NTU Nanyang Business School
Dean's Distinguished Speaker Series

AI for Social Sciences:

Generative Modeling Using LLMs and Beyond

Lin William Cong

Cornell SC Johnson, ABFER, IC3, & NBER

June 2025

Cornell University

Core Themes of Modern AI (Beyond Basic ML)

- AI (McCarthy 1955): technologies allowing machines to perform complex tasks that typically require human intelligence.
- Intelligence (HAI): the ability to learn and perform suitable techniques to solve problems and **achieve goals**, in an uncertain, ever-varying world.
- Modern AI: Instruction-driven automation/prediction $\rightarrow$
  (i) end-to-end goal-oriented optimization in enlarged modeling space,
  (ii) using generative modeling.
- The Tao/Ways of AI:
  1. Larger models, greater computation, general intelligence.
  2. Bigger data and associated computation; pre-training, self-learning.
  3. Goal-oriented algorithms (e.g., RL) as effective search/optimization.
  4. Deep generative modeling (e.g., ChatGPT), esp. suitable for social sciences.
  5. AI agents: understanding them as substitutes or complements to humans.

Core Themes of Modern AI (Beyond Basic ML)

- AI (McCarthy 1955): technologies allowing machines to perform complex tasks that typically require human intelligence.
- Intelligence (HAI): the ability to learn and perform suitable techniques to solve problems and **achieve goals**, in an uncertain, ever-varying world.
- Modern AI: Instruction-driven automation/prediction $\rightarrow$
  (i) end-to-end goal-oriented optimization in enlarged modeling space,
  (ii) using generative modeling.
- The Tao/Ways of AI:
  1. Larger models, greater computation, general intelligence.
  2. Bigger data and associated computation; pre-training, self-learning.
  3. Goal-oriented algorithms (e.g., RL) as effective search/optimization.
  4. Deep generative modeling (e.g., ChatGPT), esp. suitable for social sciences.
  5. AI agents: understanding them as substitutes or complements to humans.
- AI/ML in Finance: conventional ML $+$ (off-the-shelf) text analytics
  - Return prediction, asset pricing, investment management, etc.
  - Handles big data, but learning/training through examples/statistics.
  - Text analytics/LLMs; customized language models (Hoberg & Manela, 2025).

True Power of GenAI: (i) AI Agents as Substitutes/Complements

- Understanding the **Behavioral Economics of AI: LLM Biases and Corrections** (Bini, Cong, Huang, & Jin, 2025)

Questions that study the psychology of preferences

1. Prospect theory – diminishing sensitivity
2. Prospect theory – Loss aversion
3. Prospect theory – Probability weighting
4. Narrow framing
5. Hyperbolic discounting
6. Ambiguity aversion

Economic experiments on preferences, beliefs, and optimizations

Questions that study the psychology of beliefs

1. Sample size neglect (1) – binominal
2. Sample size neglect (2) – average estimation
3. Sample size neglect (3) – absolute vs. relative sample size
4. Base rate neglect
5. Conjunction fallacy
6. Gambler's fallacy
7. Confirmation bias
8. Anchoring
9. Overconfidence – over-precision
10. Overconfidence – over-estimation

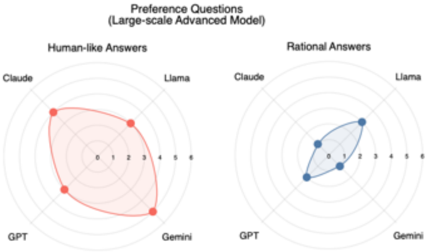
Prospect theory – diminishing sensitivity

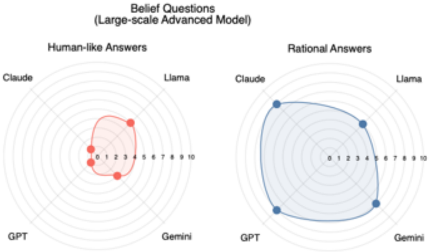
In addition to whatever you own, you have been given \$1,000. You now need to choose between the following two options:
A1: (\$1,000, 0.5), meaning winning \$1,000 with 0.5 probability and winning zero with 0.5 probability
B1: (\$500), meaning winning \$500 with certainty.

In addition to whatever you own, you have been given \$2,000. You now need to choose between the following two options:
A2: (−\$1,000, 0.5), meaning losing \$1,000 with 0.5 probability and losing zero with 0.5 probability
B2: (−\$500), meaning losing \$500 with certainty

Rational responses: A1 and A2 (less risk averse) | B1 and B2 (more risk averse)
Human responses: human participants tend to choose B1 (risk averse in gain region) and A2 (risk loving in loss region)

True Power of GenAI: (i) AI Agents as Substitutes/Complements





For experimental questions that study the psychology of beliefs:

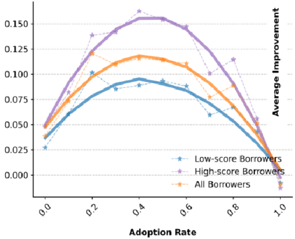
- LLMs' responses become more rational for more advanced models
- LLMs' responses become more rational for models with a larger parameter scale

For experimental questions that study the psychology of preferences:

- LLMs' responses become more irrational and human-like for more advanced models
- LLMs' responses become more irrational and human-like for models with a larger parameter scale

True Power of GenAI: (ii) Data-Driven Counterfactual Analysis

- Counterfactual equilibrium, Lucas Critique, deep parameters, and a new simulation-based micro-foundation.
- Writing Quality and Soft Informationin the GenAI Age: Evidence from Online Credit Markets (Cong, Guo, Zhao, & Zhou, 2024); Data-Driven Generative Equilibrium (Cong, 2025).
  - Quality distribution can be resampled from historical cases.
  - Borrowers as goal-oriented generative agents.
  - Lenders as data-driven discriminative agents.
  - Both lenders and borrowers' actions should be optimal in equilibrium.
  - GE features fixed points (e.g., LLM/tech adoption or interest rate).



The set of agents $\mathcal{N} = \mathcal{G} \cup \mathcal{D}$ consists of:

- **Generative agents** $i \in \mathcal{G}$ with compact high-level action sets $A_i^H \subset \mathbb{R}^{k_i}$ and low-level policies $a_i^L \sim G_i(\cdot|a_i^H)$.
- **Discriminative agents** $j \in \mathcal{D}$ updating beliefs $\mu_j(\cdot|\mathbf{D}^C)$ on counterfactual data $\mathbf{D}^C = \Delta(\mathbf{a}^H, \mathbf{a}^L, \omega)$.

Each agent's payoff is:

$$u_i\left(a_i^H, a_i^L; \mathbf{a}_{-i}^H, \mathbf{a}_{-i}^L, \{a_j\}, \omega\right) \quad \text{(Generative)}$$

$$\mathbb{E}^{\mu_j(\cdot|\mathbf{D}^C)}\left[u_j(a_j; \mathbf{a}^H, \mathbf{a}^L, \omega)\right] \quad \text{(Discriminative)}$$

DDGE exists If

1. The strategy space $\prod_{i\in\mathcal{G}}\{0,1\}$ is a complete lattice,
2. The best response correspondence $\mathrm{BR}_i(\mathbf{a}_{-i}^H)$ is monotone

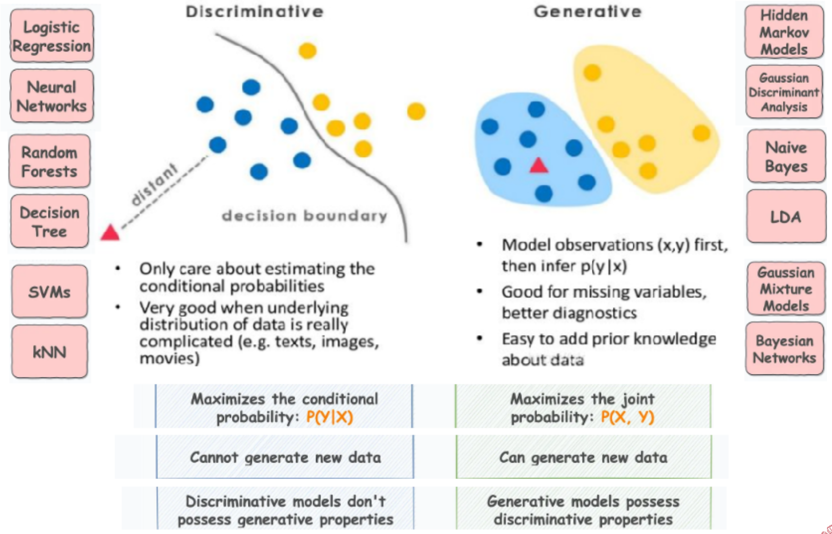
Sufficient Condition for Monotonicity — If payoffs u_i satisfy:

$$u_i(1, a_i^L; \mathbf{a}_{-i}^H) \geq u_i(0, a_i^L; \mathbf{a}_{-i}^H) \quad \forall \mathbf{a}_{-i}^H,$$

then BR_i is monotone.

Cornell University

True Power of GenAI: (iii) Goal-Oriented Algorithms in Large Spaces



Cornell University

Generative Modeling Beyond LLMs

- "Big Four" Families of Generative Models
  1. Variational Autoencoders (VAEs)
  2. Adversarial Networks (GANs)
  3. Autoregressive Models
  4. Diffusion Models / Score-Based Models

Generative Modeling Beyond LLMs

- “Big Four” Families of Generative Models
  1. Variational Autoencoders (VAEs)
  2. Adversarial Networks (GANs)
  3. Autoregressive Models
  4. Diffusion Models / Score-Based Models
  5. Others: Normalizing Flows, Energy-Based Models, etc.

Generative Modeling Beyond LLMs

- "Big Four" Families of Generative Models
  1. Variational Autoencoders (VAEs)
  2. Adversarial Networks (GANs)
  3. Autoregressive Models
  4. Diffusion Models / Score-Based Models
  5. Others: Normalizing Flows, Energy-Based Models, etc.

- Types of generative models:
  - **An unconditional generative model** learns the overall data distribution $p(x)$ without relying on any additional input or condition. Its primary objective is to generate samples that resemble the entire data set.
  - **A conditional generative modeling** incorporate extra information or conditions. It learns the conditional distribution $p(x \mid y)$, where y can be any auxiliary information (e.g., class labels, attributes).

- GOALS constitute a new type of generative model beyond LLMs.

Generative Modeling Beyond LLMs

- "Big Four" Families of Generative Models
  1. Variational Autoencoders (VAEs)
  2. Adversarial Networks (GANs)
  3. Autoregressive Models
  4. Diffusion Models / Score-Based Models
  5. Others: Normalizing Flows, Energy-Based Models, etc.

- Types of generative models:
  - **An unconditional generative model** learns the overall data distribution $p(x)$ without relying on any additional input or condition. Its primary objective is to generate samples that resemble the entire data set.
  - **A conditional generative modeling** incorporate extra information or conditions. It learns the conditional distribution $p(x \mid y)$, where y can be any auxiliary information (e.g., class labels, attributes).

- GOALS constitute a new type of generative model beyond LLMs.

- Hallucination or creativity?
  Robust control balances creativity and reliability.

Goal-Oriented Portfolio Management Through Transformer-Based Reinforcement Learning

Lin William Cong, Cornell University
Ke Tang, Tsinghua University
Jingyuan Wang, Beihang University

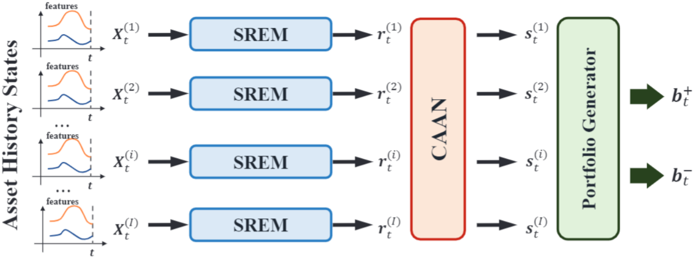
Goal-Oriented Portfolio Management Through Transformer-Based RL

- First "large" model in finance ($>$ 10 million parameters) with Transformer/offline DRL in 2019.
  - Virtual of complexity/large models (Belkin et al., 2019, Kelly, Malamud, & Zhou, 2022, Fan et al., 2022).
  - Share how humans learn the spirit of Grounded Theory ($> 20,000$ citations).
  - Non-textual, proprietary, but not too large.
- Advantages:
  - Anything objective that can be written as a cumulative reward function; not just utility over terminal wealth.
  - High-dimensional and nonlinear data approximated by deep neural nets instead of linear functions or index functions.
  - Action-state interaction, partial/non-equilibrium, unknown states and payoffs.
  - Cutting-edge sequence modeling (i.e., Transformer) plus cross-sectional information extraction through attention mechanisms.
- Bridge between engineering, AI architecture, and finance.
- RL uniquely fit for portfolio management problems.

Economic Question and AI for Finance

- DRL comes to the rescue:
  - Anything objective that can be written as a cumulative reward function; not just utility over terminal wealth.
  - High-dimensional and nonlinear data approximated by deep neural nets instead of linear functions or index functions.
  - Action-state interaction, partial/non-equilibrium, unknown states and payoffs.
  - Cutting-edge sequence modeling (i.e., Transformer) plus cross-sectional information extraction through attention mechanisms.

Architecture Innovation: Transformer + Cross-Asset Attention Network + RL



CAAN & Winner Score Estimation

- Cross-Asset Attention Network (CAAN)
  - Built on self-attention mechanism (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, & Polosukhin, 2017).
- Query vector $\boldsymbol{q}^{(i)}$, key vector $\boldsymbol{k}^{(i)}$, value vector $\boldsymbol{v}^{(i)}$:

$$\boldsymbol{q}^{(i)} = \boldsymbol{W}^{(Q)}\boldsymbol{r}^{(i)}, \quad \boldsymbol{k}^{(i)} = \boldsymbol{W}^{(K)}\boldsymbol{r}^{(i)}, \quad \boldsymbol{v}^{(i)} = \boldsymbol{W}^{(V)}\boldsymbol{r}^{(i)}, \tag{1}$$

- Attenuation score:

$$\boldsymbol{a}^{(i)} = \sum_{j=1}^{I} \mathrm{SATT}\left(\boldsymbol{q}^{(i)}, \boldsymbol{k}^{(j)}\right) \cdot \boldsymbol{v}^{(j)}, \tag{2}$$

$$\mathrm{SATT}\left(\boldsymbol{q}^{(i)}, \boldsymbol{k}^{(j)}\right) = \frac{\exp\left(\beta_{ij}\right)}{\sum_{j'=1}^{I} \exp\left(\beta_{ij'}\right)}; \quad \beta_{ij} = \frac{\boldsymbol{q}^{(i)\top} \cdot \boldsymbol{k}^{(j)}}{\sqrt{d_k}}. \tag{3}$$

- Winner score $s^{(i)} = \mathrm{sigmoid}\left(\boldsymbol{w}^{(s)\top} \cdot \boldsymbol{a}^{(i)} + e^{(s)}\right)$, where $\boldsymbol{w}^{(s)}$ and $e^{(s)}$ are the connection weights and the bias to be learned.

Application to U.S. Equities

- CRSP, Compustat, EDGAR, etc.
- July 1965-June 2016: pre-1990, 1990-2016, post 2001
- 51 characteristics/signals with 1-12 month lags, similar to (Freyberger et.al., 2019).
    - Past-return based.
    - Investment-related characteristics: annual percentage change in total assets (**Investment**),change in inventory over total assets (**IVC**),
    - Profitability-related characteristics such as gross profitability over the book-value of equity (**Prof**) or return on operating assets (**ROA**),
    - Intangibles such as operating accruals (**OA**) and tangibility (**Tan**)
    - Value-related characteristics such as the book-to-market ratio (**BEME**) and earnings-to-price (**E2P**)
    - Trading frictions such as the average daily bid-ask spread (**Spread**) and standard unexplained volume (**SUV**).
- Macro variables and alternative data: MD&A (10-K & 10-Q), Risk Factor Discussions (1A of 10-K), Analyst Reports, etc.

AlphaPortfolio Performance on Test Samples

	AP Performance				AP Excess Alpha					
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
Firms	All	$> q_{10}$	$> q_{20}$	Factor Models	All $\alpha(\%)$	All R^2	$> q_{10}$ $\alpha(\%)$	$> q_{10}$ R^2	$> q_{20}$ $\alpha(\%)$	$> q_{20}$ R^2
Return (%)	17.00	17.10	18.10	CAPM	13.9***	0.005	12.2***	0.088	14.0***	0.102
Std.Dev. (%)	8.50	7.70	8.20	FFC	14.2***	0.052	13.4***	0.381	14.7***	0.465
Sharpe	2.00	2.31	2.21	FFC+PS	13.7***	0.054	12.3***	0.392	13.3***	0.480
Skewness	1.42	1.74	1.91	FF5	15.3***	0.12	13.8***	0.426	14.7***	0.435
Kurtosis	6.33	5.70	5.97	FF6	15.6***	0.128	14.5***	0.459	15.8***	0.516
Turnover	0.26	0.24	0.26	SY	17.4***	0.037	15.8***	0.332	17.0***	0.394
MDD	0.08	0.02	0.02	Q4	16.0***	0.121	15.0***	0.495	16.2***	0.521

	AP Performance				AP Excess Alpha					
	(1)	(2)	(3)		(4)	(5)	(6)	(7)	(8)	(9)
Firms	All	$> q_{10}$	$> q_{20}$	Factor Models	All $\alpha(\%)$	All R^2	$> q_{10}$ $\alpha(\%)$	$> q_{10}$ R^2	$> q_{20}$ $\alpha(\%)$	$> q_{20}$ R^2
Return(%)	11.11	17.86	14.08	CAPM	9.64***	0.041	19.01***	0.025	15.51***	0.049
Std.Dev.(%)	6.82	6.74	6.00	FFC	10.91***	0.093	19.29***	0.316	15.48***	0.326
Sharpe	1.63	2.65	2.35	FFC+PS	10.86***	0.093	18.45***	0.352	14.33***	0.413
Skewness	0.12	0.42	-0.19	FF5	10.73***	0.134	19.21***	0.401	15.85***	0.342
Kurtosis	0.38	0.76	0.43	FF6	10.65***	0.135	18.94***	0.413	15.56***	0.357
Turnover	0.20	0.24	0.24	SY	-	-	-	-	-	-
MDD	0.04	0.02	0.02	Q4	11.67***	0.234	21.90***	0.308	17.45***	0.232

AlphaManager: A Data-Driven-Robust-Control Approach to Corporate Finance

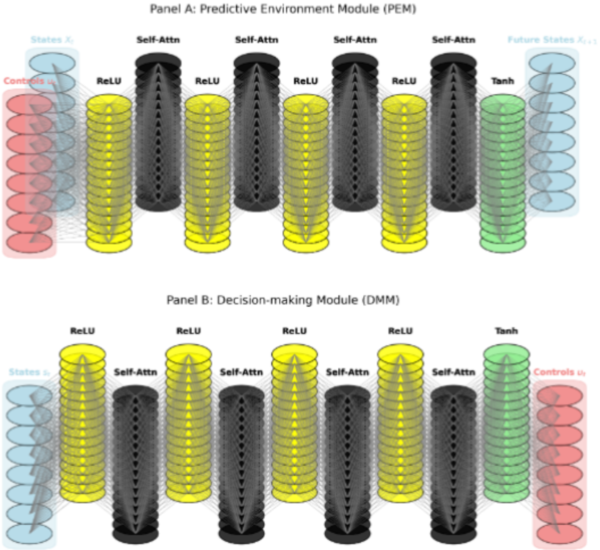
Murillo Campello, University of Florida and NBER
Lin William Cong, Cornell University and NBER
Luofeng Zhou, NYU Stern

A Data-Driven-Robust-Control Approach to Corporate Finance

Understanding Underlying Economic Mechanisms	←		→	Empirically Explaining Outcomes & Guiding Actions
	Reduced Form	**Structural Estimation**	???	
	transparent easy to estimate locally causal	holistic theory-based counterfactual	data-driven unrestricted modeling counterfactual	

- What are managers' optimal actions given a certain objective?
  - Counterfactual statements are essential for aiding CF decisions
    - Experiments & surveys: costly, infrequent, sometimes ethically questionable.
    - Reduced-form: restrictive models, limited performance, partial equilibrium.
    - Causal inferences: local and low-dimensional; fragmented knowledge.
    - Structural approaches: need a theory that accounts for the general dynamics.
  - Manager actions $\Leftrightarrow$ environment (e.g., Edmans, Goldstein & Jiang, 2015).
  - Environment stochastic, complex, often unknown.
- CF fundamentally a stochastic control problem:
  - Info set: internal states (accounting, audits, etc.) and market environments.
  - Dynamically learned, high-dimensional and nonlinear.
  - Action space: high-dimensional policy $\rightarrow$ need robust counterfactuals!

AlphaManager (AM) Architecture



Two Baseline Modules + Robust Control

Predictive Environment module (PEM, 10 auxiliaries, 512+64+5):

- Roughly 30% training. 1976Q1 - 1991Q4; Transformer-based; quarterly rolling.

Decision-Making Module: (DMM), RL/policy-gradient, 256:64+5:

$$\max_{\{u_{t_0},\ldots,u_{t_0+T}\}} \mathbb{E}_{t_0} \sum_{t=t_0}^{t_0+T} r(X_t, u_t) \qquad s.t.\ \Delta X_{t+1} = f(X_t, u_t) + \varepsilon_{t+1} \quad \& \quad u_t = g(X_t)$$

Robust Control and Ambiguity (Hansen and Sargent, 2023):

- Overfit and data shifts.
- Misspecification - limited power of the model class
- Risk - in-model stochastic innovation
- Ambiguity - uncertainty about model choice
- Inspiration from climate finance (Barnett, Brock, and Hansen, 2020): max-min + relative-entropy punishment + probability adjustment.

Empirical Results: PEM's Predictions of Firm Outcomes

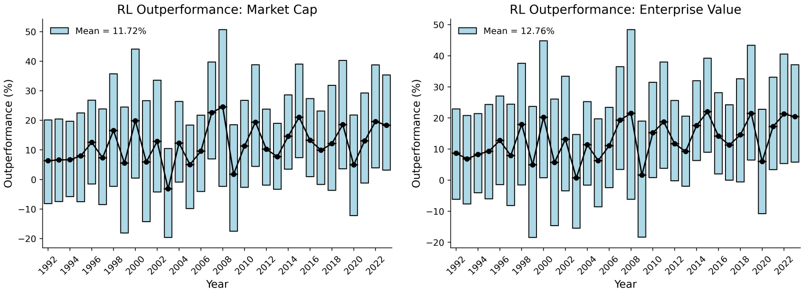
- High-dimensional, high-fidelity OOS, reduce costly experiments.

	Ignoring Control Information		Including Control Information	
State Variable	Training R^2	Test R^2	Training R^2	Test R^2
Total Assets	–4.09%	–8.15%	55.44%	62.56%
Current Assets	–3.58%	–7.10%	44.49%	51.21%
Sales	29.54%	28.68%	31.33%	30.88%
Payables	21.46%	24.43%	24.40%	27.64%
COGS	25.68%	26.76%	27.00%	28.56%
InterestExp	73.26%	77.17%	73.36%	77.28%
Inventory	12.78%	13.71%	17.04%	18.92%
CurrentLiability	8.88%	7.72%	21.89%	22.69%
Receivabless	17.52%	18.77%	21.59%	23.20%
Revenue	29.51%	28.59%	31.31%	30.80%
MarketCap	1.32%	–3.33%	9.32%	7.07%
EnterpriseValue	–0.97%	–5.73%	14.61%	13.14%
Volume	12.81%	16.53%	15.77%	20.75%
Return	47.90%	45.27%	50.04%	48.19%

- Controls more important for some state evolution.
- Consistent with known local (causal) patterns from the literature.

Out-Performance of AlphaManager (DMM)

- Objectives: next Q and next 8Q market cap and enterprise value.
- Next Q market cap increase (short-termist)
- Overall short-horizon performance: **11.72%** and 12.76%.
- Long-horizon objective: 8.73% and 4.43%
- Alternative objectives and IRL/GANs.



Optimal Actions Versus Historical Actions

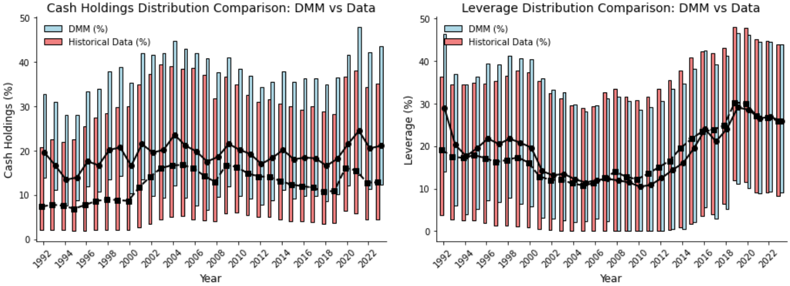


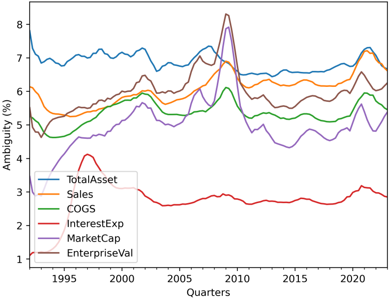
Figure: Optimal decisions (blue lines) vs real decisions (red lines): cash holdings (left) and leverage (right)

Maximizing **next Q enterprise value**: more acquisitions, increasing cash holdings more, keeping the same leverage, paying out more dividend, and increasing investment, especially in R&D, allowing more variations in investments, and more repurchases during bad times.

Economic Questions DDRC Enables Asking for the First Time

Piecing together CF research and answering new questions

- Ambiguity and the need for theory/reduced-form/structural models.
  - Boundary of data-driven approach.



Economic Questions DDRC Enables Asking for the First Time

Piecing together CF research and answering new questions

- Ambiguity and the need for theory/reduced-form/structural models.
  - Boundary of data-driven approach.
- Ambiguity-guided transfer learning/pre-training.
  - Combining insights and predictions from other approaches/applications.
  - For example, investment-CAPM (Liu, Whited, and Zhang, 2009) links investment, sales, profitability and expected returns
  - For each observation, use investment-CAPM to synthesize counterfactual observations where sales, profitability and expected returns are impacted by a counterfactual investment ratio
  - Then, use these synthetic data to train PEM
- Holistically studying managerial objectives.
  - Actions are not additive $\rightarrow$ cannot simply study partial objectives.

Revealed Objectives and Inverse Reinforcement Learning

- Revealed preference argument and Min-Max formulation:

$$\underset{r(\cdot,\cdot)}{\text{argmin}}\ \max_{g(\cdot)} \mathbb{E}\left[\int_{t_0}^{t_0+T} r(X_t, g(X_t)) - r(X_t, u_t^*)\, dt\right]$$
$$s.t.\ dX_t = f(X_t, g(X_t))dt + \Sigma_x dB_t. \tag{4}$$

  - u_t^* – real/optimal managerial decisions
  - r reflects real objectives $\Longleftrightarrow$ g^* yields zero difference

- Min-Max $\Rightarrow$ Min + Max + $g(X_t) = u_t^*$ or $D(g) = 0$, where

$$D(g) = \min_{r(\cdot,\cdot)} \mathbb{E}\left[\int_{t_0}^{t_0+T} r(X_t, g(X_t)) - r(X_t, u_t^*)\, dt\right] \tag{4}$$
$$s.t.\ dX_t = f(X_t, g(X_t))dt + \Sigma_x dB_t \tag{5}$$

- Generative Adversarial Networks (GANs)

Panel Trees as Effective Greedy GOALS

- Interpretable and not-so-large AI models (not supervised/unsupervised).
- Human intelligence and *The Art of War*: divide and conquer.
- Conventional trees not designed for panel data, economic objectives, and interpretability comes at the cost of overfitting.
- Panel trees: panel data analytics with global split criteria using economic goals instead of local split criteria based on statistical error minimization.

Panel Trees as Effective Greedy GOALS

- Interpretable and not-so-large AI models (not supervised/unsupervised).
- Human intelligence and *The Art of War*: divide and conquer.
- Conventional trees not designed for panel data, economic objectives, and interpretability comes at the cost of overfitting.
- Panel trees: panel data analytics with global split criteria using economic goals instead of local split criteria based on statistical error minimization.
- Creating test assets and latent factor models:
  - "**Growing the Efficient Frontier on Panel Trees**" (Cong, Feng, He, & He, 2025 JFE).
- Uncommon factors and macro regimes for asset pricing:
  - "**Sparse Modeling Under Grouped Heterogeneity with an Application to Asset Pricing**" (Cong, Feng, He, & Li, 2022).
- Heterogeneous predictability and trading:
  - "**Mosaics of Predictability**" (Cong, Feng, He, & Wang, 2024).

Growing the Efficient Frontier on Panel Trees

Lin William Cong, Cornell University

Gavin Feng, City University of Hong Kong

Jingyu He, City University of Hong Kong

Xin He, University of Science and Technology of China

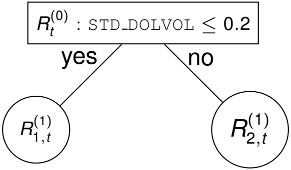
Growing A P-Tree

- Security sorting routinely used for testing pricing models and investment.
- Gaps in pricing test assets and spanning the efficient frontier.
- P-Tree generating test assets and latent factors.

Growing A P-Tree

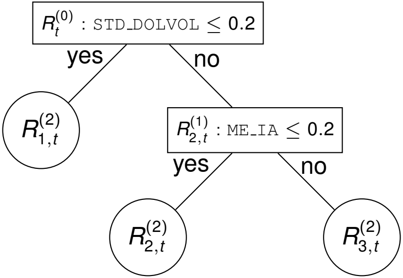
- Security sorting routinely used for testing pricing models and investment.
- Gaps in pricing test assets and spanning the efficient frontier.
- P-Tree generating test assets and latent factors.

Consider a split point candidate



- $R_t^{(0)}$ denotes vector of market returns (value weighted) at the root.
- $R_{j,t}^{(k)}$ is the leaf-basis portfolio of the j-th terminal node after the k-th split.
- Leaf-basis portfolios can be value / equally weighted – the panel data structure for returns.

P-Tree Test Assets and Factor Models

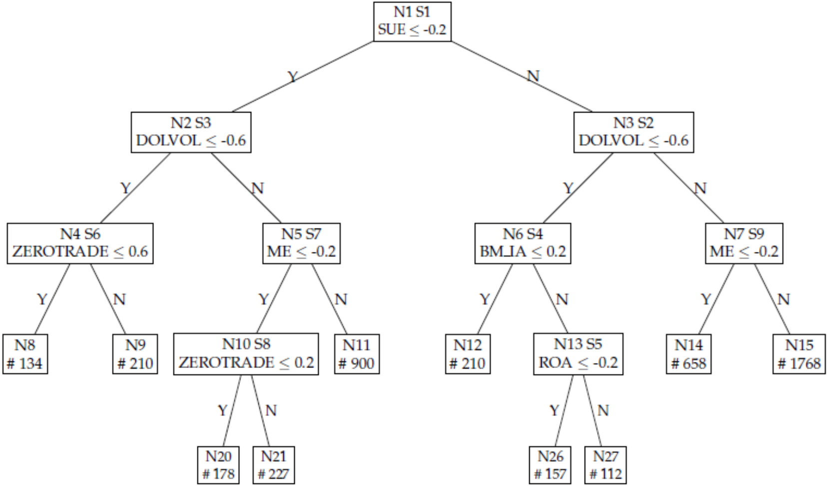


- The second split gives us three leaf basis portfolios and a updated SDF:

$$f_t^{(2)} = \widehat{\Sigma}_2^{-1} \widehat{\mu}_2 R_t^{(2)} = w_{21} R_{1,t}^{(2)} + w_{22} R_{2,t}^{(2)} + w_{23} R_{3,t}^{(2)},$$

- For the second split, the algorithm searches over all leaf nodes, characteristics, and breakpoints.
- The split criterion is calculated based on the entire cross section, thus P-Tree and its SDF are global.

A P-Tree Grown on U.S. Equities



A Grown P-Tree

	SR	α_{CAPM}	α_{FF5}	α_{Q5}	α_{RP5}	α_{IP5}	SR	α_{CAPM}	α_{FF5}	α_{Q5}	α_{RP5}	α_{IP5}
	Panel B1: 20 Years In-Sample (1981-2000)						Panel B2: 20 Years Out-of-Sample (2001-2020)					
P-Tree1	7.12	1.85	1.88	1.71	1.64	1.24	3.24	1.34	1.31	1.23	1.30	0.91
P-Tree1-5	12.14	0.80	0.79	0.77	0.75	0.71	2.98	0.52	0.47	0.49	0.48	0.32
P-Tree1-10	19.16	0.64	0.65	0.62	0.61	0.57	2.57	0.36	0.29	0.33	0.34	0.19
P-Tree1-15	27.74	0.53	0.53	0.52	0.52	0.49	2.48	0.30	0.23	0.27	0.27	0.16
P-Tree1-20	41.12	0.47	0.47	0.46	0.46	0.45	2.57	0.26	0.20	0.24	0.25	0.15
	Panel C1: 20 Years In-Sample (2001-2020)						Panel C2: 20 Years Out-of-Sample (1981-2000)					
P-Tree1	5.82	1.50	1.46	1.50	1.37	1.36	4.35	1.50	1.59	1.35	1.43	1.10
P-Tree1-5	9.61	0.66	0.65	0.66	0.64	0.62	4.11	0.57	0.71	0.51	0.60	0.30
P-Tree1-10	14.24	0.46	0.46	0.46	0.45	0.45	4.07	0.40	0.49	0.35	0.44	0.23
P-Tree1-15	20.52	0.38	0.38	0.38	0.37	0.37	3.98	0.36	0.45	0.31	0.40	0.18
P-Tree1-20	26.95	0.34	0.34	0.34	0.33	0.33	3.82	0.31	0.40	0.26	0.35	0.14

Uncommon Factors and Asset Heterogeneity in the Cross Section and Time Series

Lin William Cong, Cornell University
Gavin Feng, City University of Hong Kong
Jingyu He, City University of Hong Kong
Junye Li, Fudan University

(Futile?) Search for a Universal Model

- Einstein's quest for the "Theory of Everything."
- Sparsity for tractability and interpretation, but not necessarily universal models.
- Standard factor modeling in empirical asset pricing:

$$\begin{aligned} r_{1,t} &= \alpha_{1,t} + \beta_{1,1,t} f_{1,t} + \cdots + \beta_{1,k,t} f_{k,t} + \epsilon_{1,t} \\ &\vdots \\ r_{n,t} &= \alpha_{n,t} + \beta_{n,1,t} f_{1,t} + \cdots + \beta_{n,k,t} f_{k,t} + \epsilon_{n,t} \end{aligned}$$

  - LHS observations/assets are heterogeneous; grouped heterogeneity.
  - Burden all on RHS model estimation and selection.
  - Sorting/test asset construction for common factors & cross-cluster spread.
- *Joint* observation clustering and heterogeneous model selection:
  - Model selection on RHS: observations following one common factor model.
  - Observation clustering on LHS: split the panel allowing uncommon factors.
  - Data-driven, economically guided tree method preserving interpretability and guarding against overfitting.

Factor Selection in BCM Leaf Clusters

Prob.	N8	N18	N19		N48	N49	N50	N51	N52	N53	N54	N55
	Panel A: I(SVAR $\leq$ -0.2)I(ME $\leq$ 0.2)				Panel C: I(SVAR $>$ -0.2)I(SVAR $\leq$ 0.6)							
$>$ 0.95	MKTRF IVOL BETA	MKTRF IVOL BETA BAB SUE UMD STR	MKTRF IVOL BETA STR HML RMW SMB UMD		MKTRF STR IVOL	MKTRF STR HMLM	MKTRF IVOL	MKTRF IVOL ROE STR BETA SMB	MKTRF STR BETA SMB UMD HML RMW IA	MKTRF SMB BAB	BETA STR IVOL IMD	BETA STR IVOL UMD MKTRF BAB
$>$ 0.9	LTR	LTR						SUE				LTR
$>$ 0.7	BAB SUE					BETA BAB SMB	UMD	UMD		BETA	SMB	
In-Sample α	-0.36	0.36	0.37		-0.07	0.20	0.41	0.03	0.21	0.07	0.91	-0.10
Out-of-Sample α	0.45	0.33	0.14		1.12	0.18	0.38	0.08	0.08	-0.17	-0.13	-0.09

Prob.	N20	N21	N22	N23			N56	N57	N29	N30	N62	N63
	Panel B: I(SVAR $\leq$ -0.2)I(ME $>$ 0.2)						Panel D: I(SVAR $>$ -0.2)I(SVAR $>$ 0.6)					
$>$ 0.95	MKTRF IVOL HML SMB UMD	MKTRF HML SMB STR BETA BAB ROE HMLM	MKTRF SMB UMD	MKTRF SMB HMLM LIQ RMW ROE STR				UMD MKTRF SMB	MKTRF IVOL	MKTRF IVOL STR	MKTRF IVOL SMB	IVOL SMB STR QMJ
$>$ 0.9												
$>$ 0.7			BAB	REG						NI	NI	MKTRF
In-Sample α	0.07	-0.22	0.12	-0.24			1.61	-1.05	0.25	1.20	-0.88	-0.85
Out-of-Sample α	0.12	0.14	-0.02	-0.10			1.68	0.02	0.38	0.38	-0.14	0.27

Common and Uncommon Factors

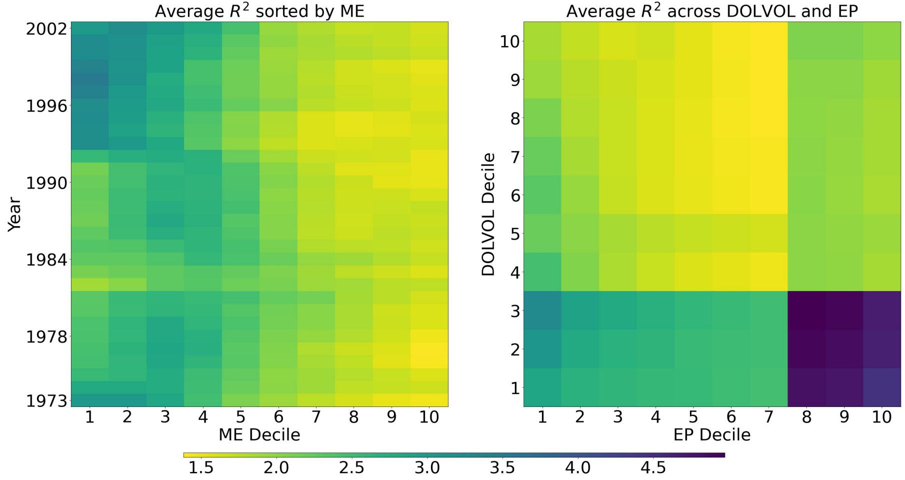
- Common factors:
  - MKTRF has been selected with a probability of 100% in most leaf clusters except for N54, N56, and N63.
  - Size (SMB) is also important, selected with probability $>$ 70% in 13 out of 21 clusters.
  - STR and BETA are important, $>$ 70% in 12 and 10 clusters.
  - Value (HML) is $>$ 70% only for 4 out of 21 clusters.
- Odd-ball out: What makes N56 so special? No factor is selected!
  - Stocks with highest SVAR and lowest EP, MOM, and ME, basically all junk stocks.
- Factor usefulness and grouped heterogeneity in disguise.
- Evaluate Economic Fitness by Leaf Clusters: Agnostic performs slightly better; N21, N30, N54 have $R^2 > 24\%$ OOS.

Mosaics of Predictability

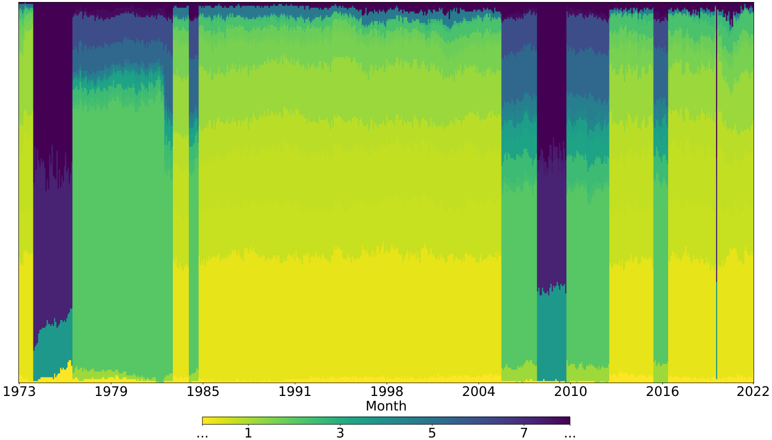
Lin William Cong, Cornell University
Gavin Feng, City University of Hong Kong
Jingyu He, City University of Hong Kong
Yuanzhi Wang, City University of Hong Kong

Goal-Oriented Clusters of Differential Return Predictability

Return predictability taken as attribute of predictors or models; but is unobervable, heterogeneous, and time-varying.



Mosaics of Predictability by Calendar Months



- S&P 500 dividend yield and Pastor-Stambaugh liquidity predicts regime switches.
- Some regimes correspond to events (Oil Crisis, COVID-19, etc.).

Model Disagreement Anomaly & Predictability-Based Investment:

Heterogeneity disagreement and mispricing $R^2_{\text{CMG},j} = R^2_{\text{C},j} - R^2_{\text{G},j}$
Large and significant OOS Alphas, dominated by long leg.

	1973 - 2002 (in-sample)			2003 - 2022 (out-of-sample)		
	T3	B3	T3 - B3	T3	B3	T3 - B3
			Panel A: Performance			
Avg (%)	2.75	0.40	2.35	2.10	1.07	1.03
Ann. SR	1.56	0.25	2.33	1.18	0.52	0.75
			Panel B: Unexplained monthly alphas (%)			
CAPM	2.35***	-0.01	2.35***	1.18***	0.03	1.15***
FF3	1.94***	-0.32**	2.26***	1.22***	0.07	1.14***
FF5	1.89***	-0.37***	2.26***	1.31***	0.11	1.20***
FF5+MOM+IVOL	2.09***	-0.24*	2.33***	1.39***	0.19	1.20***
Q5	2.19***	-0.05	2.24***	1.43***	0.09	1.34***
BS6	1.84***	-0.32**	2.15***	1.37***	0.10	1.26***
DHS3	2.68***	0.17	2.51***	1.44***	0.26	1.19***
SY4	2.05***	-0.26*	2.31***	1.89***	0.46	1.43***

Generative Modeling Using LLMs and Beyond

- Modern AI: goal-oriented search in enlarged modeling space to provide end-to-end solutions to complex problems.
- **Deep RL as heuristic search and panel trees as (interpretable) greedy search; both generative GOALS providing better solutions.**
  - AlphaPortfolio for Investment Management
  - AlphaManager for corporate finance
  - Panel Trees for test asset and latent factor generation
  - Panel trees for modeling grouped heterogeneity
- GenAI enables data-driven counter-factual equilibrium analyses.
  - Data-Driven Generative Equilibrium (DDGE)
  - Online lending market in the presence of LLMs
- Understanding GenAI and its interaction/complementarity with humans.
  - Behavioral economics of GenAI agents
  - Causal human+machine learning
  - AI copilots, etc.

AI for Business Research & Social Sciences

- What is special about business and social science research?

AI for Business Research & Social Sciences

- What is special about business and social science research?
- I. Complexity, Heterogeneity, and Purpose-Driven
  1. Probabilistic, context-specific (heterogeneous), mergent interactions.
  2. Non-stationary, variable, adaptive, etc.
  3. Value-laden with high ethical risks.
  4. (Hetergoeneous) goal-oriented; stakeholder interactions and agencies.
- II. Methods and Approach
  1. Multi-modal, qualitative, unstructured, not strictly ordered, non-systematic.
  2. Interpretation and causality.
  3. What-if questions. The future of economics lies not in explaining the past or predicting the world, but in re-imagining and generating the future.
- III. Data limitations:
  1. Scarce (mostly observational), missing, etc.
  2. Privacy-sensitive.
  3. Endogenous and time-varying data generation.
- Limitations of extant generative modeling in economics or social sciences (e.g., DSGE models, Computable General Equilibrium Models, Structural Equation Models, Agent-Based models).

GenAI for Social Sciences

- Generative AI comes to the rescue:
  1. Enlarged modeling space with end-to-end architecture: better latent representation, richer contextual details and variability, and data-driven heterogeneity modeling (VAEs, GANs, Diffusion, etc.)
  2. Realistic data generation: augmenting qualitative and quantitative data, privacy-preserving synthetic data, DDGE analysis, etc.
  3. AI-Agent-Based Simulation/Modeling:
     - Human proxies in surveys and experimentation.
     - Computable ground-up and in aggregate; accommodates interactions, adaptation, heterogeneity and emergence.
     - Counterfactual and causal analysis; policy simulation/stress-testing.
  4. Other new developments:
     - Sparse AI, reasoning, causal ML, etc.
     - Multi-objective generative model, bias detection tools, etc.
     - Data (human-generated and historical) VS. experience (machine generated).
     - Understanding AI agents and their interactions with humans.
- Unsupervised/supervised $\rightarrow$ goal-oriented (i) can flexibly guide the generation of end-to-end solutions; (ii) generates counterfactual data and enables data-driven dequilibrium analysis; (iii) behavior of AI($+$humans).

Q&A and More on AI, Digital Economy, and Financial Technology

FinTech@Cornell and DEFT Lab
https://www.comm.business.cornell.edu/l/296062/2021-09-02/7c8l15