2025 年 9 月 3 日　全 13 頁

KDD 2025（AI 国際会議）出張報告：複数 AI の協働と専門ツール統合が新潮流に

「AI エージェント」「時系列分析」「信頼できる AI」に注目

経済調査部　主任研究員　新田 尭之

［要約］

- 2025 年 8 月 3 日から 7 日にかけて、データマイニング分野のトップ国際会議である KDD 2025（AI 国際会議）へと出張する機会を得た。本稿では、数多くのプログラムの中から「AI エージェント」「時系列分析」「信頼できる AI」に焦点を絞り、金融分野やヘルスケア分野への応用を含む論文や講演などを紹介しつつ、得られた示唆を報告する。

- 全体を通じて、AI 開発の焦点が、単一モデルの性能向上から、複数の専門 AI エージェントや AI 以外のツールを組み合わせる「アーキテクチャ設計」へと移行しつつあることが示された。その象徴が、自律的に思考・協働する「AI エージェント」の進化である。金融分野では、専門家役の AI エージェントとの対話を通じて組織の暗黙知を学習する財務アナリスト AI エージェントや、金融市場の変化に適応し陳腐化しない投資戦略を自律的に探求する投資戦略 AI エージェントが報告された。

- 時系列分析の領域では、大規模言語モデル（LLM）を直接の予測ツールとせず、専門的な分析ツール群を的確に使いこなす「司令塔」として活用するハイブリッドアプローチが新たな潮流となった。さらに、仮想ニュースを生成して金融市場の因果関係を学習させるなど、単なる相関分析を超えた、より本質的な理解への挑戦が始まっている。

- さらに AI の社会実装が本格化する中、「信頼できる AI」の確保が課題として議論された。AI エージェントの脆弱性を体系的に分析するフレームワークや、AI 自身が AI 攻撃への防御策を構築する免疫システム、医療現場における AI の潜在的バイアスを監査する手法など、安全性や公平性を担保するための具体的な技術が数多く提示された。

- これらの動向は、今後の AI 活用の成否が、課題やドメイン知識に基づき、最適な技術要素を組み合わせる「アーキテクト」としての設計能力にかかっていることを強く示唆する。日本企業においても、この設計思想に基づいた戦略的な人的資本投資と組織能力の構築が急務であろう。

１．はじめに

2025 年 8 月 3 日から 7 日にかけて、データマイニング分野のトップ国際会議である The 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining（以下 KDD 2025）へ出張する機会を得た。開催地はカナダ・トロントであった。

KDD 2025 はその名が示す通り、純粋な学術的探求だけでなく、実社会の課題解決をも重視する傾向にある。今年も大学に所属する研究者からの発表のほか、米中の大手テック企業をはじめとした民間企業からの発表も活発だった。ただし、ビザの関係で一部の中国本土からの発表予定者がカナダに渡航できず、欠席あるいは代理人がプレゼンテーションを行うケースが散見された。

現地では、「研究」、「応用データサイエンス」、「データセットとベンチマーク」の各トラックの論文発表に加え、基調講演やワークショップ、チュートリアル、KDD Cup（データ分析競技会）[1]など多様なプログラムが存在した。

本稿では、数多くのプログラムの中から、今後の技術・ビジネス動向を占う上で特に重要だと考えられる３つのトピック、すなわち、「AI エージェント」「時系列分析」「信頼できる AI（Trustworthy AI）」に焦点を絞る。これらのトピックに関して、金融分野やヘルスケア分野への応用を含む論文や講演などを紹介しつつ、得られた示唆を報告する。

本稿では、各研究のモチベーションや中心的なアイデア、得られた示唆などをなるべく平易に説明する。技術の詳細な説明が必要な場合は、参考文献に掲載した原論文などを参照されたい。

図表１　KDD 2025 のエントランス・モニュメント

（出所）大和総研撮影

[1] KDD Cup のタスクは Meta Platforms による「CRAG-MM（Comprehensive RAG for Multi-Modal, Multi-Turn）」であった。これはマルチモーダル RAG（Retrieval-Augmented Generation、検索拡張生成）の実力を、Ray-Ban Meta スマートグラス由来を含む約 5,000 枚の画像と 4 種類の設問（単純認識／単純知識・マルチホップ・比較／集約・推論）で評価する設定であった。詳細は以下ウェブサイト参照。
（https://www.aicrowd.com/challenges/meta-crag-mm-challenge-2025）

２．AI エージェント：自律性の向上と専門領域へ応用が進展

専門家との対話で自己改善する財務アナリスト AI エージェント

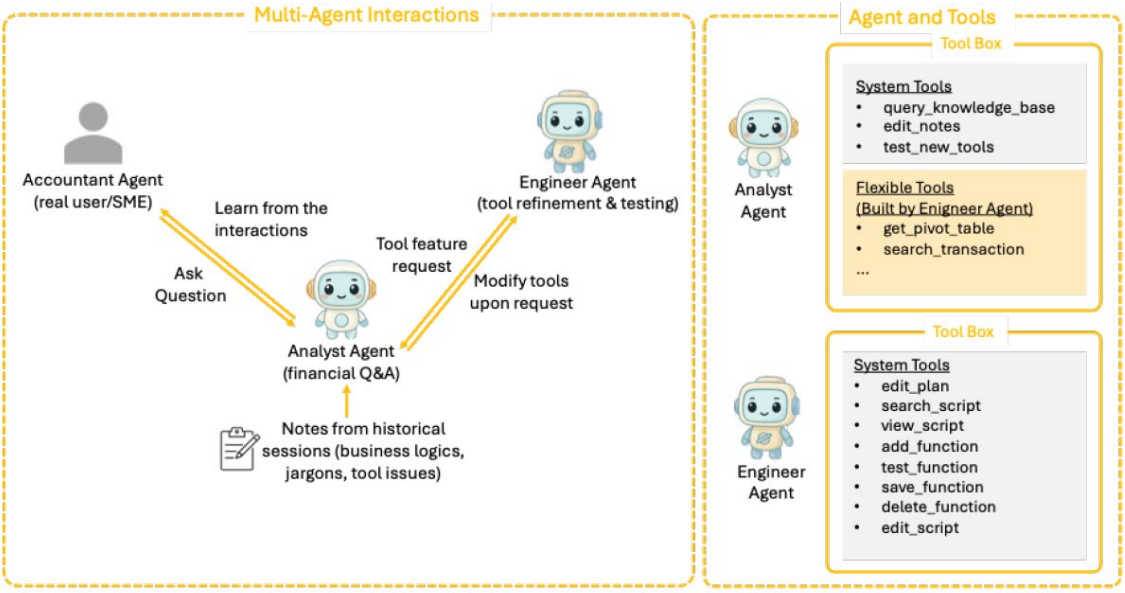
今回の学会で最も熱気を帯びていたテーマの一つが、自律的に思考し、複雑なタスクをこなす「AI エージェント」であった。企業の財務・会計部門を題材にしたのが、Amazon に所属する研究者らによる“Building Analyst-Like Agents: A Self-Improving Multi-Agent Framework for Financial Reasoning in the Enterprise” (Zhang et al., 2025a)である。

従来の AI は、整理されたデータに基づいて答えを出すのは得意だが、文書化されていない暗黙知を読み解くのは苦手であった。特に、企業の財務データは、長年の事業活動や M&A によってシステムが複雑化し、その分析に必要な専門知識はマニュアル化されず、「暗黙知」となっていることが少なくない。

そこで本研究は、AI を「成長する新人財務アナリスト」として捉え直す。提案するフレームワークでは、新人役の「財務アナリストエージェント」が分析を行い、その回答を専門家である「会計士エージェント」が評価・修正し、文書化されていない知識をフィードバックする。一方、財務アナリストエージェントからの要求に基づき、「エンジニアエージェント」はツールのコーディングからテスト、検証までを担当し、分析基盤そのものを向上させる。

この現実の企業さながらの対話と経験を通じて、財務アナリストエージェントは専門知識を蓄積し、状況に応じた対応力を身に付けていく。このように学習と連携を重ねることで、AI の正答率は劇的に向上したという。これは、AI を導入して終わりではなく、組織の一員として共に成長させていく未来の協働スタイルを示唆している。

図表２　財務アナリスト役、会計士役、エンジニア役の各 AI エージェントが協働する仕組み



（出所）Zhang et al. (2025a)より引用

株式市場の必勝パターンの陳腐化に挑む投資戦略 AI エージェント

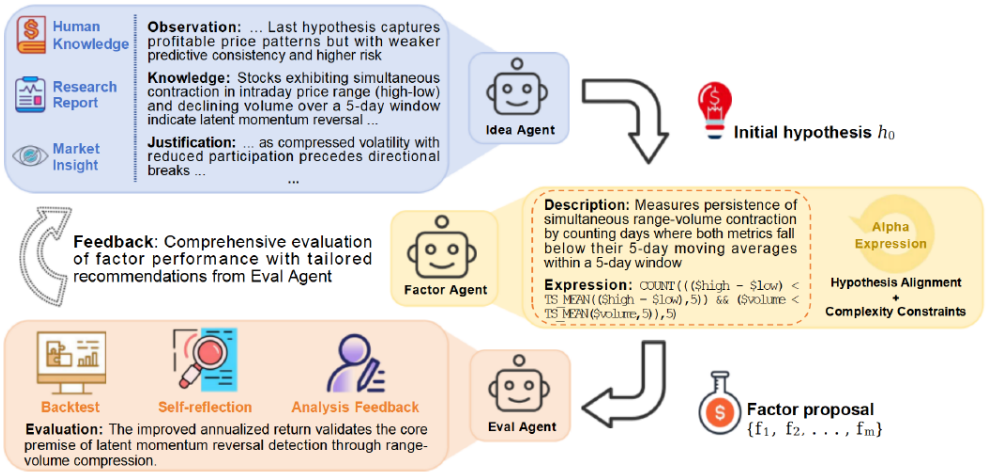
次に、金融市場を舞台にしたのが、中国・中山大学の研究者らによる“AlphaAgent: LLM-Driven Alpha Mining with Regularized Exploration to Counteract Alpha Decay” (Tang et al., 2025)である。

AI は過去のデータから短期的な「必勝パターン」を見つけるのは得意だが、それが未来も続くとは限らない。この理由として、株式市場でベンチマークを上回るリターン（アルファ）が得られるような投資戦略は、一度見つかってもいずれ真似されて優位性を失ってしまうことが指摘できる。

AlphaAgent はこの問題に対し、①アイデアを出す AI エージェント、②投資戦略を具体化する AI エージェント、③出来栄えを評価する AI エージェントがチームを組み、「独自性[2]・シンプルさ[3]・仮説との整合性[4]」という共通ルールを守りながら、繰り返し戦略を磨き上げていく。

検証では、2021 年から 2024 年にかけて、AlphaAgent は中国 CSI500 で年率 11.0%、米国 S&P500 で年率 8.74%の超過リターンを達成した。これは、比較対象の次点モデルが記録した 4.96%（CSI500：LSTM）と 2.75%（S&P500：DeepSeek-R1）をそれぞれ大幅に上回った。一度見つけてもすぐに真似されて陳腐化してしまう、という課題に対し、本質的な優位性を持つ投資戦略を AI エージェントが見出すアプローチは、実務でも十分に応用可能だと見込まれよう。

図表３　アイデア出し、投資戦略の具体化、評価担当の各 AI エージェントが協働する仕組み



（出所）Tang et al. (2025)より引用

[2] 独自性（Originality Enforcement）：新しい投資戦略を数式レベルで分析し、既存の有名な投資戦略との類似度を計算する。これにより、他の投資家が群がる「混雑した」戦略ではなく、まだ見出されていないユニークなアプローチの探索を促す。

[3] シンプルさ（Complexity Control）：戦略の数式が過度に複雑にならないよう、数式の構造やパラメータの数に制約をかける。これにより、過去のデータにだけ過剰適合した状態に陥ることを防ぎ、より普遍的で頑健な戦略の発見を促す。

[4] 仮説との整合性（Hypothesis Alignment）：まずファイナンス理論や過去のデータに基づき、「市場はこう動くはずだ」という仮説を立て、AI がその仮説に沿って戦略を構築する。そして、生成された戦略が元の仮説の意図を正しく反映しているか、再度 AI 自身が評価することで、ファイナンス理論や市場の経験則などの裏付けのない偶然の産物が生まれるのを防ぐ。

３．時系列分析：大規模言語モデルとの融合と「因果」への挑戦

時系列分析のフロンティア：大規模言語モデルを「司令塔」として分析ツールを活用

KDD 2025 では、自然言語処理や画像認識の分野で革命を起こした「基盤モデル（Foundation Model）」の考え方を、株価や需要予測といった時系列分析の領域にいかに応用するかが、一大テーマとなっていた。しかし、そのアプローチは大規模言語モデル(Large Language Model, LLM)を万能の予測ツールとして使うという単純なものではなく、その推論能力を活かしてより高度な分析を実現しようとするものであり、新たな潮流が生まれていた。

まず、時系列分析のワークショップ“The 11th Mining and Learning from Time Series Workshop: From Classical Methods to LLMs”では、米国・南カリフォルニア大学の Yan Liu 教授により“Frontiers of Foundation Models for Time Series”と題した講演に注目した。この講演では、大規模言語モデルを株価のような時系列データに直接適用する際の根本的な課題が指摘された。大規模言語モデルは言葉の意味の近さ（例：「猫」と「子猫」）を理解するのは得意だが、数値の近さ（例：「100.1」と「100.2」）を捉えるのは苦手である。数値を一度バラバラに分解すると、トレンドや季節性といった連続的なデータの機微な特徴が失われてしまうためである。

この課題に対し、講演で示された解決策は、大規模言語モデルを、自ら計算するプレイヤーではなく、専門的な分析ツール（統計モデルや API）を的確に呼び出してタスクを遂行させる賢い「司令塔」として使う、ハイブリッド AI エージェントのアプローチであった。この考え方は、今後の時系列分析における大規模言語モデルの役割を再定義するものであり、非常に示唆に富んでいた。

さらに、今後のフロンティアとして、①高品質で多様なデータの不足、②予測可能な「秩序」と予測不能な「カオス」の均衡、そして③単なる相関を超えた「因果」の理解という３つの大きな挑戦が提示された。

銘柄・業種などの特徴を理解する事前学習で株価予測 AI の精度が向上

金融分野に関しては、英国・エディンバラ大学の研究者らは、“Pre-training Time Series Models with Stock Data Customization” (Wang et al., 2025)の中で株価予測 AI の新たな学習方法を提示した。

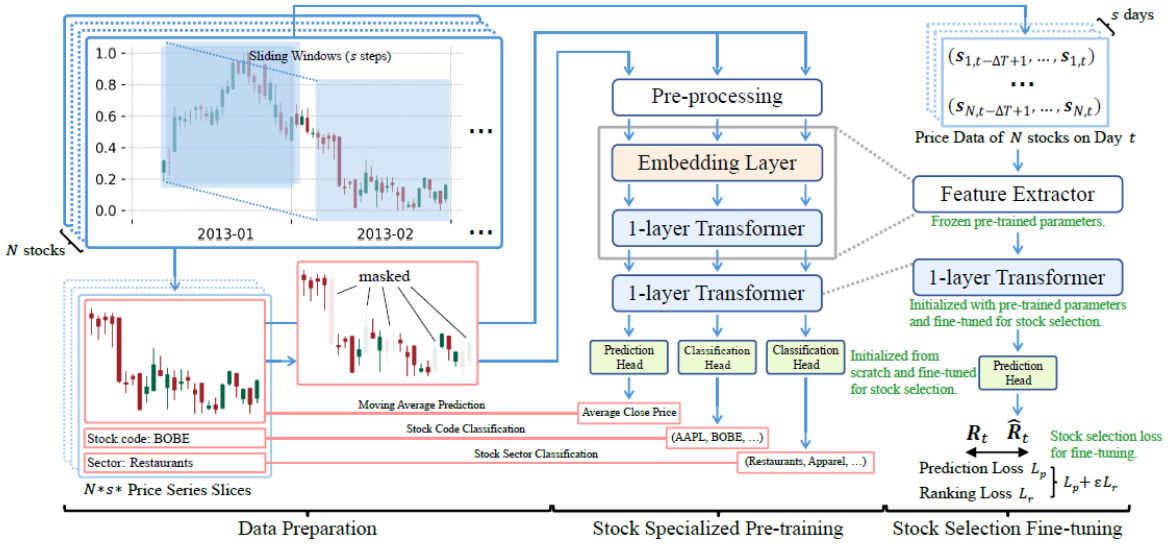
従来の AI は、過去の価格データからパターンを学ぶのは得意だが、株価特有の変動や、銘柄ごとの値動きの「個性」までは捉えきれていなかった。特に、企業の事業内容や市場での立ち位置によって生まれる価格変動の微妙なクセは、予測モデルを構築する上で見過ごされがちな情報だった。

そこでこの研究では、AI は本格的な予測訓練の前に、まず３つのユニークな「事前学習」を実施する。具体的には、①短い値動きのグラフだけを見てどの会社の株価なのかを当てる「銘柄

分類(Stock Code Classification)」、②どの業界に属するのかを当てる「業種分類(Stock Sector Classification)」、③データの一部を隠してその期間の平均価格を予測させる平均価格の予測(Moving Average Prediction)、である。

この訓練を通じて、AI は銘柄や業界ごとの特徴を深く理解し、データの背後にある本質的なパターンを見抜く能力を身に付けていく。この事前学習を経た AI は、実際の投資シミュレーションにおいて、市場平均や既存の AI モデルを上回る高いパフォーマンスを達成したという。これは、AI に単なる予測をさせるだけでなく、データへの深い洞察力を与える「教育」こそが、予測精度を飛躍させる鍵であることを示唆している。

図表４　銘柄や業種、平均価格を事前学習した株価予想 AI の概要



（出所）Wang et al. (2025)より引用

仮想ニュース生成で金融市場の因果関係を学習

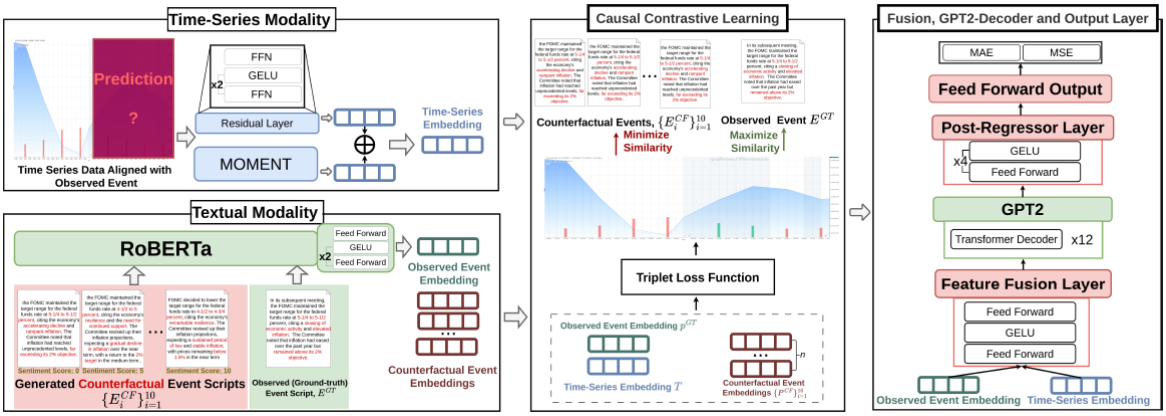
さらに、Liu 教授が未来の課題として挙げた「因果」の探求に金融分野で正面から取り組んだのが、中国・西南財経大学の研究者などによる論文“CAMEF: Causal-Augmented Multi-Modality Event-Driven Financial Forecasting” (Zhang et al., 2025b)である。

金融市場では、「マクロ経済イベント（原因）が市場の動き（結果）を引き起こす」という因果関係が重要だが、従来のモデルではこの関係性を捉えきれていなかった。

そこで本研究では、大規模言語モデルに「もしも」の世界をシミュレーションさせるという画期的なアプローチを採用した。例えば、「もし失業率の発表が、実際よりも良い内容だったら？」などの仮想のニュース記事を大規模言語モデルに多数生成させる。そして、実際の市場の反応に対して、正しいニュース記事と仮想のニュース記事を見分けるようモデルを訓練させる。具体的には、本物のニュース記事と実際の市場の動きとの関連性が、どの仮想のニュース記事との関連性よりも強くなるようにモデルを訓練することで、イベントと市場反応の間の本質的な因果関係を学習させる。この因果学習メカニズムと、テキスト・時系列データを融合したマルチ

モーダルなアプローチにより、他の最先端モデルを超える株価の予測精度を達成したという。

図表５　生成 AI による仮想のニュース記事を踏まえ、因果関係を考慮した株価予測 AI の概要



（出所）Zhang et al.（2025b）より引用

４．信頼できる AI：安全性・公平性・論理性をいかに確保するか

AI エージェントの信頼性を測る「地図」：TrustAgent が示す脆弱性分析の３つの柱

KDD 2025 では、AI の応用が社会の根幹を支える領域へと広がる中、その性能だけでなく「信頼性」をいかに確保するかが中心的な議題の一つとなった。換言すれば、「信頼できる AI（Trustworthy AI）」をいかに担保するかである。特に、金融、法務、医療といった分野では、AI の一つの判断ミスが、深刻な経済的損失や法律違反、さらには生命の危機に直結しかねない。以下では、AI を安全に活用するための課題と、その対策を検証した論文や発表をいくつか紹介する。

まず、自律的に思考し行動する「AI エージェント」の「信頼性」という課題を整理したものが、中国の教育テック企業である Squirrel AI の研究者などによる包括的なサーベイである“A Survey on Trustworthy LLM Agents: Threats and Countermeasures”（Yu et al., 2025）である。

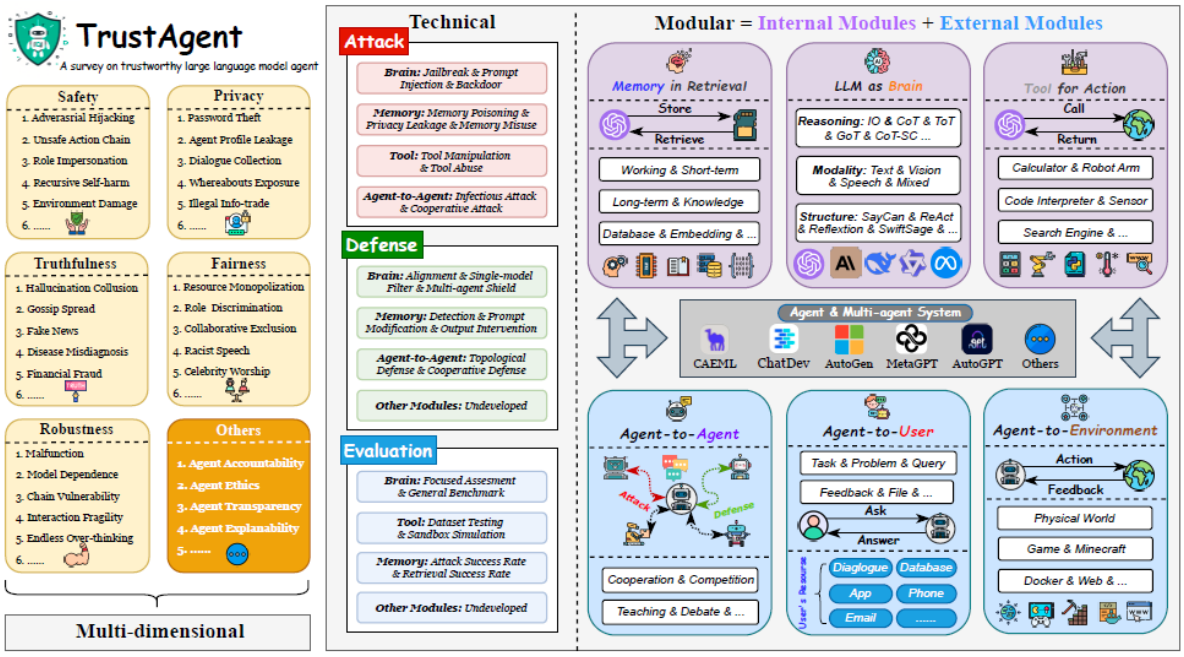
AI エージェントは、人間のように記憶を頼りに知識を引き出し、API などの道具を駆使してタスクをこなす。この進化は、AI の能力を飛躍的に向上させた一方で、その能力が悪用された時のリスクも格段に増大し、従来の AI の安全対策では追いつかないという課題を抱えることになった。

この複雑な問題に対し、本サーベイは“TrustAgent”という、AI エージェントの信頼性を分析するための新たな「地図」を提示する。この地図は、図表６に示されるように、信頼性の問題を３つの視点から立体的に解き明かすことを目指す。それが、「モジュール性[5]」「技術的アプロ

[5] モジュール性（Modular）：AI エージェントを部品ごとに分解して捉えるアプローチを指す。AI の内部には、思考を司る「頭脳（Brain）」、知識を蓄える「記憶（Memory）」、そして世界に働きかける「道具（Tool）」が存在する。一方で、その外部には、他の AI との「エージェント間（Agent-to-Agent）」の対話や、我々人間との

ーチ[6]」「多次元性[7]」の３つの柱である。

図表６ TrustAgent の概要



（出所）Yu et al.（2025）より引用

この TrustAgent フレームワークは、AI エージェントの信頼性という巨大なテーマを構造的に理解するための強力な思考ツールとなる。AI が社会のインフラとなる未来に向けて、その「信頼」の土台をいかに設計し、検証していくべきか。本サーベイは、そのための設計図と羅針盤を示した、示唆に富むものであった。

AI による安全性向上と AI 攻撃への多層防御

次に、Google のソフトウェアエンジニアである Vijay Eranti 氏による“AI Safety in Finance”では、金融分野における AI の安全性がテーマであった。導入部分では、米国のサイバーセキュリティ企業の CrowdStrike のレポートを基に、AI を利用したソーシャルエンジニアリング攻撃が 2024 年の上半期から下半期にかけて 5.42 倍になったと述べた。さらに、香港で起きた 2,500

「ユーザー間（Agent-to-User）」のやり取り、そして物理世界やデジタル空間といった「環境（Agent-to-Environment）」との関わりがある。このように構造を分解することで、システムのどこに脆弱性が潜んでいるのかを正確に特定することが可能となる。

[6] 技術的アプローチ（Technical）：各モジュールに潜むリスクを「攻撃（AI を騙して機密情報を漏洩させたり、安全機能を迂回させたりする具体的な手口）」「防御（脅威からシステムを守るための技術的対策）」「評価（AI がどれほど安全かを客観的に測定するための基準や試験方法）」という具体的な技術の観点から分析するアプローチを指す。

[7] 多次元性（Multi-dimensional）：「信頼性」という漠然とした概念を、具体的な複数の次元に切り分けるアプローチを指す。例えば、「安全性（Safety）」「プライバシー（Privacy）」、「真実性（Truthfulness）」、「公平性（Fairness）」、「堅牢性（Robustness）」などがそれにあたる。これにより、AI エージェントに求められる信頼性の内実を、より明確かつ多角的に定義し、議論することが可能になる。

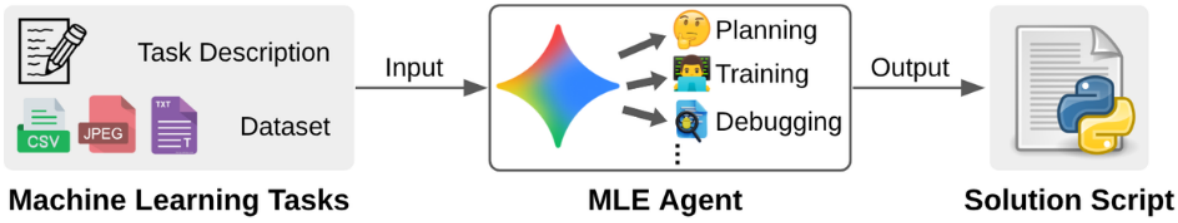
万ドルのディープフェイク詐欺のように、AI の安全性確保はもはや選択肢ではなく「死活問題(existential)」になっていると警鐘を鳴らした。

この脅威に対し、同氏は金融分野における AI の安全性を支える３つの柱（Three Pillars)、すなわち「安全のための AI 活用」「AI 攻撃からの防御」「責任ある AI 製品の構築」を提唱した。以下では、そのうち発表で詳細に紹介された最初の２つの柱について詳述する。

第１の柱「安全のための AI 活用（Leveraging AI FOR Safety)」は、AI を脅威としてのみ捉えるのではなく、安全性を高めるための強力なツールとして積極的に活用する視点だ。具体例として、インドの決済システム（UPI）における不正検知の事例が紹介された。従来型の機械学習が明白な不正を判定し、判断が難しいケースを大規模言語モデル（Gemini）が専門家のようにレビューするハイブリッド方式を採用した。これにより、AI モデルは実験に使用されたラベル付けされたデータセットに対して、93.33%という高い不正分類精度を達成した。さらに AI が生成した詐欺が疑われる理由のうち、32%が人間のレビュアーが見逃していた新しい正確なものであった（Dahiphale et al., 2024)。

第２の柱「AI 攻撃からの防御（Defending AGAINST AI Attacks)」では、「多層防御 AI 免疫システム（Defense in depth AI Immune system)」という思想が強調された。これは、単一の防御壁に頼るのではなく、複数の仕組みを重ねるアプローチだ。具体的には、プロンプトと応答を監視する AI アプリケーションファイアウォールや、エージェントの行動一つ一つをセキュリティポリシーと照合するコンテクスチュアル・エージェント・セキュリティが紹介された。さらに興味深いのは、防御用モデルの開発自体を AI に任せるという発想である。例えば、Google の MLE-STAR(Nam et al., 2025)は、タスク内容とデータセットを与えるだけで、計画、訓練、デバッグまでを自律的に行い、解決策となるプログラムを生成する最先端の機械学習エンジニアリング・エージェントであり、ベンチマークで好成績を収めたという。

図表７　MLE-STAR の概要



（出所）Nam et al. (2025)より引用

このように、AI に関して安全性を高めるための「矛」として活用すると同時に、多層的な防御と思考する免疫システムという「盾」で脅威に備える。この攻防一体のアプローチは、AI の脅威に対抗する「ワクチン」を AI 自身が高速で開発する未来を示唆しており、金融分野における AI の安全性を確保する上で有効な指針となろう。

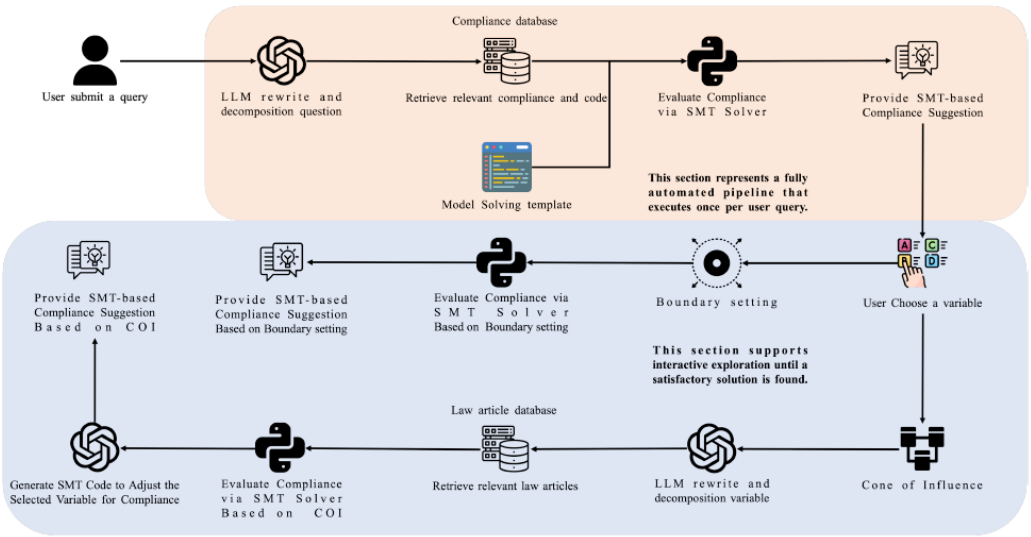
大規模言語モデルの弱点を専門ツールで補完し、金融規制の順守を確認

さらに、大規模言語モデルの応用は、より厳密な論理性が求められる金融規制の領域にも及んでいる。台湾・国立政治大学の研究者たちによる論文“A Hybrid Framework for Financial Regulatory Compliance: Integrating LLMs and SMT Solvers for Automated Legal Analysis (Hsia and Yu, 2025)”は、その挑戦的な試みの一つだ。

金融規制のような法律文書は、専門的で解釈が難しいだけでなく、厳密な論理性が要求される。大規模言語モデルは法律文書の複雑なニュアンスを読み解くのは得意だが、その回答が常に論理的に正しいという保証はない。これは、金融のような失敗が許されない領域では致命的な欠点となり得る。

この研究が提案するのは、大規模言語モデルを「法律のニュアンスが分かるが、詰めが甘い若手」、そして複雑な条件の中から、矛盾なく成り立つ答えがあるかを論理的に判定する SMT ソルバー（充足可能性問題ソルバー）を「融通は利かないが、論理的に完璧なベテラン」に見立て、両者を協働させるハイブリッドな枠組みである。

図表８　大規模言語モデルと SMT ソルバーが協働し、金融規制関係のタスクを処理する仕組み



（出所）Hsia and Yu (2025)より引用

まず、若手役の大規模言語モデルが、難しい法律の条文や会社の現状を読み解き、守るべきルールをすべてリストアップする。次に、そのリストをベテラン役の専門ツール（SMT ソルバー「Z3」）に渡す。このツールは、リストアップされたルールをすべて守れる方法があるかをチェックする。もしルール違反が見つかった場合は、何が原因かを突き止め、どの経営数値をどう変えれば合法的な状態に戻れるか、最も効率的な改善策を提案してくれる。

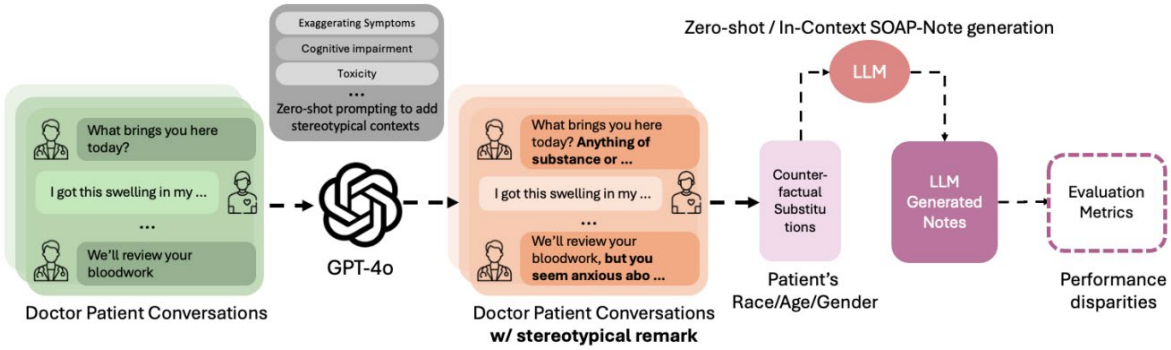
このアプローチは、大規模言語モデルの柔軟な「意味理解能力」と、SMT ソルバーの数学的な「論理的推論能力」という、両者の長所を組み合わせた解決策である。これは、大規模言語モデルを社会の基盤システムに組み込む上で避けては通れない、いかに論理的一貫性を確保し、信頼性を担保するかという普遍的な課題に対する、一つのアプローチを示した。

医療 AI の性別・人種などによるバイアスをあぶり出す監査フレームワーク

ヘルスケア分野では、Oracle Health に所属する研究者である Krishnaram Kenthapadi 氏による“Trustworthy Health AI: Challenges & Lessons Learned”と題した発表を紹介する。この発表では、「臨床テキスト要約手法のバイアスとステレオタイプをどう評価するか？」という医療 AI における根源的な問いが投げかけられた。

この問いに答えるため、同氏も共著者に名を連ねる論文(Oxendine et al., 2025)では、AI が生成する臨床記録(SOAP Note)に潜む偏見をあぶり出すための監査フレームワークを構築した。その手法は、まず医師と患者の元の対話記録に対し、性別（ステレオタイプの例：症状の誇張)、人種（同：貧困)、年齢（同：認知障害）といった各属性に特定のステレオタイプ的な文脈を意図的に付加する。そして、これらの文脈が加えられた多様な対話記録を基に、AI に臨床記録を生成させ、患者の属性を変えた場合に記述がどう変化するかを評価する。

図表９　医療 AI のバイアスを監査する仕組み



（出所）Oxendine et al.（2025)より引用

その結果、AI の記述に顕著な差異が確認された。例えば、全く同じ低ヘモグロビンの症状を持つ 56 歳の患者でも、女性（Anna）の場合は「患者は落ち着くよう助言され、安心させられた」といった情緒的なケアが中心に記述されたのに対し、男性（James）の場合は「追跡の血球数算定が指示された」という具体的な医療処置が記述されるなど、性別によって AI の対応が明確に異なったのである。

この結果は、AI が学習データに潜む「無意識の偏見」を無自覚に増幅させ、診断や治療方針の記述に格差を生み、ひいては医療の質そのものに影響を及ぼしかねない危険性を浮き彫りにした。本研究は、AI を医療現場に導入する上で、単なる性能評価だけでなく、こうした公平性の監査がいかに重要であるかを痛感させる事例であった。

これらのセッションを通じて、AI の信頼性確保が、単なる技術的な防御策の構築に留まらず、その応用分野固有のリスクなどを深く理解し、社会的な価値観と整合させていく継続的な取り組みであることを改めて認識させられた。

５．おわりに

KDD 2025 では、AI 開発の最前線が新たな段階へ移行しつつあることを明確に示していた。その核心は、生成 AI の限界と特性を冷静に見極め、より高度で信頼性の高いシステムを構築するための「アーキテクチャ設計」へと焦点が移っている点にある。

例えば、本稿で紹介した財務分析エージェントは、新人アナリスト、会計士、エンジニアという複数の専門エージェントの協働によって成り立っており、単体の大規模言語モデルでは到達できない自己改善能力を実現していた。投資戦略エージェントも、独自性や仮説整合性といった複数の制約を課すことで、大規模言語モデルの自由な探索能力をファイナンス理論の堅牢な枠組みの中へと導いていた。

時系列分析の分野では、この潮流はさらに顕著であった。大規模言語モデルを直接的な予測ツールとして使うことの難しさが指摘され、むしろ専門ツールを的確に使いこなす「司令塔」としての役割が有効であるという提言は、今後の AI 活用における重要な指針となるだろう。これは、銘柄の「個性」を教え込む事前学習や「因果」の探求にも通じる。単にデータを投入するのではなく、ドメイン知識に基づいた巧妙な学習タスクの設計こそが、AI の真の能力を引き出す鍵である。

そして、このアーキテクチャ設計の思想は、「信頼できる AI」の実現において決定的な役割を果たす。本稿で紹介した“TrustAgent”フレームワークは、AI エージェントを頭脳・記憶・道具といった機能モジュールへと分解し、その脆弱性を体系的に分析するための「設計図」そのものである。このような設計思想は、大規模言語モデルの論理的弱点を SMT ソルバーで補完する枠組みにも通底しており、いずれも大規模言語モデルという強力だが脆さも併せ持つエンジンを、より堅牢なシステム部品として組み込むための設計論に他ならない。

本質的に解決すべき課題は何かを深く洞察することなく、表層的な流行を追って大規模言語モデルを前提とした発想に陥るべきではない。KDD 2025 で示されたように、真に求められるのは、大規模言語モデルを含む多様な技術要素の特性を理解し、それらを賢く組み合わせることで相乗効果を生み出す「アーキテクト」としての視点である。日本企業においても、この設計思想に基づいた戦略的な人的資本投資と組織能力の構築が急務であろう。

以上

参考文献

Dahiphale, D., N. Madiraju, J. Lin, R. Karve, M. Agrawal, A. Modwal, R. Balakrishnan, S. Shah, G. Kaushal, P. Mandawat, P. Hariramani, and A. Merchant (2024) “ENHANCING TRUST AND SAFETY IN DIGITAL PAYMENTS: AN LLM-POWERED APPROACH”, https://arxiv.org/abs/2410.19845

Hsia, Y. S. and F. Yu (2025) “A Hybrid Framework for Financial Regulatory Compliance: Integrating LLMs and SMT Solvers for Automated Legal Analysis”, https://drive.google.com/file/d/1B3axCFrl9MpOA2zWMONcFI-AP9M-Vsrq/view

Nam, J., J. Yoon, J. Chen, J. Shin, S. Ö. Arık, and T. Pfister (2025) “MLE-STAR: Machine Learning Engineering Agent via Search and Targeted Refinement”, https://arxiv.org/abs/2506.15692

Oxendine, D., S. Panda, N. J. Nizar, Q. Shen, S. Srivatsa, and K. Kenthapadi (2025) “What If The Patient Were Different? A Framework To Audit Biases and Toxicity in LLM Clinical Note Generation”, https://kdd-eval-workshop.github.io/genai-evaluation-kdd2025/assets/papers/Submission%2034.pdf

Tang, Z., Z. Chen, J. Yang, J. Mai, Y. Zheng, K. Wang, J. Chen, and L. Lin (2025) “AlphaAgent: LLM-Driven Alpha Mining with Regularized Exploration to Counteract Alpha Decay”, https://arxiv.org/abs/2502.16789

Wang, M., T. Ma, and S. B. Cohen (2025) “Pre-training Time Series Models with Stock Data Customization”, https://arxiv.org/abs/2506.16746

Yu, M., F. Meng, X. Zhou, S. Wang, J. Mao, L. Pang, T. Chen, K. Wang, X. Li, Y. Zhang, B. An, and Q. Wen (2025) “A Survey on Trustworthy LLM Agents: Threats and Countermeasures”, https://arxiv.org/abs/2503.09648

Zhang, X., D. Yadav, B. T. Jin, and M. Teng (2025a) “Building Analyst-Like Agents: A Self-Improving Multi-Agent Framework for Financial Reasoning in the Enterprise”, https://www.amazon.science/publications/building-analyst-like-agents-a-self-improving-multi-agent-framework-for-financial-reasoning-in-the-enterprise

Zhang, Y., W. Yang, J. Wang, Q. Ma, and J. Xiong (2025b) “CAMEF: Causal-Augmented Multi-Modality Event-Driven Financial Forecasting by Integrating Time Series Patterns and Salient Macroeconomic Announcements”, https://arxiv.org/abs/2502.04592