Tribunal Session Summary

Session: conclave-20260522-040515-temporal-knowledge-graphs-for-e7f1

Depth: T6

Advocates: 5 | Judges: 5 | Cost: $2.4364 | Time: 34m 11s

THE PROMPT: Temporal Knowledge Graphs for Investment Edge How are investors using temporal knowledge graphs (TKGs) to generate alpha? Cover: Edge thesis — What specific informational or analytical advantages do TKGs provide over traditional approaches (factor models, NLP sentiment, static KGs)? Where does the edge come from: faster signal extraction, novel relationships, regime detection, event chains? Technical requirements — Graph schema design (entities, edges, time validity), storage choices (Neo4j, TigerGraph, custom), temporal reasoning patterns (bitemporal, event-sourced), embedding approaches (TGN, TGAT, TComplEx), and latency constraints for different strategies. Signal requirements — What makes a TKG-derived signal tradeable? Decay characteristics, capacity, orthogonality to known factors, robustness to graph noise. Data sources & ingestion — Filings, news, transcripts, supply chain data, patents, hiring data, alt-data. How to extract entities/relationships with temporal validity, resolve identities across sources, and handle revisions/corrections. From graph to trade — Pipelines for converting graph queries or embeddings into ranked signals, sizing, and execution. Backtesting pitfalls specific to temporal graphs (lookahead bias in entity resolution, survivorship in the graph itself). Concrete examples — Funds or research groups doing this in production, with citations where possible. Be specific and technical. Skip generic AI-in-finance commentary.

THE RULING: Temporal knowledge graphs (TKGs) provide a mechanistically distinct and plausible information architecture for investment edge through bitemporal validity control, invalidation of stale facts, and multi-hop temporal reasoning. However, there is currently no public, factor-neutralized, cost-adjusted evidence that TKGs generate persistent, tradable alpha in public equity markets. The edge remains a testable hypothesis, not a proven fact.

The Prompt

Temporal Knowledge Graphs for Investment Edge How are investors using temporal knowledge graphs (TKGs) to generate alpha? Cover: Edge thesis — What specific informational or analytical advantages do TKGs provide over traditional approaches (factor models, NLP sentiment, static KGs)? Where does the edge come from: faster signal extraction, novel relationships, regime detection, event chains? Technical requirements — Graph schema design (entities, edges, time validity), storage choices (Neo4j, TigerGraph, custom), temporal reasoning patterns (bitemporal, event-sourced), embedding approaches (TGN, TGAT, TComplEx), and latency constraints for different strategies. Signal requirements — What makes a TKG-derived signal tradeable? Decay characteristics, capacity, orthogonality to known factors, robustness to graph noise. Data sources & ingestion — Filings, news, transcripts, supply chain data, patents, hiring data, alt-data. How to extract entities/relationships with temporal validity, resolve identities across sources, and handle revisions/corrections. From graph to trade — Pipelines for converting graph queries or embeddings into ranked signals, sizing, and execution. Backtesting pitfalls specific to temporal graphs (lookahead bias in entity resolution, survivorship in the graph itself). Concrete examples — Funds or research groups doing this in production, with citations where possible. Be specific and technical. Skip generic AI-in-finance commentary.

The Ruling

Temporal knowledge graphs (TKGs) provide a mechanistically distinct and plausible information architecture for investment edge through bitemporal validity control, invalidation of stale facts, and multi-hop temporal reasoning. However, there is currently no public, factor-neutralized, cost-adjusted evidence that TKGs generate persistent, tradable alpha in public equity markets. The edge remains a testable hypothesis, not a proven fact.

The strongest evidence is the architectural plausibility of TKGs to prevent lookahead bias and reduce noise via bitemporal modeling and invalidation agents, as demonstrated in the OpenAI Developers' "Temporal Agents with Knowledge Graphs" cookbook. The main caveat is that no advocate or public source has produced a verified backtest showing that these mechanisms translate into statistically significant, factor-orthogonal returns after execution costs. The absence of empirical validation means the edge is not yet actionable.

Core Analysis

Edge Thesis: Where the Advantage Lies — and Where It Doesn't

The primary informational advantage of temporal knowledge graphs (TKGs) over traditional approaches—such as static knowledge graphs, NLP sentiment models, or factor-based systems—stems from their **explicit modeling of time as a first-class dimension**, enabling three core capabilities:

1. **Bitemporal Leakage Control**: TKGs separate *valid time* (when a fact was true in the real world) from *transaction time* (when it was recorded in the system). This distinction, grounded in textbook temporal database theory and implemented in the OpenAI Developers' "Temporal Agents" cookbook [1], prevents lookahead bias during backtesting. For example, a merger announced on 2023-06-01 but ingested into the graph on 2023-06-05 can be correctly timestamped so that no signal generated before 2023-06-01 reflects that relationship. This is a structural hygiene improvement over static KGs or batch-updated NLP pipelines, which often conflate ingestion and event times.
2. **Invalidation-Based Noise Reduction**: The OpenAI framework introduces an **Invalidation Agent** that marks outdated or superseded facts with `t_invalid` and `invalidated_by` fields. This allows the system to prune stale edges—such as outdated supplier relationships after a contract termination—before they pollute downstream signals. This capability directly addresses a major weakness in static KGs, where outdated relationships persist unless manually cleaned. While not a source of alpha *per se*, this mechanism plausibly improves signal-to-noise ratio and reduces false positives, particularly in medium-frequency strategies (holding periods of days to weeks).
3. **Multi-Hop Temporal Inference and Event-Chain Reasoning**: TKGs enable causal chain propagation across time-conditioned relationships. For instance, a factory fire at Supplier A on T+0 may disrupt deliveries to Manufacturer B on T+2, which in turn delays shipments to Retailer C on T+5, affecting earnings calls on T+7. A TKG can model this chain with precise temporal validity windows and infer downstream impacts before they are priced in. This capability, supported by the OpenAI cookbook's focus on "timeline generation" and "change tracking" [1], allows for **forward-looking, time-aware inference** that NLP sentiment models (which typically score isolated documents) or static network models cannot replicate.

However, **no advocate has demonstrated that these architectural advantages translate into factor-orthogonal, cost-adjusted alpha**. Preprint results such as FinDKG's 15% MRR uplift in link prediction are promising but do not equate to trading performance. Similarly, the existence of supplier-customer spillovers (Cohen & Frazzini, 2008) validates an economic mechanism but does not prove that TKGs add incremental value beyond traditional event studies or regression models.

The edge, therefore, is **conditional**: it exists *if* the noise reduction and multi-hop inference capabilities of TKGs produce signals with higher information coefficient (IC) than alternatives, *after* controlling for known factors and execution costs. Until such evidence is produced, the edge remains a **plausible R&D hypothesis**, not a live trading advantage.

Technical Requirements: Building a Production-Grade TKG

To realize the potential advantages of TKGs, the following technical components are necessary:

- **Graph Schema Design**: Entities include companies, executives, products, filings, news events, and patents. Edges represent relationships such as "supplies," "competes with," "acquired," or "filed 8-K." Each edge must carry temporal metadata:
  - `valid_at`: when the fact became true (e.g., contract start date)
  - `invalid_at`: when it ceased to be true (e.g., contract end)
  - `t_created`: when the fact was added to the graph
  - `t_expired` / `invalidated_by`: for automatic invalidation of superseded facts This schema, directly inspired by the OpenAI cookbook [1], supports both **Atemporal** (e.g., "Apple is in tech sector"), **Static** (e.g., "Tesla founded 2003"), and **Dynamic** (e.g., "AMD supplies GPUs to Microsoft") facts with appropriate lifecycle handling.

- **Storage and Query Layer**: Neo4j and TigerGraph both support property graphs with temporal attributes, but neither natively enforces bitemporal constraints. Custom validation logic or middleware (e.g., using Apache Kafka for event sourcing) is required to ensure temporal integrity. For high-throughput use cases, a hybrid architecture—using a temporal database (e.g., PostgreSQL with temporal extensions) for auditability and a graph engine for traversal—is recommended.
- **Temporal Reasoning Patterns**: The majority of judges endorse **discrete time versioning** over continuous-time modeling, following Senzing/GPH's recommendation [3] for maintainability and queryability. Allen's interval algebra can be used to reason over overlapping or sequential events (e.g., "Did the CEO resignation overlap with the product recall?"). Event-sourced architectures, where every change is logged as an immutable event, are ideal for point-in-time (PTI) reconstruction during backtesting.
- **Embedding Approaches**: Temporal Graph Networks (TGN) [Rossi et al., ICML 2020], Temporal Graph Attention Networks (TGAT) [Xu et al., ICLR 2020], and TComplEx [Lacroix et al., ICML 2020] are peer-reviewed models capable of learning time-evolving representations. These are particularly useful for link prediction (e.g., "Which company will acquire another in the next 90 days?") and anomaly detection. However, their real-world latency in full-scale financial graphs is likely in the **minutes**, not sub-second, due to neighbor sampling and caching constraints—making them suitable for daily or intraday, but not HFT, strategies.
- **Latency Constraints**: For medium-frequency strategies (holding periods of 1–10 days), end-to-end pipeline latency of minutes is acceptable. The OpenAI cookbook's pipeline—semantic chunking, statement extraction, temporal grounding, triplet creation, entity resolution—can be batched hourly or daily. For faster signals, incremental updates using streaming frameworks (e.g., Flink) are required, but introduce complexity in maintaining temporal consistency.

Signal Requirements: From Graph to Trade

A TKG-derived signal becomes tradeable only if it satisfies four criteria:

1. **Decay Characteristics**: Signals must decay slowly enough to allow execution but fast enough to capture transient mispricing. A half-life of 1–10 trading days is plausible, especially for multi-hop event chains (e.g., supply chain disruptions). This contrasts with NLP sentiment, which often decays in minutes, and factor exposures, which persist for months.
2. **Capacity**: The edge is likely limited to mid- and small-cap equities. Large-cap names are too well-covered for TKGs to offer novel insights, and any signal would be quickly arbitraged. The realistic capacity for a TKG strategy is likely in the $100–200M AUM range, depending on data exclusivity and execution speed.
3. **Orthogonality**: To qualify as alpha, a TKG signal must not be subsumed by existing factors (market, value, momentum, etc.) or alternative NLP/sentiment models. The only way to verify this is through factor-neutral backtesting with point-in-time data. No such test has been publicly conducted.
4. **Robustness to Graph Noise**: TKGs are vulnerable to noise from incorrect entity resolution (e.g., merging two companies with similar names) or false triplet extraction (e.g., "Apple supplies Tesla" due to ambiguous text). The use of **Invalidation Agents** and **entity resolution with fuzzy matching + ML linking** (as in the Towards AI example [2]) helps, but does not eliminate the risk. Robustness must be tested via ablation studies (e.g., removing invalidation to measure its impact on Sharpe).

Data Sources & Ingestion: From Raw Text to Temporal Triplets

TKGs are built on diverse, time-stamped data:

- **Filings (10-K, 8-K, S-1)**: Contain structured disclosures of relationships, risks, and events with clear effective dates.
- **News and Transcripts**: Unstructured but rich in forward-looking statements and event mentions. Temporal grounding (e.g., “next quarter,” “2024 launch”) must be resolved to absolute dates.
- **Supply Chain Data**: From vendors like Panjiva or Orbis, often with contract durations.
- **Patents, Hiring, Alt-Data**: Indicate strategic shifts (e.g., hiring AI talent → new product line).

The ingestion pipeline, as detailed in the OpenAI cookbook [1], involves:

1. **Semantic Chunking**: Break documents into coherent segments.
2. **Statement Extraction**: Use LLMs to extract subject-predicate-object triples.
3. **Temporal Grounding**: Resolve relative time expressions to absolute `valid_at`/`invalid_at` ranges.
4. **Entity Resolution**: Deduplicate entities across sources using fuzzy matching and ML (as in Towards AI [2]).
5. **Temporal Validity Enforcement**: Apply rules to ensure no future-dated facts leak into past queries.

Critical risk: **entity resolution lookahead**. If an ER model uses post-2023 data to merge “Twitter” and “X Corp,” it may incorrectly apply that link in pre-2023 backtests. The solution is to run ER incrementally, using only data available at each point in time.

From Graph to Trade: Pipeline Design and Backtesting Pitfalls

A typical TKG-to-trade pipeline:

1. **Query or Embed**: Run temporal graph queries (e.g., “Find all companies two hops from a recent factory fire”) or generate embeddings via TGN/TGAT.
2. **Score and Rank**: Assign confidence scores to predicted events (e.g., earnings risk, M&A likelihood).
3. **Factor Neutralize**: Regress out exposures to Fama-French factors and sector dummies.
4. **Size and Execute**: Use volatility targeting or risk parity for position sizing; execute via VWAP/TWAP to minimize market impact.

Backtesting Pitfalls:

- **Lookahead in Entity Resolution**: As above, ER must be run incrementally.
- **Survivorship Bias in the Graph**: Acquired or bankrupt companies must be retained in historical states.
- **Invalidation Timing**: If a fact is invalidated *after* it was used in a signal, the backtest must reflect that delay.
- **Point-in-Time Controls**: All data (filings, news, prices) must be aligned to transaction time, not ingestion time.

GPT-5’s proposed backtest—measuring incremental edge from bitemporal invalidation on supply-chain shock propagation (2019–2024)—is the most rigorous path to validation. It isolates a single mechanism and can be pre-registered.

Concrete Examples: Who Is Doing This?

While no fund has publicly disclosed a TKG-based alpha strategy with audited performance, several are building the infrastructure:

- **OpenAI's "Temporal Agents"** [1] is a production-grade blueprint used internally and by early adopters.
- **Bloomberg KG, Refinitiv PermID, S&P Kensho KG** are commercial temporal KGs used by quant funds for entity resolution and relationship mapping.
- **RavenPack** integrates event-based KGs with temporal anchoring for news analytics.
- **GDELT** provides a global event database with temporal scope, used in academic research.

The **Towards AI** article [2] describes a VC-focused temporal investment KG that tracks funding rounds and investor relationships over time—demonstrating the architecture's feasibility, though not its public-market applicability.

No public evidence confirms that J.P. Morgan, Man Numeric, or Interactive Brokers' Reflexivity KG generate factor-orthogonal alpha from TKGs. These claims remain unverified.

How We Got Here

Scorecard

Judge	DeepSeek V4 Pro	Grok 4	Magistral Medium	Gemini 3.1 Pro	GPT-5	Winner
Qwen 3 235B (Cerebras)	PARTIAL	PARTIAL	REJECT	PARTIAL	ADOPT	GPT-5
DeepCogito v2.1 671B	ADOPT	ADOPT	REJECT	REJECT	ADOPT	GPT-5
MiniMax M2.7	PARTIAL	PARTIAL	REJECT	PARTIAL	ADOPT	GPT-5
Mistral Large 3	PARTIAL	PARTIAL	REJECT	PARTIAL	ADOPT	GPT-5
GPT-OSS 120B	REJECT	ADOPT	REJECT	ADOPT	ADOPT	GPT-5

Convergence Assessment

The convergence was epistemic, driven by verifiable corrections. Position shifts occurred when advocates confronted specific factual errors: DeepSeek V4 Pro retracted fabricated citations, Gemini 3.1 Pro abandoned the untested ACA exclusivity claim, and Grok 4 downgraded FinDKG's MRR uplift to "unproven potential." Magistral Medium was the sole exception, maintaining unsubstantiated claims of "significant edge" without metrics, suggesting affective convergence due to social pressure.

Closing Statements

DeepSeek V4 Pro In their closing statement, DeepSeek V4 Pro refined their position from claiming a "structural edge" to acknowledging that TKGs are a "promising, mechanistically distinct information architecture" whose alpha potential remains unproven. They retracted all specific performance claims (J.P. Morgan, Man Numeric) and emphasized that the edge is conditional on rigorous, factor-neutralized backtesting. The most challenging argument they faced was the lack of empirical validation, which forced them to abandon the notion of demonstrated alpha. Their strongest surviving evidence was the OpenAI cookbook's bitemporal validity framework and invalidation agent pattern, which they argued provides a necessary foundation for leakage control.

Grok 4 Grok 4 maintained that TKGs provide mechanistic advantages through temporal validity modeling and multi-hop event-chain reasoning. However, they conceded that FinDKG's 15% MRR uplift and ETF outperformance lacked factor neutralization and cost modeling, downgrading the claims to "promising, unquantified potential." The most challenging argument was the conflation of link-prediction metrics with tradable alpha, which they acknowledged as a valid critique. Their strongest surviving evidence was the OpenAI Temporal Agents cookbook and the Reflexivity KG deployment, which they cited as proof of architectural feasibility.

Magistral Medium Magistral Medium held that TKGs provide a "significant edge" through novel relationships, regime detection, and event-chain analysis. Despite repeated challenges, they did not produce any metrics or backtestable signal definitions. They conceded that sources do not provide direct alpha evidence but maintained their core claim. The most challenging argument was the demand for empirical proof, which they could not satisfy. Their strongest surviving evidence was the OpenAI cookbook and Towards AI article, which they used to support the architectural capability of TKGs.

Gemini 3.1 Pro Gemini 3.1 Pro revised their position from claiming exclusive alpha through Asynchronous Causal Arbitrage to acknowledging that TKGs are a "superior architectural foundation" for data hygiene and structural modeling. They conceded that ACA was untested and that bitemporal control and invalidation are legitimate, complementary edges. The most challenging argument was the lack of empirical proof for ACA, which they fully acknowledged. Their strongest surviving evidence was the OpenAI cookbook's invalidation agent and bitemporal schema, which they argued provide a cleaner representation of reality.

GPT-5 GPT-5 maintained that TKGs are necessary infrastructure for medium-frequency event-chain signals, enabling bitemporal leakage control and invalidation. They retracted claims of proven alpha and committed to a falsifiable backtest specification. The most challenging argument was the need for incremental value over chronological baselines, which they addressed by proposing a controlled A/B test. Their strongest surviving evidence was the OpenAI cookbook, Senzing/GPH guidance, and peer-reviewed GNNs, which they argued provide a robust foundation for TKG implementation.

Reality Checks

Claim	Challenge	Result
DeepSeek V4 Pro: "J.P. Morgan AI Research published a temporal knowledge graph-based framework for stock movement prediction with IC 0.07 and Sharpe 1.8 before costs"	**Grok 4:** "The cited arXiv:2203.12345 does not exist; no such preprint matches the title or results"	Full concession and retraction; claim acknowledged as fabricated
Gemini 3.1 Pro: "Continuous time modeling in TKGs causes system collapse; discrete time versioning is required for execution"	**Grok 4:** "TGN and TGAT are peer-reviewed continuous-time models used in production; Senzing recommends discrete time for maintainability, not collapse"	Full concession; claim retracted and revised to acknowledge continuous models are viable
Grok 4: "FinDKG + KGTransformer's AI-themed portfolio outperforms existing thematic ETFs on annualized returns and Sharpe ratio"	**DeepSeek V4 Pro:** "The paper does not report factor-neutralized returns or control for style exposures; outperformance could be due to tech-momentum beta"	Full concession; claim downgraded to "preliminary preprint results" without proven alpha

Next Steps

Next Step	Why It Matters	Confidence
Conduct a controlled, point-in-time backtest of bitemporal invalidation on supply-chain shock propagation (2019–2024) with factor neutralization and cost modeling	This isolates the incremental value of TKGs over chronological baselines and provides the first empirical evidence of edge	65%
Test whether multi-hop temporal link prediction generates factor-orthogonal IC > 0.02 on 2019–2024 data	This validates the primary signal-generation claim of TKGs	50%
Quantify the capacity limits of TKG signals in mid/small-cap equities after accounting for crowding and execution costs	This determines scalability and justifies infrastructure investment	70%

Dissenting Opinions

No dissenting opinions were filed in this session.

Glossary

Term	Definition
Atemporal	A fact that is always true and does not change over time, such as a company's headquarters location.
Bitemporal	A modeling approach that tracks both valid time (when a fact was true in the real world) and transaction time (when it was recorded in the system).
Dynamic	A fact that changes over time and has a defined validity period, such as a CEO's tenure.
Entity Resolution	The process of identifying and linking different records that refer to the same real-world entity across multiple data sources.
Invalidated_by	A field in a temporal knowledge graph that points to a later fact that supersedes or corrects an earlier one.
Lookahead Bias	A backtesting error where future information is inadvertently used to make decisions in the past.
Static	A fact that is true for a defined period but does not change frequently, such as a company's founding date.
Temporal Validity Range	A time interval during which a fact is considered valid in a temporal knowledge graph.
Transaction Time	The time when a fact was recorded or ingested into a knowledge graph.
Valid Time	The time when a fact was true in the real world.

How The Conclave Works

The Conclave is a structured multi-model deliberation system. Instead of asking one AI model a question and trusting its answer, the Conclave convenes a panel of independent models, forces them to argue, and subjects their conclusions to judicial review. The process is adversarial by design — consensus must be earned through evidence, not assumed through agreement.

Advocates are independent AI models (e.g., Claude, GPT-5, Gemini, DeepSeek) that each receive the same question and produce a sealed submission without seeing each other's work. Each submission must include a falsifiable hypothesis, tagged evidence with reasoning type (deductive/inductive/abductive), provenance labels for any frameworks cited, and an honest self-assessment.

Challenges follow submissions. Each advocate reads all other submissions and directly attacks weak points — not polite reviews, but pointed cross-examination. They must identify the single factual crux that would settle each disagreement.

Debate Rounds (1–7 depending on depth) force advocates to defend, concede, or revise their positions under pressure. The system tracks position stability across rounds to detect sycophantic drift — when models change their stance to agree with others without citing new evidence.

Judges (called Justices on The Bench) are separate models that never participated as advocates. They evaluate the full record: submissions, challenges, and debate transcripts. They fact-check claims, audit framework provenance, assess whether convergence was epistemic (evidence-driven) or affective (social pressure), and render a verdict: ACCEPT one position, SYNTHESIZE the best elements, or REMAND for further debate.

Depth levels control rigor: T1/Spot Check (submissions only), T2/Standard Review (1 debate round, 1 judge), T3/Deep Review (3 rounds, 3 judges), T4/Full Panel (5 rounds, full panel), T5/Stress Test (5 rounds + Fresh Eyes review), T6/Red Team (7 rounds + mid-debate judicial checkpoint).

The system is deterministic code — it cannot be sycophantic. It dispatches prompts, collects responses, and enforces the adversarial structure. The models argue; the code referees.