-aktiviert werden.
- **Cooldown nach Stop-out.** Per-Symbol-Cooldown verhindert Revenge-Trade-Re-Entry.
- **Regime-Change-Exit triggert nur in Profit über Fees.** Ein Regime-Change-Exit kann einen Winning-Trade nicht zur Execution-Zeit in einen Loser verwandeln.

Das sind keine neuen Ideen — standardmässige institutional-grade Sicherheitspraxis. Was zählt: sie sind eingebaut und seit Live-Going operativ.

0.14 13. Roadmap

0.14.1 13.1 Abgeschlossene Phasen (operativ seit Januar 2026)

- **Phase 0 — Foundation.** DB-Schema, Exchange-API-Client, Candle-und-Indikator-Pipeline.
- **Phase 1 — Daily Intelligence.** Pre-Filter, LLM-Analyser, Paper-Trading-Sandbox, Alerts, Briefings.
- **Phase 2 — Strategy-Modes.** Settings-Store, operator-switchable Posture, Modus-Tagging auf Trades.
- **Phase 3 — Dashboard.** TOTP-authentifizierte Web-UI, Public-/Private-Split.
- **Phase 4 — Macro.** Wochen-Macro-Report-Generator, Macro-Snapshot-Persistenz.
- **Phase 4.5 — On-Chain und Regime.** Sechs Free-Tier-On-Chain-Quellen; 5-State-Regime-Klassifizierer auf 4-Stunden-Candle-Stream.
- **Phase 5 — Live-Trading auf Spot.** Echtes Kapital, echte Fees, echte Slippage. Operativ seit Anfang 2026.
- **Phase 6 — Live-Trading auf Futures (Long und Short).** Bidirektionale Execution. Operativ.
- **Phase 7 — Capital-Event-aware Wallet-Buchhaltung.** Geometrisches Period-Linking für Netto-PnL %, Internal-Transfer-Bridge, Capital-Events-Sync. Operativ seit April 2026.
- **Phase 8 — Regime-Change-Exit.** Auto-Close auf adversem Regime-Flip in Profit. Operativ seit April 2026.
- **Phase 9 — Original-SL-Preservation, TP1-aware Sizing, Dust-Threshold-Portfolio-View.** Operativ seit April 2026.

0.14.2 13.2 Kontinuierliche Verbesserung (in Arbeit)

- **Production-Cron-Migration.** Der MVP läuft auf 15-Minuten-Core-Loop. Die Produktions-Deployment für die Vier-Investor-Cohort migriert auf 1-Minuten-Core-Loop, mit feinerer Candle-Ingest und höherem Inferenz-Budget pro Cycle. Operatives Rollout, kein Forschungsprojekt.
- **Pooled-Execution-Rollout.** Der Fan-out über mehrere Investor-API-Keys ist gegen Test-Accounts parallel zur Produktions-Execution-Schicht validiert. Onboarding des ersten investor-eigenen Binance-Accounts in den Live-Fan-out ist eine Deployment-Übung (Vault-Aufnahme + IP-Whitelist + Dry-Run-Signal-Echo + erster Live-Cycle), keine Build-out.
- **Adaptives Threshold-Tuning.** Strategie-Modus-Thresholds werden gegen Live-Daten auf festem Schedule reviewed. Der MVP-Paper-Trader ist die Validierungs-Bench für Revisionen.
- **Regime-bedingte Execution-Verfeinerungen.** Der Regime-Change-Exit-Branch ist operativ. Zukünftige Verfeinerungen umfassen regime-bedingtes Position-Sizing und regime-bedingte Take-Profit-Ladders.

0.14.3 13.3 Zukünftige Erweiterungen (kein Blocker für den Produktions-Pool)

- **Paid-Tier-On-Chain-Integration.** Glassnode Pro, Arkham-Intelligence-Label-Datenbank zur Lösung der Coinbase/Kraken-Smart-Contract-Custody-Lücke.
- **Social-Sentiment-Integration.** Twitter / Reddit-Sentiment als zusätzlicher Signal-Stream, vorbehaltlich Quality- und Dedupe-Disziplin. Semantische Deduplizierung von News-/Social-Signalen, um Übergewichtung von Echo-Chamber-Narrativen zu vermeiden.
- **Multi-Chain-Expansion.** Solana, Base, BSC als First-Class-Chains neben Bitcoin und Ethereum.
- **Multi-Cluster-Scale-out.** Über vier Investoren pro Cluster werden zusätzliche Cluster mit derselben Architektur parallel deployed.
- **Compliance- und Reporting-Paket.** Für Investoren, die die Plattform unter EU-MiFID-II-equivalent oder Schweizer FINMA-equivalent betreiben möchten.

Die Framing-Änderung gegenüber früheren Versionen dieses Dokuments ist materiell: die Roadmap ist kein Build-out-Plan mehr, sondern ein Continuous-Improvement- und Scale-out-Plan gegen ein operatives Baseline.

0.15 14. Das Investmentmodell

0.15.1 14.1 Was wir anbieten

KonnectAI Trader hat die Build-out-Phase abgeschlossen. **Wir suchen kein Kapital, um ein Team aufzubauen.** Wir öffnen einen kleinen, hart begrenzten Pool von Trading-Kapital-Slots auf dem operativen System.

Vier Investoren-Slots × CHF 100’000 = CHF 400’000 Trading-Kapital total pro Server-Cluster.

Jeder Slot ist in zwei Tranchen strukturiert:

- **Tranche 1: CHF 50’000 direkt bei Vertragsunterzeichnung.** Finanziert die Produktions-Cluster-Hardware-Beschaffung und das Per-Investor-Onboarding (Vault-Setup, Key-Aufnahme, IP-Whitelist, Dry-Run-Validierung). Das eigene Binance-Account des Investors wird vom Investor direkt mit Binance eröffnet; KonnectAI eröffnet oder hält kein Account im Namen des Investors.
- **Tranche 2: CHF 50’000, fällig einen Monat nach Bereitstellung der Produktions-Hardware und funktionalem Live-Trading unter Vertragsbedingungen.** Diese Tranche ist die zweite Hälfte des Trading-Kapitals. Nach Wahl des Investors kann sie in der Zwischenzeit über **Treuhandkonto (Escrow)** oder **Bankgarantie** abgesichert werden.

Total Kapital pro Investor: CHF 100’000. Total finanziertes Trading-Kapital pro Cluster, bei voller Belegung: **CHF 400’000**.

0.15.2 14.2 Was Investoren erhalten

Pro Investor:

- **Trading auf seinem eigenen Binance-Account, jederzeit voll im Eigentum und unter Kontrolle des Investors.** KonnectAI hält zu keinem Zeitpunkt Custody. Der Investor eröffnet das Account direkt mit Binance (ein Standard-Retail-Account ist ausreichend — kein VIP-Status nötig), durchläuft Binance’ KYC/AML, zahlt CHF/USD via einen Schweizer Fiat-On-Ramp seiner Wahl ein und konvertiert innerhalb des Accounts in USDT. Das Account verbleibt jederzeit auf den Namen des Investors, unter dessen direktem rechtlichen Eigentum.
- **Eine gescopte API-Integration mit KonnectAI.** Der Investor stellt KonnectAI einen API-Key mit folgendem Permission-Profil zur Verfügung (siehe Abschnitt 14.7 für die explizite Auflistung): Spot-&-Margin-Trading aktiviert, USDS-M-Futures-Trading aktiviert, `Universal Transfer` / Binance Pay aktiviert (für den monatlichen Operating-Cost-Auto-Deduct), **Withdrawal permanent deaktiviert**, IP-restricted auf den KonnectAI-Execution-Server. Der Investor kann diesen Key jederzeit über das Binance-UI widerrufen.
- **Eine Equity-Curve, Ein-/Auszahlungs-Log und PnL-Historie**, berechnet gegen die Wallet-Historie seines eigenen Accounts. Jede Investor-Sicht im Dashboard ist auf

sein Account gefiltert; Segregierung ist strukturell, nicht buchhalterisch.
- **Zugriff auf die Signale des operativen Systems, executiert gegen sein Account.** Auf jedem Cycle wird die LLM-Analyse via die vier Investor-API-Keys auf die vier Investor-Accounts gefächert, mit Position-Sizing gegen die jeweils verfügbare Equity.
- **Monatliches Compounding by default.** Netto-PnL verbleibt jeden Cycle im Account des Investors; es gibt keinen "Lock-up" der Profite, und das Default-Verhalten ist, dass die Renditen auf der Trading-Basis compounden. Der Investor kann optional in den Settings eine monatliche Auto-Distribution aktivieren — entweder als fixer Prozentsatz oder als fixer Betrag des Netto-PnL — auf eine externe Cold-Wallet seiner Wahl.
- **Quartalsweise Investor-Reports** mit dem wallet-basierten Netto-PnL %, der im Zeitraum beobachteten Regime-Verteilung und einer narrativen Zusammenfassung.

Was die Plattform **nicht** ist:

- **Kein Subscription-Produkt.** Der Investor zahlt keine monatliche Gebühr für "Signale".
- **Kein Signal-Service.** Signale werden nicht als kopierbares Artefakt geliefert; sie werden gegen das Account des Investors executiert.
- **Kein Fonds.** Kein Fondsvehikel, keine NAV-Berechnungs-Methodik, keine Fonds-Level-Gebührenstruktur. Jeder Investor ist rechtlicher Eigentümer der Assets in seinem eigenen Binance-Account.
- **Kein Custodian.** KonnectAI hält zu keinem Zeitpunkt Investor-Funds. Funds verbleiben jederzeit im direkt im Eigentum stehenden Binance-Account des Investors. API-Permissions sind ausschliesslich auf Trading und Binance-interne Transfers gescoped — Withdrawal-Permissions bleiben jederzeit deaktiviert.
- **Keine Vermögensverwaltung unter FINMA-Lizenz.** Es ist ein non-custodial Execution-Service; wir sind keine lizenzierten Vermögensverwalter und stellen uns auch nicht so dar. Der Investor behält jederzeit das rechtliche Eigentum und die operative Kontrolle über sein Account. Diese Positionierung stellt sicher, dass KonnectAI Trader **nicht den FINMA-lizenzierten Vermögensverwaltungs-Pflichten unterliegt**, weil wir keine Custody über Investor-Funds halten.

0.15.3 14.3 Operative Kosten

Das geteilte Cost-Modell:

- **CHF 6'250-equivalent USDT / Monat pro Investor** (CHF 75'000-equivalent / Jahr pro Investor).
- **Total bei voller Belegung: CHF 25'000-equivalent USDT / Monat** (CHF 300'000-equivalent / Jahr, das operative Envelope des Clusters).
- **Auto-Pay-Mechanik.** Am 1. jedes Monats um 09:00 Europe/Zurich wird der Operating-Cost-Betrag automatisch via den Binance-Pay-Endpoint (`/sapi/v1/pay/transactions`) vom Investor-Account zum dedizierten **KonnectAI-Operator-Wallet** transferiert. Der Transfer wird im Audit-Speicher mit derselben `external_id`-Disziplin wie jedes

andere Capital-Event geloggt.

- **Investor-kontrollierter Settings-Toggle.** Der Auto-Pay-Flow wird durch einen per-Investor-Settings-Toggle (Default ON) gesteuert, den der Investor in seiner Dashboard-Sicht kontrolliert. Schaltet der Investor den Toggle auf OFF, pausiert das System jegliches Trading auf seinem Account, bis der Operating-Cost-Betrag manuell beglichen wurde. Das ist ein **investor-souveränes Control-Feature**, kein Compliance-Workaround: der Investor entscheidet, ob die Operating-Costs jeden Monat bezahlt werden, und das System verweigert Trading ohne diese Autorisierung.
- **Pay-as-you-go.** Keine Operating-Cost wird vorausbezahlt. Bei Vertragsende ist kein Refund-Mechanismus nötig: der Investor widerruft einfach den API-Key und stoppt die Autorisierung der nächsten Auto-Pay; das Trading endet sofort.
- **Netto-PnL nach Kosten.** Was nach der monatlichen Auto-Pay im Account des Investors verbleibt, ist der Netto-PnL des Investors für den Zeitraum.

Die CHF 6’250/Monat Per-Investor-Zahl ist die Steady-State-Voll-Cluster-Operating-Cost vier-fach geteilt. Sie deckt aggregiert über das Cluster:

- LLM-Inferenz auf der Produktions-1-Minuten-Cron-Kadenz (grösster Einzelposten).
- Server-Hosting, einschliesslich dedizierter Produktions-Hardware und Managed-Services.
- Ein Operator-Stipendium für Monitoring, On-Call-Response, Threshold-Tuning, Prompt-Revisionen und Continuous-Improvement-Arbeit.
- Datenfeeds, Alerting-Infrastruktur und Observability-Tooling.

Die Kosten sind **fix pro Investor**, ungeachtet des Trading-PnL. Es gibt keine Performance-Gebühr, kein Carry, keine Hürde. Die Interessens-Alignment ist strukturell: das Operator-Einkommen ist die Operating-Cost-Rückerstattung, die beschränkt ist; der Investor behält 100 % des Netto-PnL nach Kosten.

Währungsnotiz. Trading-Kapital wird in USDT auf Binance gehalten. CHF-Beträge oben sind Referenzwerte für Schweizer Investoren; der tatsächliche Auto-Pay-Transfer erfolgt in USDT gegen den prevailing CHF/USDT-Spot-Kurs zum Zeitpunkt des Transfers.

0.15.4 14.4 Performance-Projektionen

Wir projizieren Performance unter drei Regime-Szenarien, auf MVP-Daten-Basis:

Marktregime	Daily-ROI-Projektion	Annualisiert (linear, 365 Tage)
Drawdown	0.35 %	+127 % p.a.
Sideways	0.50 %	+180 % p.a.
Bullish	0.65 %	+237 % p.a.

Das sind **Projektionen auf MVP-Daten-Basis**, keine historischen Track-Records. (Siehe Abschnitt 9.5 für die operativen Differenzierer gegenüber dem akademischen 15–30 %

Korridor.)

0.15.5 14.5 Capital-Deployment-Szenarien

Drei Deployment-Szenarien werden für Due-Diligence-Framing präsentiert. Jedes Szenario wird unter den drei Regime-Projektionen aus Abschnitt 14.4 evaluiert. Die CHF-100’000-Investition pro Slot ist der **Investor-Beitrag**; die **Trading-Kapital**-Spalte reflektiert Szenarien, in denen Investoren zusätzliches Kapital über den vertraglichen CHF-100’000-Slot allozieren.

Szenario 1 — Basisfall: CHF 100’000 Investition, CHF 100’000 Trading-Kapital.

Die Investition CHF 100’000 ist gleichzeitig das Trading-Kapital; kein zusätzliches Eigenkapital wird eingesetzt.

	Y1 PnL (nach Kosten)	Y2 kumulativ PnL (nach Kosten)
Drawdown 0.35 %	−CHF 49’000	+CHF 2’000
Sideways 0.50 %	+CHF 5’000	+CHF 110’000
Bullish 0.65 %	+CHF 59’000	+CHF 218’000

Kritische Beobachtung: Szenario 1 + Drawdown ist das Worst-Case-Argument. Im pessimistischsten Regime-Szenario endet Jahr 1 bei −CHF 49’000. Bis Jahr 2 kumulativ erreicht die Position Breakeven (+CHF 2’000). **Selbst im Worst-Case-Projektionspfad ist der Investor an der 24-Monats-Marke ungefähr ganz.** Das ist die wichtigste Zahl der Tabelle für risiko-averse Due-Diligence: der Downside-Pfad auf einer eigenständigen CHF-100’000-Verpflichtung ist auf einem Zwei-Jahres-Horizont unter der projizierten Regime-Mischung näherungsweise erholbar.

Szenario 2 — Verdoppelte Trading-Basis: CHF 200’000 Trading-Kapital, CHF 100’000 Investition.

Der Investor alloziert zusätzlich CHF 100’000 Eigenkapital neben dem Slot, was die Trading-Basis verdoppelt, während der Operating-Cost-Anteil unverändert auf CHF 75’000/Jahr bleibt.

	Y1 PnL (nach Kosten)	Y2 kumulativ PnL (nach Kosten)
Drawdown 0.35 %	+CHF 77’000	+CHF 254’000
Sideways 0.50 %	+CHF 185’000	+CHF 470’000
Bullish 0.65 %	+CHF 293’000	+CHF 686’000

Die Hebelwirkung zusätzlichen Eigenkapitals gegen den fixen Kostenanteil ist der strukturelle Vorteil von Szenario 2: die Operating-Cost ist gleich, aber die Trading-Basis, von der Prozent-Renditen generiert werden, ist verdoppelt.

Szenario 3 — Hohe Trading-Basis: CHF 500’000 Trading-Kapital, CHF 100’000 Investition.

	Y1 PnL (nach Kosten)	Y2 kumulativ PnL (nach Kosten)
Drawdown 0.35 %	+CHF 455’000	+CHF 1.01 M
Sideways 0.50 %	+CHF 725’000	+CHF 1.55 M
Bullish 0.65 %	+CHF 995’000	+CHF 2.09 M

Diese Werte sind vor Steuern. Sie stellen keine Prognose des Outcomes eines individuellen Investors dar; sie sind Projektionen unter dem MVP-abgeleiteten Regime-Modell. Szenario 3 illustriert den Operating-Cost-Amortisations-Vorteil bei Skalierung: die Per-Investor-Kosten bleiben CHF 6’250/Monat, aber die Trading-Basis dagegen ist die Fünf-Fache der vertraglichen Slot-Grösse.

Alle Werte in CHF sind Referenzwerte für Schweizer Investoren; tatsächliche Positionsgrössen, PnL und Operating-Cost-Transfers erfolgen in USDT gegen den prevailing CHF/USDT-Spot-Kurs. Investoren können einen beliebigen Betrag in ihr Binance-Account einzahlen; die CHF 100’000-Werte sind Beispiel-Deployments, keine vertraglich vorgeschriebenen Mindestguthaben.

0.15.6 14.6 Onboarding-Flow

Der Onboarding-Flow reflektiert die non-custodial Architektur: KonnectAI empfängt, hält oder bewegt zu keinem Zeitpunkt Investor-Funds. Der Investor kontrolliert das Account, den Einzahlungsweg und den API-Key. KonnectAI kontrolliert nur die Trading-Logik, die *durch* den API-Key operiert.

1. **Vertragsunterzeichnung, Tranche 1 bezahlt.** Der Investor unterzeichnet den Plattform-Zugangsvertrag und zahlt Tranche 1 (CHF 50’000) per Schweizer Bank-Überweisung oder Wire an die Banking-Entität des Operators. Das ist die Plattform-Access-Gebühren-Struktur, getrennt vom Trading-Kapital.
2. **Investor eröffnet / verwendet sein eigenes Binance-Account.** Ein Standard-Retail-Binance-Account ist ausreichend. Binance handhabt KYC und AML direkt mit dem Investor; KonnectAI ist an diesem Prozess nicht beteiligt.
3. **Investor zahlt CHF/USD via Schweizer Fiat-On-Ramp ein** (z.B. SEBA, Sygnum, Bitstamp, Kraken Pro oder einen anderen regulierten Weg seiner Wahl) und konvertiert in seinem Binance-Account in USDT. Der Betrag liegt vollständig im Ermessen des Investors; CHF 100’000 ist die in den Projektionen verwendete Beispiel-Baseline.
4. **Investor erstellt einen API-Key mit dem gescopten Permission-Profil**, dokumentiert in Abschnitt 14.7: Spot-Trading, Futures-Trading, `Universal Transfer` / Binance Pay, **Withdrawal deaktiviert**, IP-Whitelist auf den KonnectAI-Execution-Server.
5. **Investor liefert API-Key und Secret über einen verschlüsselten Kanal an KonnectAI**

(PGP, Signal, Wire oder äquivalent). KonnectAI nimmt den Key in den verschlüsselten Vault auf.

6. **KonnectAI registriert den Slot und führt einen Dry-Run-Validierungs-Cycle durch.** Ein Signal-Echo-Run bestätigt, dass der Key korrekt gescoped und IP-ge-whitelisted ist, ohne Live-Orders zu platzieren.
7. **Live-Trading startet innerhalb von 24 Stunden** nach Vault-Aufnahme und Dry-Run-Validierung.
8. **Einen Monat nach stabilem Produktions-Trading ist Tranche 2 fällig** (CHF 50'000), womit das vertragliche Plattform-Access-Total beglichen ist.

In jedem Schritt kontrolliert der Investor das Account, den Einzahlungsweg und den API-Key. Der Investor kann den API-Key in Sekunden über das Binance-UI widerrufen; der Widerruf stoppt sofort jegliches KonnectAI-Trading auf seinem Account.

0.15.7 14.7 Investor-API-Key-Permission-Profil

Der API-Key, den der Investor an KonnectAI ausstellt, hat folgendes Permission-Set, direkt im Binance-Account des Investors konfiguriert:

```
Spot & Margin Trading:    ENABLED
Futures Trading:          ENABLED  (USDS-M)
Universal Transfer:      ENABLED  (für monatlichen Operating-Cost-Auto-
Deduct via Binance Pay)
Withdrawal:          DISABLED (KonnectAI hat zu keinem Zeitpunkt Withdraw-
Rechte)
IP Restrictions:          [KonnectAI-Server-IP]
```

Dieses Profil ist die vertragliche Schnittstelle zwischen Investor und Plattform. Es ist **die** Custody-Garantie: selbst mit voller Kontrolle des API-Keys kann KonnectAI keine Funds vom Exchange wegbewegen. Der Worst-Case-Schaden eines Compromises auf der KonnectAI-Seite ist durch den No-Withdrawal-Scope begrenzt: ein bösartiger Akteur im Besitz des Keys könnte nur Trades auf dem Investor-Account ausführen — auf welchem der Investor den Key in Sekunden widerrufen kann.

0.15.8 14.8 Risiko-Disclosure (gekürzt; volle Disclosure im Investor-Subscription-Paket)

- **Drawdown-Risiko.** Selbst ein gut-getuntes System wird Drawdown-Phasen durchlaufen. Die Spalte Szenario 1 + Drawdown in Abschnitt 14.5 ist der projizierte Worst-Case; tatsächliche Drawdown-Phasen könnten tiefer oder länger sein als projiziert.
- **Regime-Shift-Risiko.** Ein Regime-Übergang, den der Klassifizierer nicht erfasst, kann ein überproportional verlustreiches Cluster produzieren. Daher wird der Klassifizierer als First-Order-Diagnostik überwacht (Abschnitt 11.2).

- **Operatives Risiko.** Das Vier-Investor-Cluster teilt eine einzige operative Umgebung. Operative Vorfälle (LLM-Provider-Outage, Exchange-API-Outage, Operator-Fehler) treffen alle Investor-Accounts simultan. Mitigations in Abschnitt 12.4.
- **Liquiditäts-Risiko.** Bei sehr grossen Account-Equities wird Slippage auf dünn gehandelten Altcoins zu materiell grösserem Drag, was effektive Renditen mindert. Watchlist-Komposition und Sizing sind cluster-level-kapital-skalen-bedingt.
- **Regulatorisches Risiko.** EU-MiCA, Schweizer FINMA und US-equivalent regulatorische Entwicklungen könnten beeinflussen, welche Venues und Instrumente zugänglich sind, und könnten die rechtliche Positionierung von non-custodial Execution-Services beeinflussen. Aktuelle rechtliche Positionierung ist non-Fonds, non-custodial, non-Vermögensverwaltung; zukünftige regulatorische Entwicklungen könnten Restrukturierung erfordern.
- **Provider-Risiko.** Abhängigkeit von Anthropic (LLM), Binance (Exchange) und On-Chain-Datenprovidern. Mitigation: Multi-Provider-Abstraktions-Schichten in der Architektur, vorgebaute Switchover-Pfade.
- **Cluster-Level-Konzentration.** Alle vier Investoren in einem Cluster teilen die gleiche operative Logik. Diversifikation über Cluster-Operatoren ist nicht im Scope dieses Dokuments; das Zweit-Cluster-Scale-out (Abschnitt 11.4) ist die architektonische Antwort auf nachfrageseitige Diversifikation.

0.15.9 14.9 Warum das Buy-and-Hold schlägt

Im Prinzip hat Buy-and-Hold-BTC über Fünf-Jahres-Fenster starke absolute Renditen produziert — allerdings mit Drawdowns von über −70 % im 2022er-Zyklus. Die KonnectAI-Trader-Begründung gegenüber passivem Halten ruht auf drei Säulen:

1. **Performance-Kontinuität.** Die bidirektionale Execution-Schicht profitiert in Bull-, Sideways- und Drawdown-Regimes. Eine passive Position ist dem direktionalen Bias des Markts über den Halte-Zeitraum ausgesetzt.
2. **Drawdown-Management.** Regime-aware Execution mit Regime-Change-Exit ist darauf ausgelegt, Drawdown-Tiefe zu begrenzen. Passives Halten hat keinen solchen Mechanismus.
3. **Risk-managed, nicht direktional.** Das System ist als risk-managed Continuous-Performance-Instrument positioniert. Passives Halten ist konstruktionsbedingt eine direktionale Wette.

Für Investoren, deren Nutzenfunktion Drawdown stark gewichtet, oder die Krypto-Exposure ohne regime-blindes direktionales Risiko wollen, ist das aktive Capital-Access-Modell strukturell anders als eine passive Position.

0.16 15. MVP-Status und Track-Record

0.16.1 15.1 Operativ seit Januar 2026

KonnectAI Trader läuft live, gegen Kai Zehs eigenes Kapital, auf Binance Spot und Binance Futures, seit **Januar 2026**. Der aktuelle Feature-Stand ist seit **April 2026** in kontinuierlichem Produktionsbetrieb. Dieser Abschnitt fasst zusammen, was heute, vor Onboarding der Vier-Investor-Produktions-Cohort, operativ und verifizierbar ist.

0.16.2 15.2 Was heute in Produktion ist

Trading-Schicht.

- Live-Trading auf Binance Spot (Long-only, in ausgewählten USDT-Paaren).
- Live-Trading auf Binance Futures (Long und Short), mit Hebel und Margin pro Trade getrackt.
- Bidirektionale Execution: der Analyser emittiert BUY (Long) oder SELL (Short) Actions; die Execution-Schicht routet entsprechend.
- Partial-Take-Profit-mit-Runner über Spot und Futures.
- Original-SL-Preservation über alle Trades; TP1-Trail überschreibt nicht den originalen Stop im Audit-Record.
- TP1-aware Position-Sizing: Risiko pro Trade gegen Entry-zu-Original-SL-Distanz berechnet.
- Regime-Change-Exit: Auto-Close auf adversem Regime-Flip, wenn hypothetischer PnL Round-Trip-Fee-Threshold überschreitet.
- Cooldown-Enforcement nach Stop-out, pro Symbol.

Buchhaltungs-Schicht.

- Wallet-basierter Netto-PnL %, berechnet über geometrisches Period-Linking über Capital-Event-Grenzen.
- Capital-Events-Sync mit Binance: Ein-, Auszahlungen und Spot↔Futures-Internal-Transfers alle in die Equity-Curve rekonziliert, mit idempotentem Re-Sync über UNIQUE `external_id`.
- Internal-Transfer-Bridge-Logik: Phantom-Dip-Eliminierung auf der Aggregat-Spot-/Futures-Equity-Curve.
- Asset-Breakdown-Stacked-Area-Chart auf der Dashboard-Portfolio-View, mit Capital-Event-Markern.
- Dust-Threshold-Filter auf der Portfolio-View.
- Reconciliation-Drift-Monitoring: kontinuierlicher Vergleich der berechneten Equity gegen Binance-Wallet-Snapshot.

Audit-Schicht.

- Vier primäre Trade-und-Event-Repos: `live_trades`, `futures_trades`, `capital_events`, `internal_transfers`. Jede Zeile mit Zeitstempel, Payload, Zuschreibung und Pipeline-Version.

- Voller LLM-Analysen-Archiv mit Reasoning-Text, Regime-Label, On-Chain-Kontext und Strategie-Modus-Tag.
- Migrationen sind additiv; keine destruktiven Schema-Änderungen waren erforderlich.

Operations-Schicht.

- 15-Minuten-Core-Loop-Cron, läuft auf jedem Cycle ohne verpasste Executions in nachhaltigen Fenstern.
- Stündliche On-Chain-, Macro- und Portfolio-Snapshots.
- Zwei-Mal-Tägliche Telegram-Briefings.
- Wochen-LLM-generierter Macro-Report.
- Live-KPIs-Dashboard, das aktuelle offene Trades, jüngste Fills, Equity-Curve-Health reflektiert.
- Health-Check-Probes verifizieren geschedule-Job-Completion und alerten auf Anomalien.

Dashboard.

- TOTP-authentifizierte Private-Surface und sanitarisierte Public-Surface.
- Authentifizierte Sichten: volle Analysen, volle Trades (Live und Futures), Portfolio mit bridge-fixed Equity-Curve und Capital-Event-Markern, Capital-Events-Log, Settings, Macro-Reports-Archiv.
- Öffentliche Sichten: Watchlist, Charts, sanitarisierte Analysen-Verteilung, On-Chain-Summary, Daily-Cost-Telemetrie.

0.16.3 15.3 Position der Stärke

Die Unterscheidung dieser Whitepaper-Version gegenüber früheren Framings ist strukturell. Wir suchen kein Funding für eine 18-Monats-Build-out; die Build-out ist abgeschlossen. Wir bitten Investoren nicht, eine These auf akademischer Literatur zu validieren; die These wird heute live, auf echtem Kapital, gemessen. Der in Abschnitt 13 beschriebene Continuous-Improvement- und Production-Migrations-Plan ist inkrementelle Arbeit gegen ein operatives Baseline, kein Forschungsprojekt.

Das ist die Position, aus der wir das Vier-Slot-Capital-Access-Modell anbieten: ein laufendes System, kapazitätsbeschränkt durch die Vier-Investor-Fan-out-Ökonomik, mit sauberer Kosten-und-Einkommens-Trennung zwischen Investor-PnL (der im eigenen Account des Investors verbleibt) und Operator-Stipendium (monatlich automatisch via Binance Pay unter investor-kontrolliertem Toggle bezahlt).

0.17 16. Referenzen

Public-Datenprovider referenziert. Binance Public Market-Data API; Binance Spot, Futures und Pay APIs (`/sapi/v1/pay/transactions` für den Auto-Pay-Flow); Coin

Metrics Community API; Etherscan V2 Public API; blockchain.com Charts API; mempool.space API; Bitnodes API; DefiLlama API; Alternative.me Fear & Greed API; CoinGecko Public API.

LLM-Provider referenziert. Anthropic Claude (Opus 4-7 Familie).

Akademische Frameworks referenziert. Hidden Markov Models (Regime-Klassifikations-Literatur); Reinforcement Learning (FinRL-Familie); Reflective LLM-Agent Pattern (CryptoTrade).

0.18 Anhang A: Referenz-Implementation — Das operative System

Die vorhergehenden Abschnitte beschreiben KonnectAI Trader als konzeptuelles Framework: was es ist, warum jede Komponente existiert, wie sie zusammenspielen. Dieser Anhang dokumentiert, was wir tatsächlich gebaut haben — seit Januar 2026 live, mit dem aktuellen Feature-Stand seit April 2026 in kontinuierlicher Produktion. Der Zweck: die Investoren-Entscheidung zu de-riskieren. Wir bitten Investoren nicht, eine These auf akademischer Literatur zu finanzieren; wir bieten Capital-Access-Slots auf einem laufenden operativen System an, dessen architektonische Solidität direkt verifizierbar ist.

0.18.1 A.1 Implementations-Übersicht

Das Live-System besteht aus zwei koordinierten Services:

- **Ein Analyse- und Trading-Backend** (Python-basiert, Paket `crypto_agent`). Läuft auf einem geschedule-Job-Modell. Verantwortlich für Daten-Ingestion, Indikator-Berechnung, Regime-Klassifikation, Pre-Filter, LLM-basierte Analyse, Live- und Futures-Trading, Paper-Trading-Sandbox, Capital-Events-Sync, Regime-Change-Exit-Logik, Alerting, Briefings und Wochen-Macro-Report-Generator.
- **Ein Web-Dashboard** (Node.js / Express / EJS). Bedient die in Abschnitt 10 beschriebene Operator-UX, mit TOTP-authentifizierter Private-Surface und sanitarisierter Public-Surface. Liest aus dem gleichen Persistenzspeicher wie das Backend.

Beide Services laufen im MVP auf einer einzigen AWS-EC2-Ubuntu-Instanz, hinter einem Caddy-Reverse-Proxy mit Let's-Encrypt-verwaltetem TLS. Daten persistieren in einer PostgreSQL-Datenbank auf derselben Instanz. Observability via strukturierte Logs und journald.

Die Produktions-Deployment für die Vier-Investor-Cohort trennt Compute-Tiers (Analyse + Execution + Dashboard auf dedizierter Produktions-Hardware), läuft den Core-Loop in 1-Minuten-Kadenz (vs. 15 Minuten im MVP) und operiert auf **investor-eigenen Binance-Accounts** via per-Investor gescopter API-Keys, die in einem verschlüsselten Vault gehalten werden — KonnectAI hält zu keinem Zeitpunkt Custody über Investor-Funds (Abschnitte 11.4, 12, 14).

0.18.2 A.2 Technologie-Stack (konkret)

Backend-Services (Python). Modulares Paket-Layout unter `crypto_agent.src`:

- `market_data.py` — Binance-Candle-Ingest in mehreren Intervallen; Indikator-Berechnung (RSI, MACD, Bollinger, EMAs, ATR, Volume-Ratios) auf Completed-Candle-Baseline.
- `indicators.py` — Indikator-Bibliothek, stateless und pure für Testbarkeit.
- `prefilter.py` — deterministischer regelbasierter Pre-Filter; emittiert Kandidaten pro Cycle.
- `screener.py` — Multi-Trigger-Kompositions-Logik.
- `regime.py` — 5-State-Kalena-Klassifizierer auf 4-Stunden-Candle-Stream; emittiert eine Zeile pro `(symbol, captured_at)` pro Cycle.
- `regime_exit.py` — Regime-Change-Exit-Helper (beschrieben in A.4).
- `analyzer.py` — LLM-Analyser; konstruiert den strukturierten Prompt und dispatcht an Anthropic.
- `anthropic_client.py` — LLM-Client-Wrapper mit Retry und Schema-Validierung.
- `onchain.py` — Multi-Source-On-Chain-Snapshot-Sync.
- `macro.py` — Macro- und Sentiment-Snapshot-Sync.
- `paper_trading.py` — MVP-only Paper-Trader-Sandbox.
- `live_trading.py` — Spot-Live-Trading: Place / Track / Reconcile / Close.
- `futures_trading.py` — Futures-Live-Trading (Long und Short): Hebel, Margin, Funding.
- `capital_events.py` — Binance-Deposit-/Withdraw-/Internal-Transfer-Sync (beschrieben in A.3).
- `binance_client.py` — Binance-API-Client mit Retry, Rate-Limit und per-Investor-Key-Scoping (eine Client-Instanz pro Investor-Key, gelookuppt aus dem Vault).
- `portfolio.py` — Portfolio-Aggregation, Dust-Threshold-Filter, Asset-Breakdown-View.
- `briefings.py`, `alerts.py`, `telegram.py` — Reporting-Schicht.
- `cli.py` — Operator-CLI für Ad-hoc-Aufrufe.
- `db.py` — PostgreSQL-Connection-Pool + Helpers.

Datenbank. PostgreSQL mit relationalem Schema für Preis- und Indikator-Daten, JSONB-Spalten für reichere Payloads (Analyse-Reasoning, Regime-Metriken, On-Chain-Raw-Snapshots, Trade-Notes). Indizes auf Symbol+Zeitstempel für Ingest-Hot-Paths und auf Analyse-/Trade-ID für Dashboard-Read-Paths.

Migrationen. Append-only additive Migrationen in `agent/scripts/`. Ausgewählte Migrationen: - `migrate_v2.sql`, `migrate_phase2.sql` — Baseline-Schema-Evolution. - `migrate_regime.sql`, `migrate_regime_exit.sql` — Regime-Persistenz + Exit-Branch. - `migrate_live_trading.sql`, `migrate_futures_trading.sql` — Live-Trade-Tabellen. - `migrate_capital_events.sql` — Capital-Events- + Internal-Transfers-

Tabellen. - `migrate_partial_tp.sql`, `migrate_original_sl.sql` — TP/SL-Handling-Verfeinerungen. - `migrate_emergency_close.sql`, `migrate_emergency_close_futures.sql` — Manual-Close-Audit. - `migrate_dust_sweep.sql` — Portfolio-Dust-Threshold-Support. - `migrate_macro.sql`, `migrate_onchain.sql`, `migrate_prefilter.sql` — Begleitende Tabellen. - `backfill_original_sl.sql` — Historischer Backfill von `original_SL` für Pre-Preservation-Trades.

LLM. Anthropic Claude Opus 4-7 über die Standard-Anthropic-API. Strukturierte-Output-Prompts mit JSON-Schema; Single-Retry-Policy auf schema-invaliden Antworten (in Produktion selten ausgeübt).

Scheduling. Kombination aus systemd-Timer-Units (Long-running-Services) und cron (der Core-Loop und geschedule Briefings). MVP läuft auf 15-Minuten-Core-Loop. Produktion migriert auf 1-Minuten-Core-Loop.

Frontend. Node.js / Express / EJS server-seitig gerendertes HTML. TradingView `lightweight-charts` für Candle-/Indikator-Rendering; Chart.js für Equity-Curves und Verteilungs-Charts. Server-seitig gerendert mit gezielter Client-side-Verbesserung — bewusste Wahl für Einfachheit und Operator-freundliche Performance. Cache-Header explizit auf HTML- und API-Antworten deaktiviert, um Stale-Data-Verwirrung in der Operator-Browser-Sitzung zu vermeiden.

Deployment. AWS EC2 Ubuntu LTS. Caddy-Reverse-Proxy. Let's-Encrypt-TLS. Application-Layer-Secrets in einem Credentials-Verzeichnis mit OS-Permissions.

Observability. Strukturierte Logs via journald-Sink, plus Application-Level-Log-Datei. Periodische Health-Check-Skripte verifizieren geschedule-Job-Completion und alerten via Telegram auf Anomalien.

0.18.3 A.3 Capital-Events-Pipeline

Das `capital_events.py`-Modul versorgt den wallet-basierten Netto-PnL via geometrisches Period-Linking. Drei Binance-Quellen werden gezogen und rekonziliert:

- `get_deposit_history` (Status=1 → Erfolg).
- `get_withdraw_history` (Status=6 → abgeschlossen).
- `get_universal_transfer_history` (Spot↔Futures, Typen `MAIN_UMFUTURE` und `UMFUTURE_MAIN`) — separat als Audit-only gespeichert.

Kritische Kai-Direktive: interne Spot↔Futures-Transfers werden **nie** als Capital-Events gezählt. Sie landen ausschliesslich in der `internal_transfers`-Tabelle. Sie als Ein-/Auszahlungen zu behandeln, würde die Equity-Curve-Methodik korrumpieren.

USD-Equivalent-Pricing. Für non-USDT-Capital-Events wird das USD-Equivalent via 1-Minuten-Kline-Close auf `{asset}USDT` zum exakten `occurred_at`-Zeitstempel berechnet. Das produziert historisch akkurate USD-Werte statt Aktuell-Preis-Näherungen. Stablecoins (`USDT`, `BUSD`, `USDC`, `FDUSD`, `TUSD`, `DAI`) werden als USD-pegged auf $1 behandelt.

Idempotenz. Jede Zeile trägt eine Binance-seitige `external_id` als UNIQUE markiert. Der Sync ist beliebig oft re-runnable, ohne Capital-Events zu duplizieren. Schematisch:

```
# capital_events.py (Auszug, vereinfacht)

USDT_ASSETS = {"USDT", "BUSD", "USDC", "FDUSD", "TUSD", "DAI"}
DEPOSIT_STATUS_SUCCESS = 1
WITHDRAW_STATUS_COMPLETED = 6
TRANSFER_SPOT_TO_FUTURES = "MAIN_UMFUTURE"
TRANSFER_FUTURES_TO_SPOT = "UMFUTURE_MAIN"

def _get_usd_price_at(asset: str, ts_ms: int) -> float | None:
    """Historischer USD-Preis via 1m-Kline-Close zum ts_ms."""
    if asset in USDT_ASSETS:
        return 1.0
    # ... 1m-Kline-Lookup gegen {asset}USDT, Close zum ts_ms

def sync_capital_events() -> None:
    # 1) Deposits, Withdrawals via Binance-Spot-Endpoints ziehen.
    # 2) usd_value zum occurred_at via _get_usd_price_at berechnen.
    # 3) Upsert in capital_events (external_id UNIQUE).
    # 4) Universal-Transfer-History ziehen; nur in internal_transfers ablegen.
```

Wallet-basierte Return-Berechnung. Mit Capital-Events, die die Equity-Curve in deposit-aware Perioden partitionieren, ist die Per-Period-Rendite:

`r_i = (equity_end_i − equity_start_i − net_capital_event_i) / equity_start_i`

und die Total-Rendite die geometrische Verkettung:

`total_return = (1 + r_1) × (1 + r_2) × ... × (1 + r_n) − 1`

Das ist die auf der Dashboard-Equity-Curve-View als **wallet-basierter Netto-PnL %** berichtete Zahl. Sie übersteht Ein- und Auszahlungen, was die Eigenschaft ist, die eine defensible Renditebehauptung von einer naiven Equity-Delta-Behauptung trennt.

0.18.4 A.4 Regime-Change-Exit-Helper

Das `regime_exit.py`-Modul ist die ausführbare Form des Regime-Change-Exit-Branch (Abschnitt 5.5). Pure-Function-Design: es berechnet, ob ein Close erforderlich ist; der Caller (`live_trading.update_live_trades` und `futures_trading.update_futures_trades`) führt das tatsächliche Close durch.

Kern-Konstanten und Mappings:

```
# regime_exit.py (Auszug)
```

```
# Single Source of Truth für Taker-Fees.
SPOT_TAKER_FEE_PCT    = 0.075   # 0.10% Standard, 0.075% mit BNB-Rabatt
FUTURES_TAKER_FEE_PCT = 0.04    # 0.05% Standard, 0.04% mit BNB-Rabatt

# Interne Kalena-Labels → Grobrichtung.
_REGIME_DIRECTION = {
    "trend_up_liquid":   "bull",
    "trend_down_liquid": "bear",
    "crash":             "bear",       # als starkes Gegen-LONG behandelt
    "range":             "sideways",
    "vol_expansion":     "sideways",   # ambiguous → sicherer: keine Action
}

NEUTRAL_DIRECTIONS = {"sideways", "neutral", "unknown", None}
```

Die Exit-Entscheidungs-Logik in Klartext:

1. Map des aktuellen Regimes auf eine Richtung (`bull`, `bear`, `sideways`, unknown).
2. Falls neutral oder unbekannt: skip (keine Action).
3. Falls die Richtung jetzt gegen die Trade-Seite steht (Long-vs-Bear oder Short-vs-Bull):
   - Berechne hypothetischen realisierten PnL % zum aktuellen Preis, auf **Notional** (`qty × Preis`).
   - Vergleiche gegen Round-Trip-Taker-Fees (Entry + Exit).
   - Falls hypothetischer PnL den Round-Trip-Fee-Threshold überschreitet: Trade mit `exit_reason='regime_change'` schliessen.
   - Falls hypothetischer PnL unter Threshold: ursprünglichen Stop und TP weiter aktiv lassen.

Hebel-Anmerkung. Fee-/Funding-Mathematik ist auf Notional, was hebel-agnostisch ist. Das Futures-`leverage`-Setting tritt nicht in diese Berechnungen ein — explizit während des Live-Trading-Rollout verifiziert.

Asymmetrische-Verlust-Eigenschaft. Der Branch feuert nie auf einem verlierenden Trade. Er kann einen ausgestoppten Trade nicht zu einem früheren Verlust verwandeln. Seine einzige Funktion ist es, Profit zu sichern, wenn eine profit-tragende Position einem feindlichen frisch klassifizierten Regime ausgesetzt ist, in dem die Fortsetzungs-Erwartung negativ gekippt ist.

0.18.5 A.5 Equity-Curve-Bridge für interne Transfers

Die Produktions-Deployment liefert dem Operator eine vereinheitlichte Spot-+Futures-Equity-Curve. Interne Spot↔Futures-Transfers sind asynchrone Ereignisse, die naiv über Snapshot-Fenster summiert **Phantom-Dips** produzieren — momentane scheinbare Equity-Drops, während der überwiesene Betrag kurz aus dem Snapshot des einen Wallets

verschwunden, aber im Snapshot des anderen noch nicht angekommen ist.

Die Bridge-Logik, beim Chart-Rendering im Dashboard appliziert:

1. Alle `internal_transfers`-Zeilen mit `(occurred_at, from_wallet, to_wallet, asset, amount, usd_value)` ziehen.
2. Für jedes benachbarte Equity-Snapshot-Paar klassifizieren, ob ein Internal-Transfer dazwischen liegt.
3. Falls ja: den überwiesenen Betrag dem Source-Wallet bis zum To-Time zuschreiben — eliminiert den Phantom-Dip auf der Aggregat-Curve.
4. Die zwei Komponenten-Curves (nur Spot, nur Futures) bleiben unbridged — jede zeigt ihre eigene reale Bewegung; die Bridge gilt nur für die Aggregat-Curve.

Der praktische Effekt: eine glatte Aggregat-Curve durch den Moment der Internal-Transfer-Execution, mit Capital-Event-Markern (echte Ein-/Auszahlungen) als diskrete Punkte überlagert.

0.18.6 A.6 Original-SL-Preservation und TP1-aware Sizing

Zwei Verfeinerungen, sowohl auf Spot- als auch Futures-Live-Trading appliziert:

Original-SL-Preservation. Wenn TP1 erreicht wird, wird der Runner-Stop auf das TP1-Level gezogen (sichert den TP1-Gewinn auf dem Runner-Bein). Allerdings wird der *originale* Stop-Level als `original_SL` am Trade-Record erhalten. Zwei Gründe:

1. **Risiko-Attribution.** Position-Sizing wurde gegen die Entry-zu-Original-SL-Distanz berechnet. Post-hoc-Analyse von "wie weit ist der Trade gegen die ursprüngliche Risiko-Prämisse maximal in den Drawdown gelaufen?" erfordert, dass der originale Stop abfragbar bleibt.
2. **Audit-Integrität.** Ein Leser des Trade-Records nach Closure muss den Trade so rekonstruieren können, wie die Strategie ihn gesetzt hat, nicht wie er sich entwickelt hat.

Die `migrate_original_sl.sql`-Migration fügt die `original_SL`-Spalte hinzu; `backfill_original_` füllt sie für historische Trades.

TP1-aware Position-Sizing. Risiko pro Trade wird als `(entry − original_SL) × qty` berechnet, nicht als `(entry − stop_active) × qty`. Das ist relevant, weil sobald TP1 den aktiven Stop nach oben gezogen hat, die Active-Stop-Distanz zur Post-Trail-Distanz wird. Sizing auf dieser Post-Trail-Distanz würde neue Trades systematisch über-sizen. Sizing auf der Original-SL-Distanz hält das Risiko-Envelope konsistent mit der pre-Trade-Intention der Strategie.

0.18.7 A.7 Gepoolte Vier-Investor non-custodial Fan-out-Architektur

Die Produktions-Deployment für die Vier-Investor-Cohort läuft auf **investor-eigenen Binance-Accounts** — nicht auf Binance-Sub-Accounts unter einem KonnectAI-Master. Der Sub-Account-Weg würde Binance-VIP-Status (~$3.8 M+ Holdings oder $50 M+ monatliches Volumen) erfordern, den wir nicht haben, und würde Investor-Funds unter unsere Kontrolle bringen — genau die Custody-Posture, die wir ablehnen. Das Design ist stattdessen by construction non-custodial: jeder Investor hält und kontrolliert sein eigenes Standard-Retail-Binance-Account; KonnectAI operiert *durch* einen vom Investor ausgestellten gescopten API-Key.

Setup (pro Investor).

1. Der Investor eröffnet (oder verwendet ein bestehendes) Standard-Retail-Binance-Account. KYC und AML werden direkt zwischen dem Investor und Binance gehandhabt.
2. Der Investor zahlt CHF/USD via Schweizer Fiat-On-Ramp ein (z.B. SEBA, Sygnum, Bitstamp, Kraken Pro) und konvertiert in seinem Account in USDT. Der Betrag liegt im Ermessen des Investors; CHF 100’000-equivalent USDT ist die in den Projektionen verwendete Beispiel-Baseline.
3. Der Investor erstellt einen Binance-API-Key mit dem vertraglichen Permission-Profil: Spot-&-Margin-Trading, USDS-M-Futures-Trading, `Universal Transfer` / Binance Pay (für den monatlichen Operating-Cost-Auto-Deduct), **Withdrawal permanent deaktiviert**, IP-Whitelist auf den KonnectAI-Execution-Server.
4. Der Investor liefert API-Key + Secret über einen verschlüsselten Kanal (PGP, Signal, Wire) an KonnectAI. KonnectAI nimmt den Key in den verschlüsselten API-Key-Vault auf.
5. Der User-Record des Investors im Auth-Speicher wird an den entsprechenden Vault-Slot gebunden; seine Dashboard-Sicht filtert alle Daten nach seinem Vault-Slot-Identifier.
6. KonnectAI führt einen Dry-Run-Validierungs-Cycle (Signal-Echo, keine Live-Orders) durch, um zu verifizieren, dass der Key korrekt gescoped und IP-ge-whitelisted ist, und das Live-Trading startet innerhalb von 24 Stunden.

Operating-Loop (pro Cycle).

```
für jeden Cycle:
    candidates = screener.scan(watchlist)
    für candidate in candidates:
        analysis = analyzer.analyse(candidate)        # 1 LLM-Call
        decision = validators.evaluate(analysis)
        falls decision.action in (BUY, SELL):
            für investor_account in subscribed_investor_accounts:
          # investor_account hält (vault_slot_id, api_key_handle, equity, policy)
```

```
            size = sizer.size_for(
        investor_account.equity, decision, investor_account.policy)
            executor.place_order(investor_account, decision, size)
    persist(analysis, decision, fills)
```

Der LLM-Call ist **einmal pro Kandidat pro Cycle**, nicht pro Investor-Account. Das Fan-out erfolgt in der Execution-Schicht, via vier parallele `binance_client`-Instanzen, jede an einen anderen Investor-API-Key gebunden, wo die Grösse mit der auf dem jeweiligen Investor-Account verfügbaren Equity skaliert. LLM-Kosten pro Investor = TotalLLM / 4.

Operating-Cost-Auto-Pay (Binance Pay). Einmal pro Monat — am 1. um 09:00 Europe/Zurich — wird der Operating-Cost-Betrag (CHF 6'250-equivalent USDT) von jedem Investor-Account auf das dedizierte KonnectAI-Operator-Wallet via den Binance-Pay-Endpoint (`/sapi/v1/pay/transactions`) überführt. Der Transfer ist idempotent (UNIQUE `external_id`) und audit-geloggt identisch wie Capital-Events. Der Flow ist gegated durch einen **investor-kontrollierten Settings-Toggle** (Default ON): schaltet der Investor ihn auf OFF, pausiert das System das Trading auf seinem Account, bis die Operating-Cost manuell beglichen ist. Pay-as-you-go: keine Vorauszahlung, keine Rückerstattung, kein Carry.

Profit-Distribution. Default-mässig verbleibt Netto-PnL jeden Cycle im Account des Investors (Compound by default). Der Investor kann optional in den Settings eine monatliche Auto-Distribution aktivieren — entweder als fixer Prozentsatz oder als fixer Betrag des Netto-PnL — auf eine externe Cold-Wallet seiner Wahl. Es gibt keine fonds-artige "Distribution", weil es keinen Fonds gibt; was im Account des Investors steht, *ist* sein Saldo.

Custody-Garantie. Kein KonnectAI-gehaltenes Credential, auf keinem Investor-Account, kann Funds vom Exchange wegbewegen. Das wird auf der Binance-Permission-Ebene durchgesetzt (Withdrawal: DISABLED), nicht in unserer Application-Logik. Der Investor kann den API-Key jederzeit über das Binance-UI widerrufen; der Widerruf ist sofortig und stoppt umgehend jegliches KonnectAI-Trading auf seinem Account.

Hard-Cap. Die Architektur erzwingt ein Maximum von vier Investor-Accounts pro Cluster. Über vier hinaus degradieren Fan-out-Varianz und Round-Lot-Ineffizienzen die Execution-Qualität. Die Skalierungs-Antwort ist die Deployment eines zweiten Clusters (separate Hardware, separates Operator-Wallet, separater Vier-Investor-Pool von investor-eigenen Accounts), nicht die vertikale Aufstockung eines einzelnen Clusters.

0.18.8 A.8 Cron-Kadenz-Migration: MVP 15min → Production 1min

Der MVP läuft auf 15-Minuten-Core-Loop-Cron (96 Cycles/Tag). Die Produktions-Deployment läuft auf 1-Minuten-Core-Loop-Cron (1'440 Cycles/Tag). Die Migration ist operativ, nicht architektonisch — dieselben Code-Pfade, dasselbe Datenbank-Schema, dieselbe Execution-Logik. Was sich ändert:

- **Candle-Ingest-Kadenz.** Der 1-Minuten-Candle-Stream wird in Produktion in das Ingest-Set aufgenommen. MVP synchronisiert auf 15-Minuten-Candle-Close; Produktion auf 1-Minuten-Candle-Close.
- **LLM-Inferenz-Budget.** Der Produktions-Cluster ist mit einem Inferenz-Cost-Budget provisioniert, das mit 1'440 Cycles/Tag über den Vier-Investor-Account-Fan-out konsistent ist. Budget-Alerts werden entsprechend kalibriert.
- **Indikator-Berechnungs-Hot-Path.** Der 1-Minuten-Zeitrahmen tritt in das bestehende 15m / 1h / 4h / 1d-Set in der Produktions-Deployment ein.
- **Health-Check-Kadenz.** Health-Check-Probes laufen in Produktion häufiger, um Mean-Time-to-Detection gegen die höhere Cycle-Rate beschränkt zu halten.

Die höhere Cron-Granularität ist einer der strukturellen Gründe, warum unsere projizierten Renditen über dem akademischen 15–30 % p.a. Korridor liegen (Abschnitt 1.3, 9.5): die Inferenz-Frequenz ist zwei Grössenordnungen höher als publizierte Designs.

0.18.9 A.9 Architectural Decision Records (ausgewählt)

Die folgenden ADRs sind im Engineering-Notebook des Projekts dokumentiert. Die investor-relevantesten:

ADR — Opus-Klasse-LLM für den Analyser. Wir wählten Anthropic Claude Opus 4-7 vor günstigeren Sonnet-Klasse-Alternativen, nach internen Vergleichs-Läufen auf historischen Kandidaten-Sets. Der Cost-Delta (etwa 4–5× pro Call) ist durch materiell besseres Multi-Signal-Reasoning in unseren beobachteten Outputs ausgeglichen. Das Framework ist modell-agnostisch und Substitution ist eine Prompt- und Schema-Adaption.

ADR — Partial-TP-mit-Runner. Wir adoptierten Partial-TP-mit-Runner (Abschnitt 8.2) gegenüber Voll-Exit-bei-TP1 oder Single-Target-TP2. Getrieben durch beobachtetes Verhalten in frühem Live-Trading: TP2 wird selten ohne vorhergehendes TP1 erreicht, und Voll-Exit-bei-TP1 gibt Upside in Trend-Regimes auf.

ADR — Strategy-Mode als globales DB-Setting. Operator-switchable ohne Code-Änderung, persistiert über Service-Restarts, unterstützt Strategy-Mode-Tagging auf Trades für A/B-Analyse. Per-Symbol-Modi wurden als zu komplex zu früh verworfen; Parallel-Shadow-Strategien auf Kosten (multiplizierte LLM-Inferenz).

ADR — TOTP-only Authentifizierung. Passwörter sind der dominante Breach-Vektor in Small-Team-Web-Anwendungen; sie zu eliminieren entfernt diese Fläche. TOTP-Secret wird einmal pro User generiert, als QR für Authenticator-App-Enrolment angezeigt, server-seitig gespeichert.

ADR — Interne Spot↔Futures-Transfers sind NIE Capital-Events. Sie zu zählen würde die Equity-Curve-Methodik durch Doppel-Zählung von Kapital korrumpieren, das lediglich Wallets gewechselt hat. Sie werden für Audit-Zwecke in `internal_transfers` gespeichert, nie in `capital_events`. (Siehe A.3.)

ADR — Original-SL-Preservation unter TP1-Trail. TP1-Trail aktualisiert den aktiven Stop für Execution-Zwecke, aber der originale Stop wird auf dem Trade-Record für Risiko-Attribution und Audit-Zwecke erhalten. (Siehe A.6.)

ADR — Regime-Exit-Mathematik auf Notional, nicht Margin. Der Fee-und-Funding-Threshold für Regime-Change-Exit wird auf Notional (`qty × Preis`) berechnet, was hebel-agnostisch ist. Eine Division durch Margin würde verschiedene Exit-Trigger bei verschiedenen Hebel-Niveaus für ökonomisch äquivalente Positionen produzieren — das ist falsch.

ADR — Coin Metrics Community API als Frühphasen-Exchange-Flow-Quelle. Glassnodes Free-Tier exponiert keine programmatische API; Coin Metrics Community ist keylos und liefert die Kern-Exchange-Inflow/Outflow-Metriken. Migrationspfad zu Glassnode Pro ist Teil der Post-MVP-Enhancement-Roadmap.

0.18.10 A.10 Aktueller Produktionsstatus

Stand Mai 2026, in Produktion:

- **Live-Trading auf Binance Spot (Long-only).** Operativ seit Januar 2026. Echte Fees, echte Slippage, echtes PnL.
- **Live-Trading auf Binance Futures (Long und Short).** Operativ. Bidirektionale Execution validiert.
- **Sechs On-Chain-Quellen** stündlich gesnapshottet: blockchain.com, mempool.space, bitnodes, DefiLlama, Coin Metrics Community, Etherscan.
- **5-State-Regime-Klassifizierer** auf 4-Stunden-Candle-Stream, pro Cycle persistiert.
- **Regime-Change-Exit**-Branch verdrahtet sowohl in Spot- als auch Futures-Execution-Pfade; loggt `exit_reason='regime_change'` beim Trigger.
- **Capital-Events-Sync** läuft auf jedem Cycle; Ein-, Auszahlungen und Spot↔Futures-Internal-Transfers rekonziliert mit idempotenter `external_id`.
- **Wallet-basierter Netto-PnL %** auf der Dashboard-Equity-Curve-View angezeigt, berechnet via geometrisches Period-Linking.
- **Internal-Transfer-Bridge** zur Chart-Render-Zeit appliziert; Phantom-Dips auf der Aggregat-Equity-Curve eliminiert.
- **Asset-Breakdown-Stacked-Area-Chart** mit Capital-Event-Markern auf der Portfolio-View sichtbar.
- **Dust-Threshold-Filter** bei $5 auf dem Portfolio-Breakdown.
- **Original-SL-Preservation** auf allen neuen Trades aktiv; historische Trades backfilled.
- **TP1-aware Position-Sizing** über Spot und Futures appliziert.
- **Dashboard** live mit TOTP-Authentifizierung; Private- und Public-Surface beide funktional.
- **Telegram-Briefings** zwei-mal täglich, Wochen-Macro-Report auf Schedule.
- **Health-Check-Probes** verifizieren geschedule-Job-Completion; Alert-on-Anomaly

via Telegram.
- **Cost-Envelope** konsistent mit Operating-Cost-Target auf MVP-Cron; Production-Cron-Envelope ist für die 1-Minuten-Kadenz der Vier-Investor-Cohort provisioniert.

0.18.11 A.11 Bekannte Limitationen und laufende Arbeit

Investor-bezogene Ehrlichkeit erfordert explizite Anerkennung von Limitationen:

- **Ethereum Smart-Contract-Custody-Untertracking.** Coinbase und Kraken nutzen Smart-Contract-Custody-Architekturen, deren Adressen nicht konsistent im Free-Tier-Etherscan-Label-Set gelabelt sind. Unsere aggregierte tracked-ETH-Zahl für diese Börsen ist ein Bruchteil des wahren Custody-Levels. *Relative* Bewegungen bleiben aussagekräftig; *absolute* Niveaus sollten als Untergrenzen behandelt werden. Resolutions-Pfad: Arkham-Intelligence-Label-Datenbank-Integration (kostenpflichtig, Post-MVP-Enhancement).
- **Slippage-Modell im Paper-Trader.** Der MVP-Paper-Trader simuliert keine Slippage; er bleibt eine Forschungs-Bench für Prompt- und Threshold-Revisionen, kein Performance-Validator. Das Live-System misst Slippage Trade-für-Trade; Live-Zahlen sind die einzigen Performance-Zahlen, gegen die wir berichten.
- **Regime-bedingte Execution-Verfeinerungen.** Die Regime-Klassifikation wird im Analyser-Prompt und im Regime-Change-Exit-Branch konsumiert. Zusätzliche regime-bedingte Verfeinerungen (regime-bedingtes Sizing, regime-bedingte TP-Ladders) sind auf dem Continuous-Improvement-Track.
- **Social-Sentiment-Integration.** Nicht vorhanden. Zukünftige Post-MVP-Enhancement; erfordert semantische-Deduplizierungs-Disziplin, um Echo-Chamber-Übergewichtung zu vermeiden.
- **Aktive-Adressen-7d-Trend.** Erfordert akkumulierte historische Snapshots, bevor er als Analyser-Prompt-Input nutzbar ist. Snapshot-Akkumulations-Pfad ist in Arbeit.
- **Single-Region-Deployment.** MVP läuft in einer AWS-Region. Produktions-Deployment beinhaltet Multi-Region-Berücksichtigung, besonders für den Ingest-Tier.
- **Telegram-Media-Deduplizierungs-Patch.** Eine kleine Workaround wird auf der OpenClaw-Extension-Schicht für ein Duplicate-Send-Verhalten gepflegt. Marker-und-Verifikations-Skript-Disziplin stellt sicher, dass er bei Upstream-Resolution entfernt werden kann.

0.18.12 A.12 Evidenz der Execution

Das Live-System ist für Investor-Demonstration an einer authentifizierten Domain zugänglich. Die Public-Surface ist ohne Credentials einsehbar. Die Private-Surface — einschliesslich voller Trade-Details, Equity-Curves mit Capital-Event-Markern, das LLM-Reasoning-Archiv und die Regime-Klassifikations-Persistenz — ist unter NDA für qualifizierte Investor-Repräsentanten verfügbar.

Was die Demonstration zeigt:

- Ein System, das kontinuierlich ohne Operator-Intervention läuft.
- Live-Spot- und Futures-Trades, die öffnen, auf TP1 skalieren und bei stop / TP1 / TP2 / regime_change / timeout / manual schliessen — über die erwarteten Exit-Reasons.
- Regime-Klassifikation auf MVP-Kadenz persistiert, sichtbar pro Symbol und pro Cycle.
- LLM-Analysen pro Kandidat generiert, mit Reasoning-Text, der Regime-, On-Chain- und technischen Kontext referenziert.
- Capital-Events (Ein-/Auszahlungen) und Internal-Transfers in die Equity-Curve-Sicht rekonziliert, mit der Bridge-Logik, die Phantom-Dips eliminiert.
- Wallet-basierter Netto-PnL % prominent auf der Equity-Curve-View angezeigt, berechnet via geometrisches Period-Linking.
- Dashboard-Operator-Flows (Settings-Änderung, Trade-Close, Capital-Events-Log, Macro-Report-Browse) end-to-end funktional.
- Telegram-Briefings auf Schedule eintreffend.
- Ein Daily-Cost-Envelope in Public-Telemetrie sichtbar, das das angegebene MVP-Target erfüllt.

Das ist kein Slideware-Prototyp. Es ist ein vollständiges, laufendes operatives System, auf dem Kai Zeh seit Januar 2026 sein eigenes Kapital handelt. Die Vier-Investor-Produktions-Cohort ist eine Onboarding-und-Scale-out-Übung gegen dieses Baseline, keine Build-out.
