Assessing the Benefits of Optimized Agentic AI Systems for Asset Pricing

Ralph S.J. Koijen and Bradford Levy

2026-07-17

CHICAGO BOOTH

The University of Chicago Booth School of Business

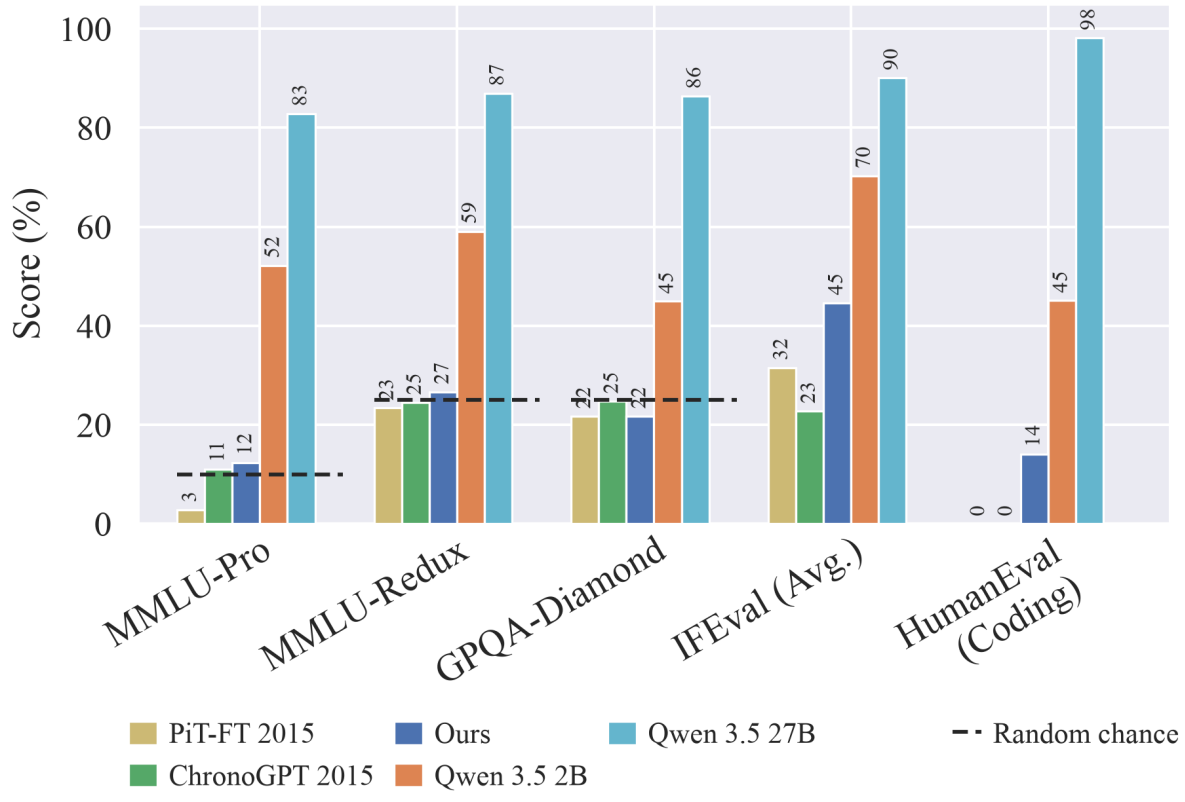
Motivation

- LLMs and ***agentic tooling*** have made rapid progress across a variety of domains/benchmarks in recent years
  - Language understanding, coding, mathematics, etc.
  - Especially true once optimization is applied to model, prompt, etc.
  - ***Particularly in domains with strong verifiability***
- Unclear how much asset management can benefit from this
  - **Look-ahead bias** – weak models or leakage exacerbated by optimization pressure
  - **Reflexivity** – value of information or technology today is likely to change tomorrow

Look-ahead Bias: Informational

Access to knowledge not available to a decision maker at the time

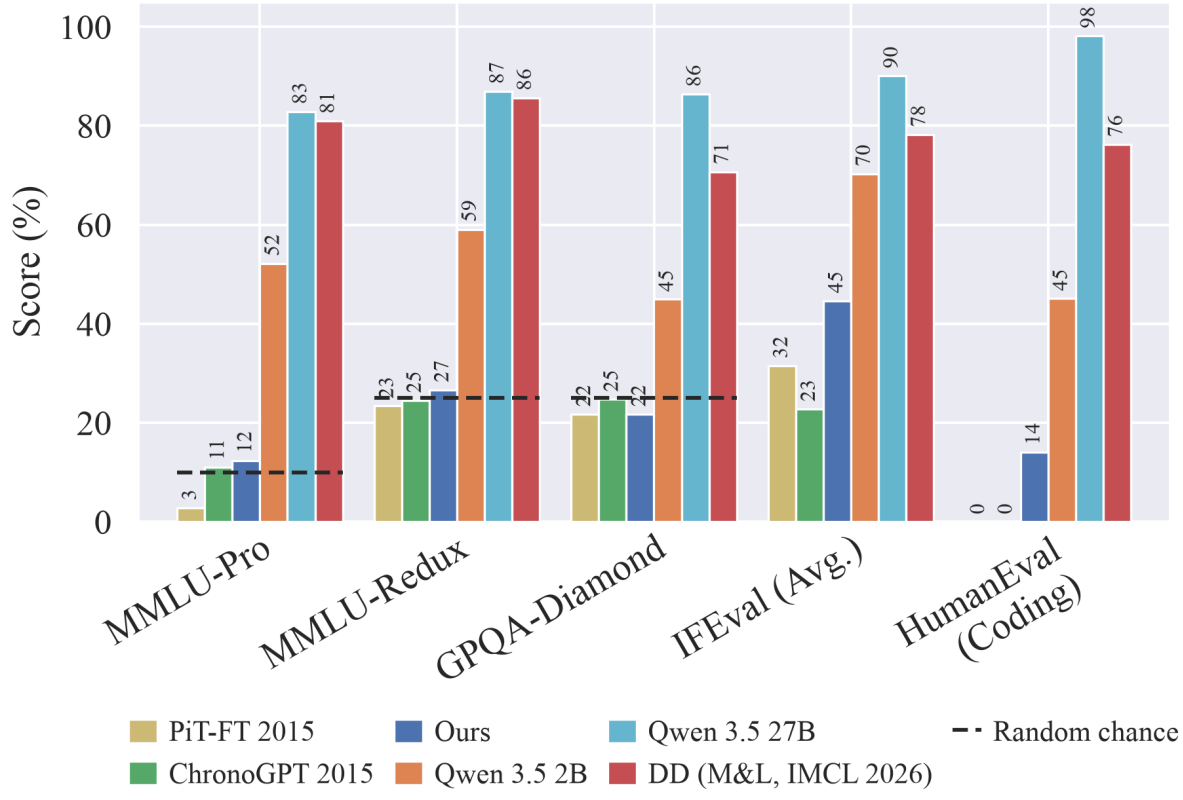
- Point-in-Time models
  - "2015 model" → trained only on content available in 2015 or earlier



Ours: ***fin-ai-lab/aux-2015***

Look-ahead Bias: Informational

- Alter distributional properties of data or model:
  - Masking (Glasserman and Lin, 2023; Engleberg et al, 2025)
  - Unlearning/Divergence Decoding (Merchant and Levy, 2026 ICML; frontiertopit.com)



Ours: ***fin-ai-lab/aux-2015***

Look-ahead Bias: Technological

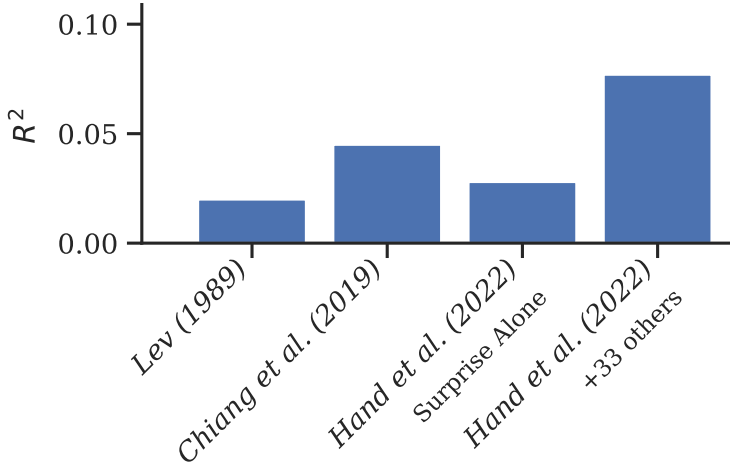
- Even if a lab did produce a dated frontier model…
  - Alec Radford released "Talkie" a 13B model trained on ~1930s era content
  - Host of techniques to "convert" a frontier model to PiT at frontiertopit.com
- Transformer architecture was not publicly disclosed until 2017
- Infrastructure—software and hardware—to train a frontier model did not exist until 2020s
- Even back tests/archival research using point in time models are likely contaminated

Motivation

- LLMs and ***agentic tooling*** have made rapid progress across a variety of domains/benchmarks in recent years
  - Language understanding, coding, mathematics, etc.
  - Especially true once optimization is applied to prompt, model, etc.
  - ***Particularly in domains with strong verifiability***
- Unclear how much asset management can benefit from this
  - **Look-ahead bias** – weak models or leakage exacerbated by optimization pressure
  - **Reflexivity** – value of information or technology today is likely to change tomorrow
- **How can we benchmark agentic AI in asset pricing?**
  - What can we learn from an AI model that "solves" the benchmark?

Benchmarking Agentic AI in Asset Pricing

- 5+ decades of explaining earnings announcement returns



- Aside from quantitative data, EAs involve lots of qualitative data
  - LLMs, VLMs, etc. excel at processing unstructured qualitative data

- 2,000+ EAs per quarter → statistical power over a short horizon
  - Can run the benchmark in real-time **purely out-of-sample**

The Benchmark

Goal: Evaluate the extent to which models can explain the effects of an information event on the focal firm (and peer firms)

- Requires some understanding of cause and effect, economic forces, networks

Implementation: Given the content of an information event, explain the return(s) near the event, e.g., a 1-day window

- Earnings Announcements (EAs)
- Given the EA (full transcript or summary), predict the market reaction
- Apply LLMs/agents in real-time for unbiased performance estimates
- Establish this as a benchmark where anyone can play

The Benchmark: Setup

Consider a stylized factor model:

$$r_{i,[t,t+h]} = \boldsymbol{\beta}_{i,t}^{\top} \mathbf{F}_{[t,t+h]} + \epsilon_{i,[t,t+h]}$$

where we are interested in the effect of an information release $\omega_{i,t}$:

$$\epsilon_{i,[t,t+h]} = g(\omega_{i,t}, \Omega_{t-}) + \nu_{i,[t,t+h]}$$

The benchmark seeks to evaluate approximations of $g(\cdot)$, i.e., the mapping from content/information to prices.

The Benchmark: Evaluation

$$\epsilon_{i,[t,t+h]} = g(\omega_{i,t}, \Omega_{t-}) + \nu_{i,[t,t+h]}$$

Can measure abnormal return $\epsilon_{i,[t,t+h]}$ but $g(\cdot)$ is unobservable

Given an approx. $\hat{g}(\cdot)$ and OOS observations $\mathcal{E} = \{(i_k, t_k)\}_{k=1}^{N}$, eval based on the R^2 from a "standard" earnings surprise regression:

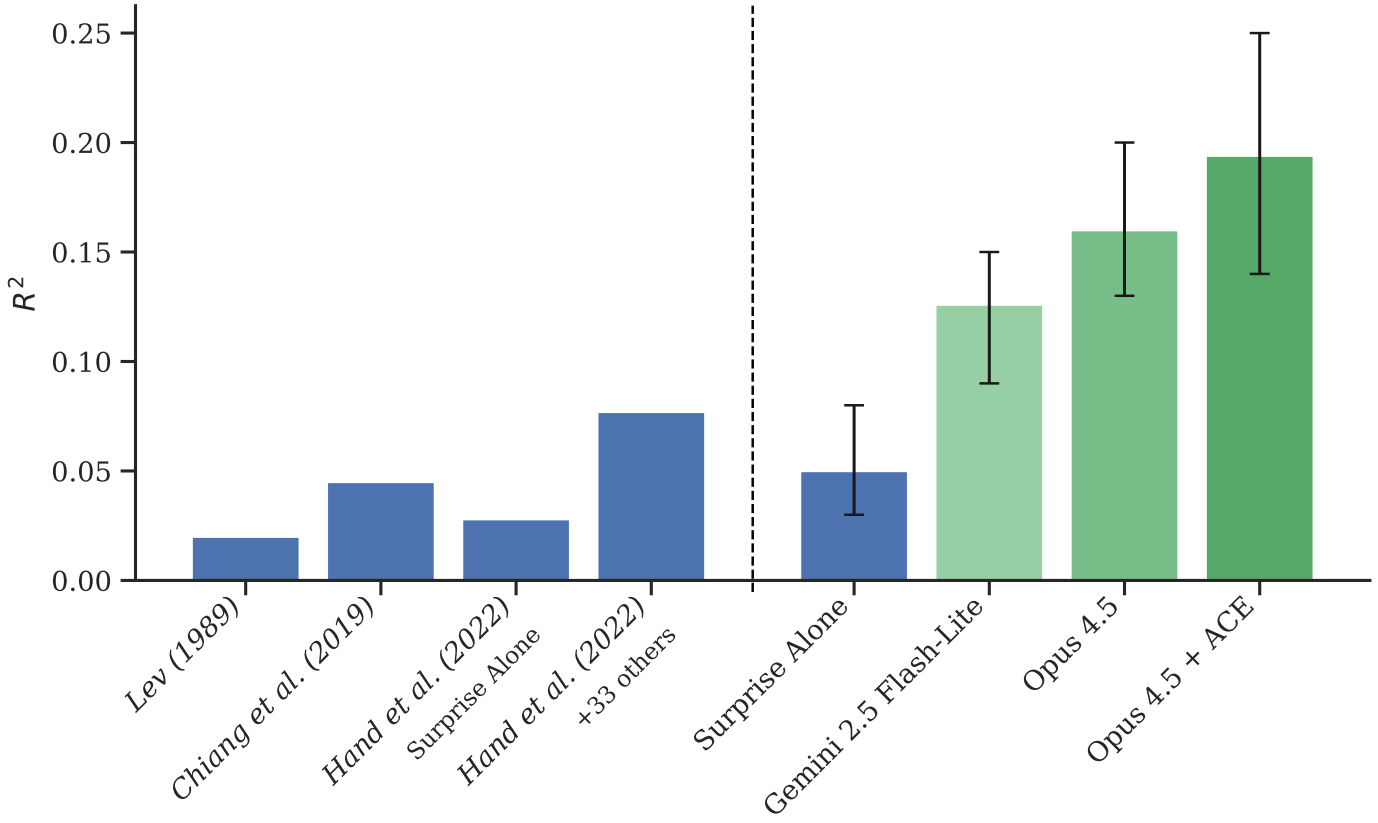
$$\epsilon_{i,[t,t+h]} = \alpha + \gamma \hat{g}(\omega_{i,t}, \Omega_{t-}) + \boldsymbol{\theta}^{\top} \mathbf{X}_{i,t} + u_{i,[t,t+h]}$$

This seems natural (to us). Upper-bound on R^2 is:

$$\overline{R}^2 = \frac{Var(g)}{Var(g) + Var(\nu)}$$

which for economic reasons we expect to be high (>> 2-8%)

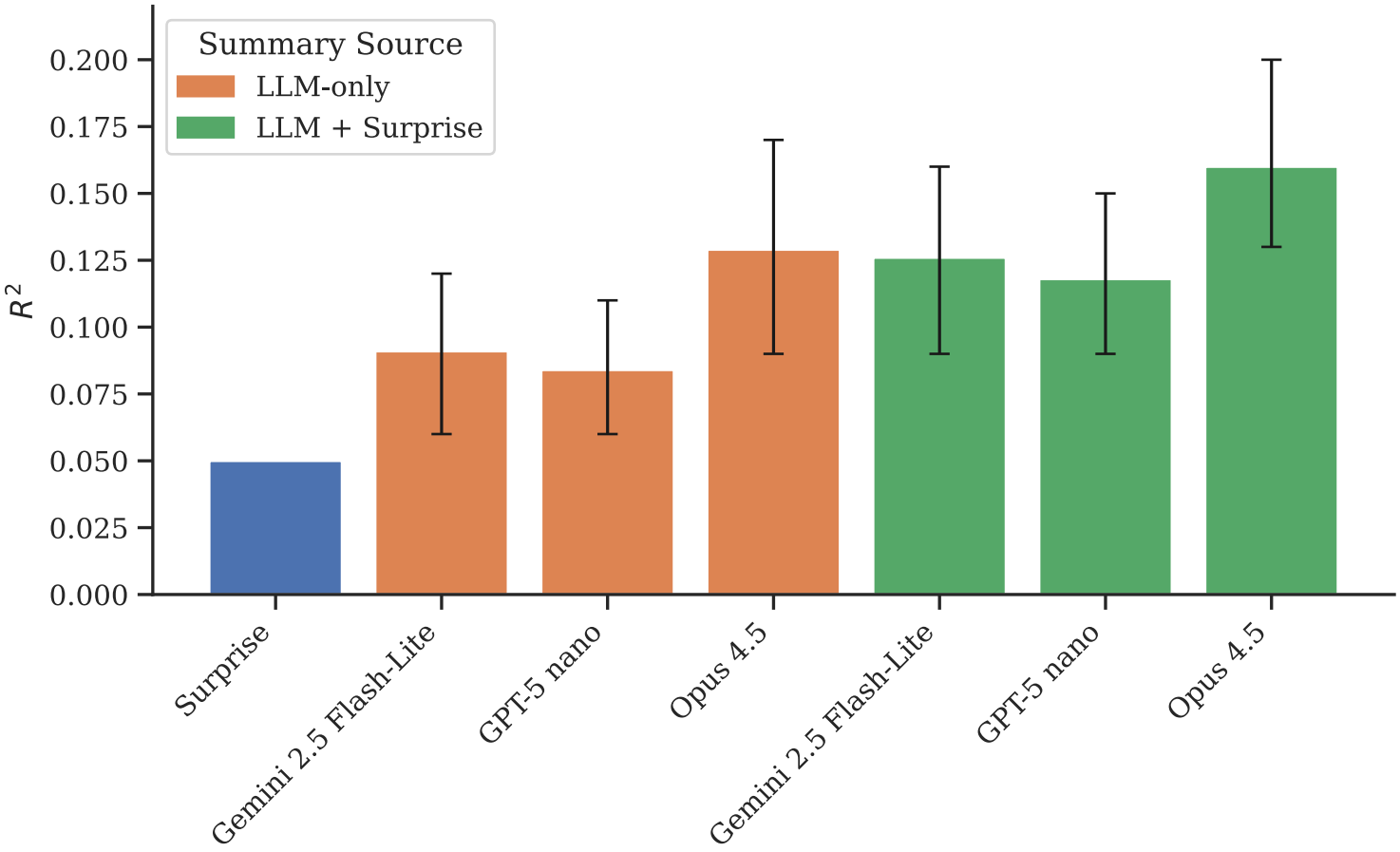
The Benchmark: Results of SOTA Applications



Koijen and Levy (2026)

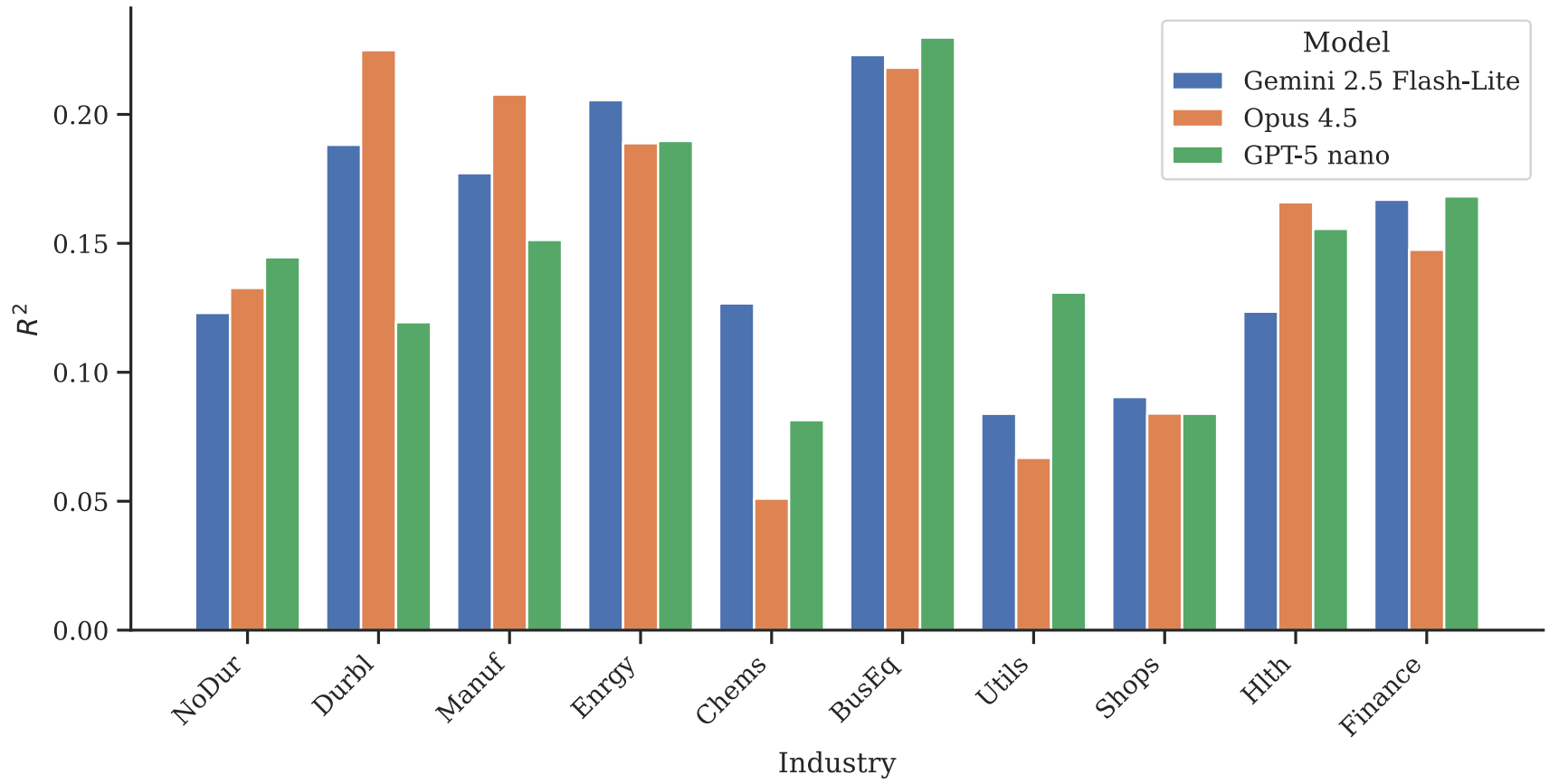
Baseline Results

- Apply LLMs to the ***full transcript in real-time***
- LLM and surprise are complementary, R^2 of 11-16%



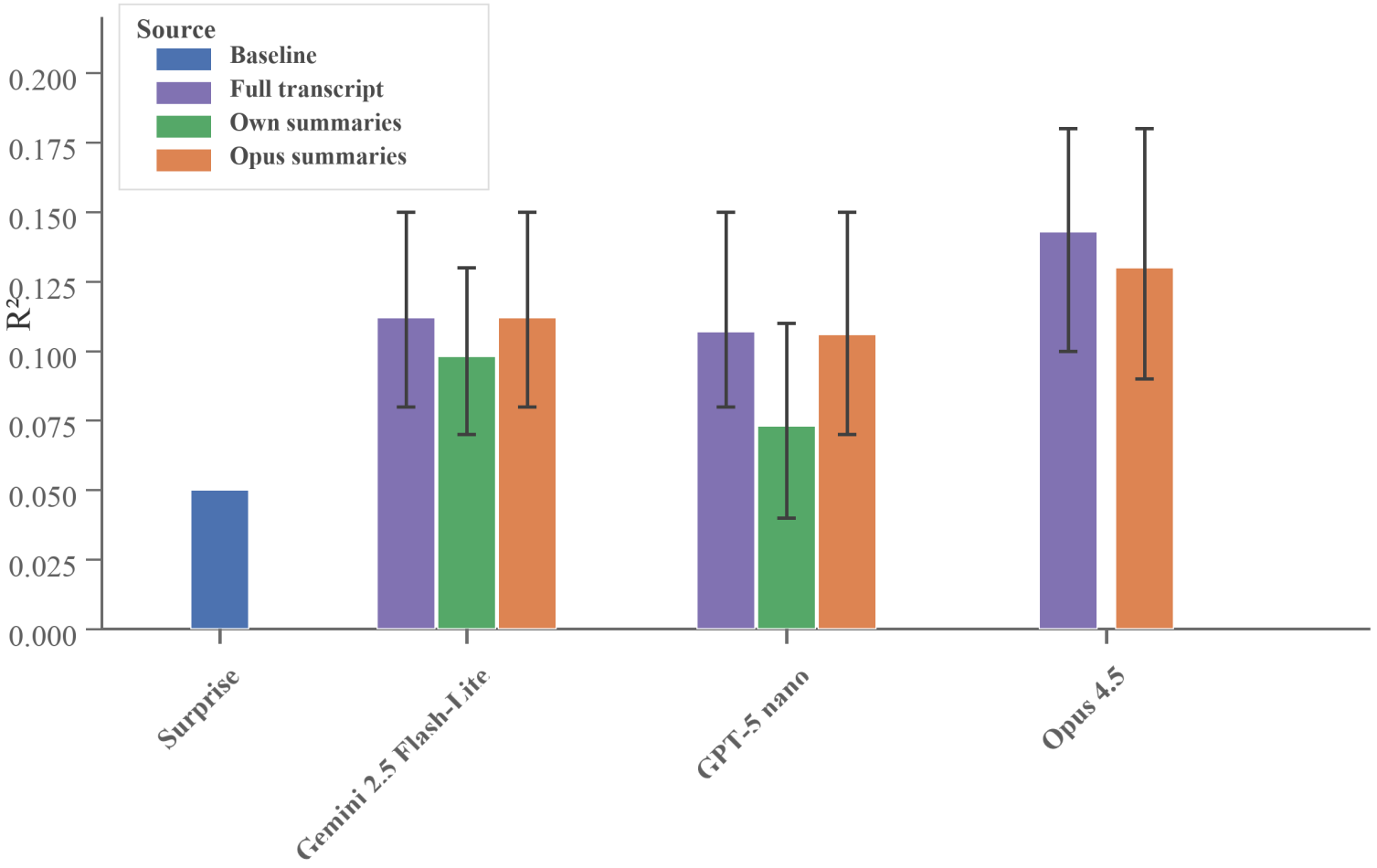
Industry and Model Heterogeneity

- Across industries, certain models are better than others
- Suggests benefits to a multi-model approach



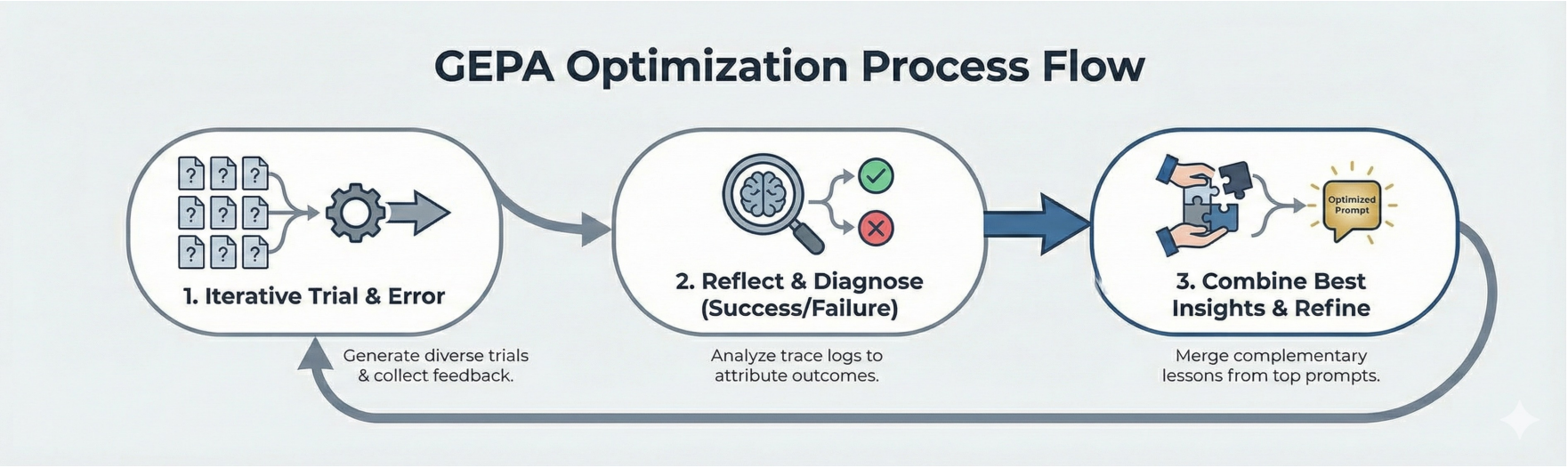
Results for Summarized Transcripts

- Summaries are ~28x shorter → Optimization cheaper, faster
- Marginal loss of explainability: R^2 is 11-15% versus 11-16%



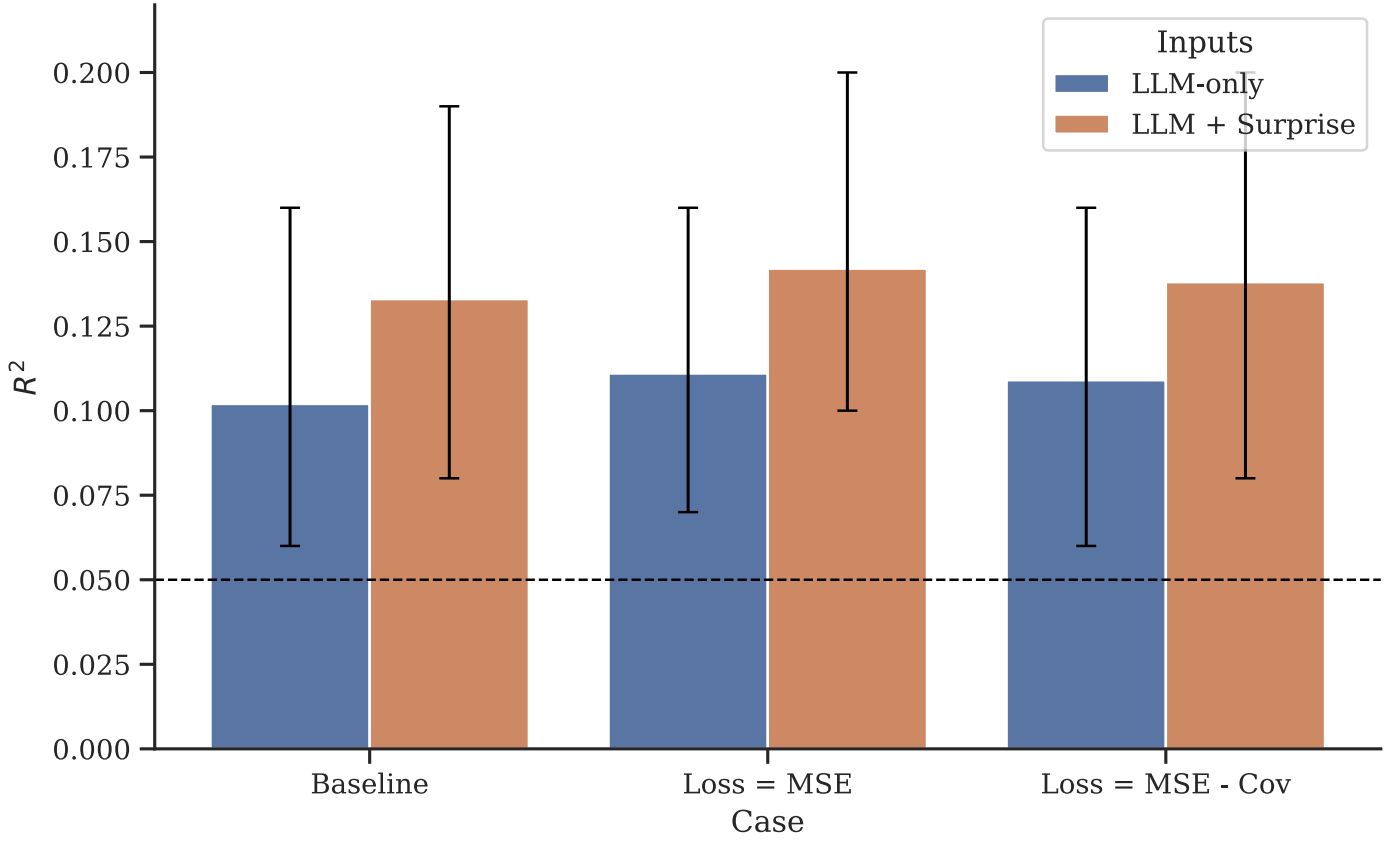
Optimization: GEPA

- Refines prompts using feedback derived from trial and error
- Use past trials to diagnose root causes of successes and failures
- Combines the best insights across a diverse pool of top-performing approaches



Optimization: GEPA Results

- During training, GEPA reduces the objective (~30% MSE decrease)
- However, it mainly debiases the LLM's predictions rather than increasing the proportion of the variance explained



Interpretability: What did GEPA learn?

The Core Directive: Assume "Good" is "Priced In"

Your previous predictions have consistently failed by overestimating the market's excitement for strong operational results.

* **The Baseline is Success:** Public companies are expected to grow, beat estimates, and improve margins. A company reporting record profits and 20% growth is usually just meeting the high expectations already built into the stock price. This is a **Neutral (0.40 - 0.60)** result, not a positive one.

* **Neutral = The Graveyard of Good News:** If a company reports a "beat and raise" that is merely incremental, or if they report strong earnings but flat guidance, the stock will likely not move or may even drift lower as traders "sell the news."

* **The Bar for "Up" is Extremely High:** To justify a prediction > 0.75, the news must be *shocking* (e.g., doubling guidance, a massive surprise buyback >10% of market cap, or a breakthrough product launch *today*).

Interpretability: What did GEPA learn? (Cont'd)

Sector-Specific Heuristics (based on past errors):

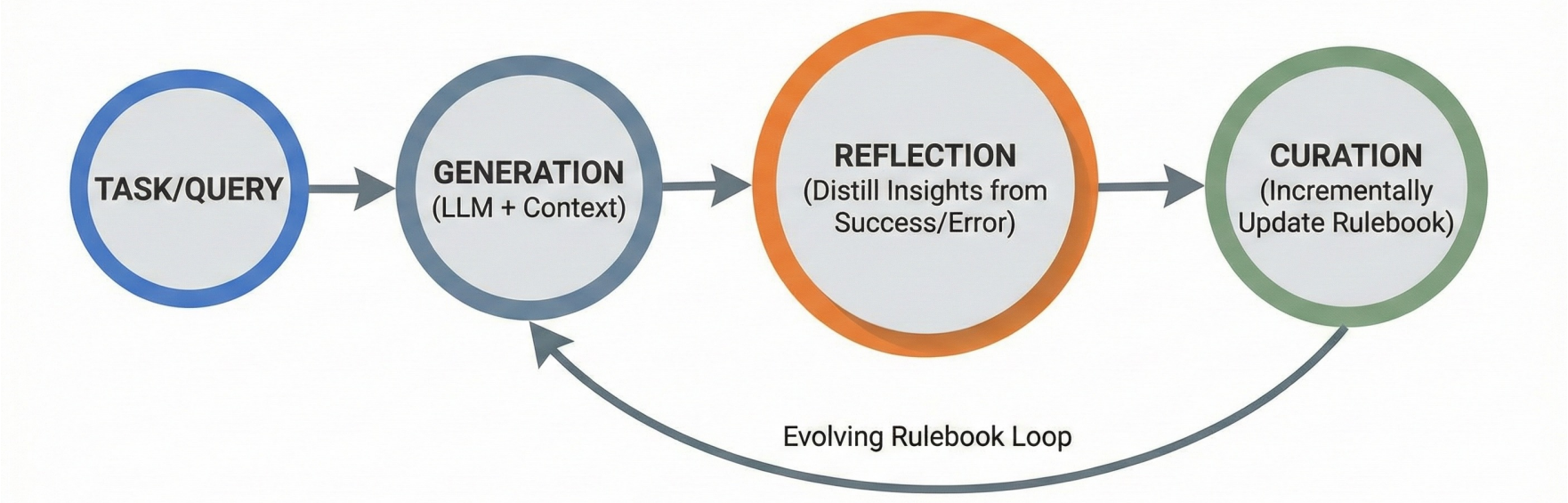
1. **Telecom/Utilities (e.g., AT&T):** These are defensive yield stocks. High subscriber growth (net adds) is often ignored if *churn* ticks up even slightly or if free cash flow guidance isn't raised. The market fears competition more than it rewards growth here. *Heuristic: If churn rises, cap prediction at 0.50.*

2. **High-Growth/Tech (e.g., Tesla):** These stocks have massive premiums. Operational improvements (cash flow, margins) are expected. "Roadmap" items (future products years away) are often discounted as "vaporware" in the short term. *Heuristic: Discount long-term promises. Only value immediate financial delivery.*

3. **Financials/Insurance (e.g., Travelers):** Buybacks are standard operating procedure, not surprises. A $3B buyback is often just capital maintenance. "Underwriting discipline" (shrinking volume to maintain price) is often viewed as a growth headwind. *Heuristic: Without a massive dividend hike or distinct catalyst, stick to 0.40-0.50.*

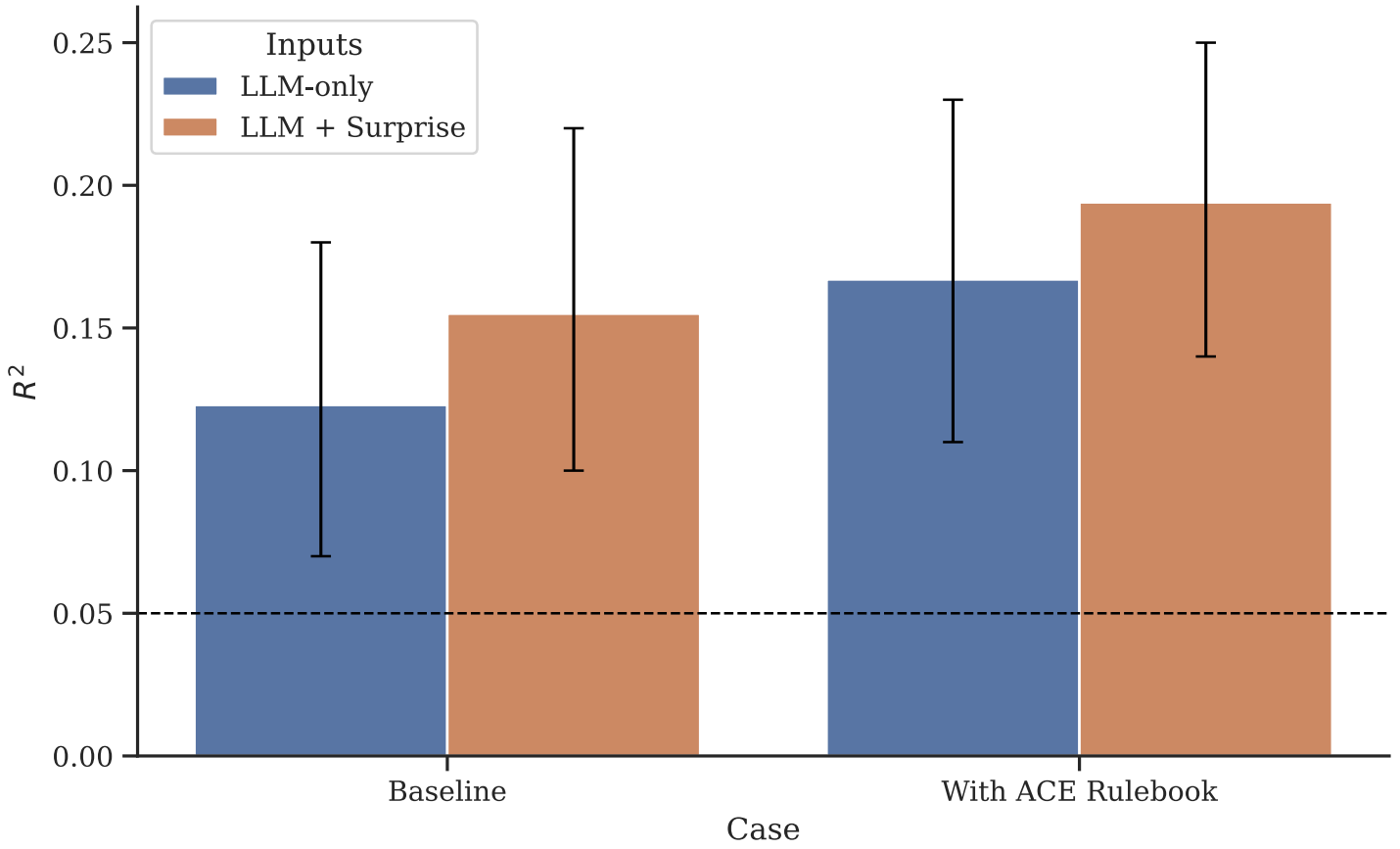
Optimization: Agentic Context Engineering (ACE)

- Imbues models with domain-specific expertise through context rather than expensive fine-tuning
- ACE treats prompts/context as an evolving rulebook of strategies
- Reflector actively distills insights from past successes and errors
- Incrementally updates rulebook to preserve details



Optimization: ACE Results

- Stronger effect on R-squared than for GEPA
- Lifts R-squared to ~17% *(19%)* for Opus 4.5



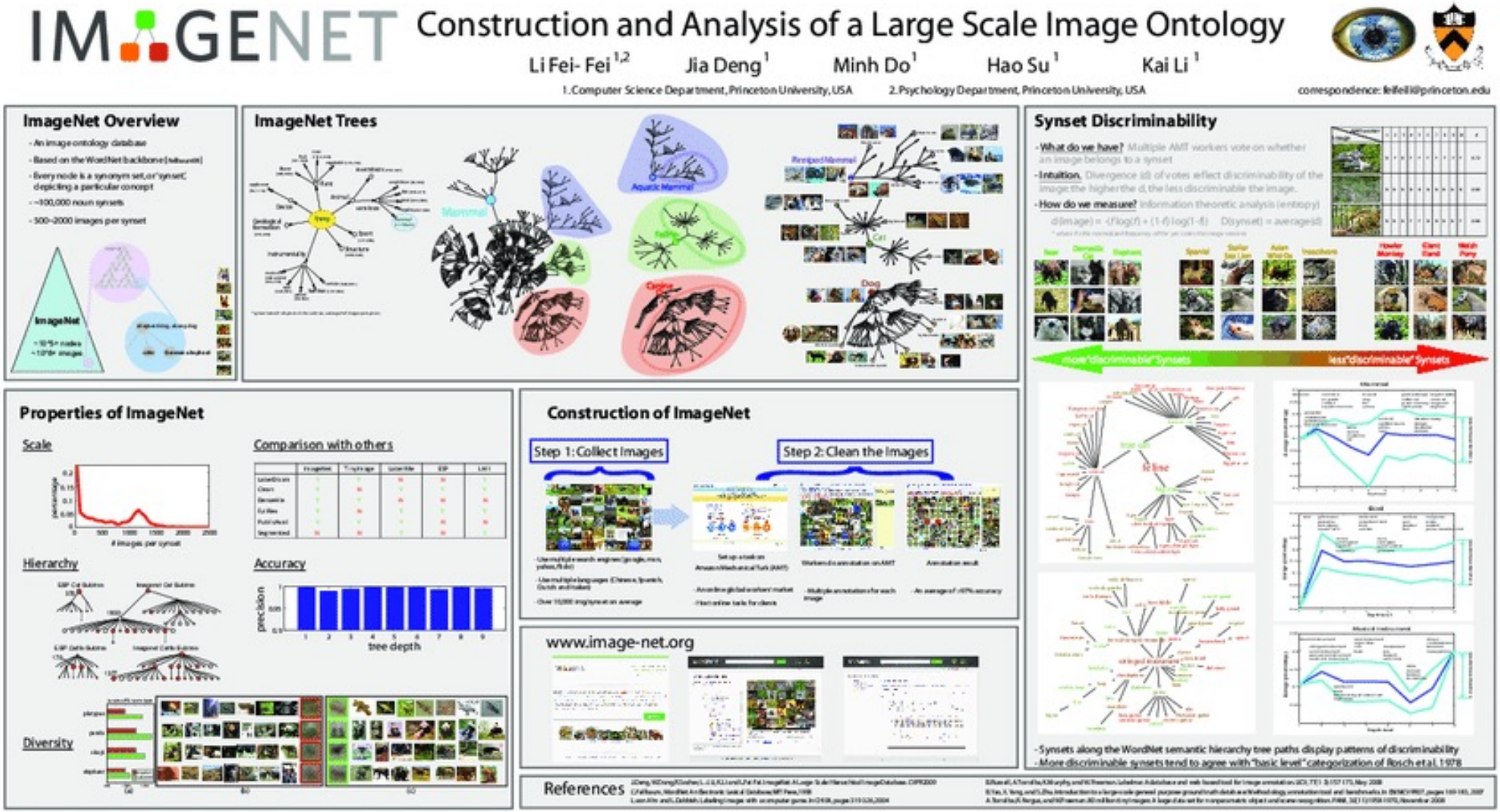
Interpretability: What did ACE learn?

* When a pre-profit company reports strong revenue growth (>30% YoY) BUT operating losses remain flat or worsen (EBITDA/operating loss not narrowing proportionally), apply 0.10-0.20 range. Markets demand that high growth translate to financial improvement. Amplified when: (1) forward guidance suggests growth momentum peaking (flat sequential); (2) in hot theme sector (LiDAR, autonomy, quantum, etc.); (3) customer conversion rates below 15%.

* When management uses cautionary forward language ("aggressive consensus," "gradual recovery," guiding to low end), apply 0.35-0.50. When forward quarter guidance represents >50% deceleration from recent organic growth, apply 0.10-0.25. When one-time items inflate headline EPS with no forward acceleration, apply 0.10-0.20. CRITICAL: premium brand guidance below framework + DTC decline = 0.05-0.15.

* When record or near-record margins include acknowledged ***non-recurring benefits*** (precious metals, inventory revaluations, one-time mix) AND there is a concrete operational execution failure in the same quarter (equipment downtime, supply disruption, facility issue) AND a major geographic segment is declining >15% from external headwinds, apply 0.10-0.20 range regardless of multi-sector demand narratives.

On the Importance of Benchmarks: ImageNet



Example of Benchmark Impact: ImageNet

- ImageNet "Full" – 14M images spanning 21K classes
  - Absolutely massive at the time. Datasets at the time spanned tens of classes
- ImageNet Large Scale Visual Recognition Challenge (ILSVRC) was run from (2010-2017)
  - 2010-2011: Error rate was around 30%
  - 2012: AlexNet reduced this to 15.3% (absolute shock at the time)
  - 2013: Refinement of AlexNet dropped it to 11.2%
  - 2015: ResNet dropped it to 3.5% (human baseline was ~5.1%)
- In five years, ILSVRC was basically solved
  - Some overfitting, see "Do ImageNet Classifiers Generalize to ImageNet?"

Conclusions and Next Steps

- First true OOS test of Optimized Agentic AI on a benchmark of broad interest to practice and academics
  - Strong emphasis on interpretability (can read the reasoning)
  - SOTA optimization & models → significant increase in the R^2. But still < 20%
- Benchmark launched in partnership with Optiver on Monday
  - 200+ people have joined so far, 40+ active submissions making predictions
  - Join/follow at explainingmarkets.ai

Conclusions and Next Steps

QUARTERLY LEADERBOARD

The standings, *Q4 2025.*

Every active submission, ranked by how well its predictions explain realized returns around earnings announcements.

THIS BOARD

Period	Q4 2025
Scorable events	1,849
Status	Finalized on Jan 8, 2026
Ranked submissions	5
Ranking metric	R^2 of explained returns
Updated	Nov 17, 2025, 21:30 UTC

Q4 2025 FINAL | Q1 2026 FINAL | Q2 2026 FINAL | Q3 2026

RANK ▲	SUBMISSION	R^2	PREDICTIONS
1	*Gemini Flash Lite — Full Transcript* BASELINE	0.112	1,849 (100%)
2	*GPT-5-nano — Full Transcript* BASELINE	0.107	1,849 (100%)
3	*Gemini Flash Lite — Summary* BASELINE	0.098	1,849 (100%)
4	*GPT-5-nano — Summary* BASELINE	0.072	1,849 (100%)
5	*Earnings Surprise* BASELINE	0.049	1,849 (100%)

FINAL RESULTS — SCORES NO LONGER CHANGE. · HOW SCORING WORKS

Conclusions and Next Steps

SUBMISSIONS

GPT-5-nano — Summary · Live

DEMO MODE

This is what your team's portal looks like once a submission is live.

Sign up to participate →

Every prediction this submission made, matched against realized market outcomes.

Scored data as of 11/17/2025, 4:30:00 PM — final.

Q4 2025 | Q1 2026 | Q2 2026 | Q3 2026 | Refresh

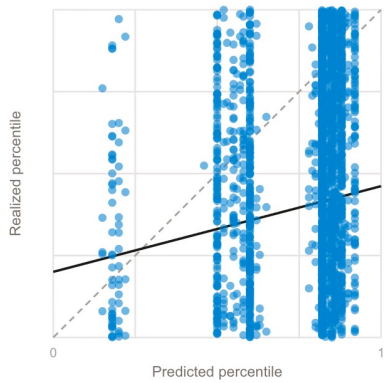
Quarter scoring — Q4 2025

OLS of realized percentile on predicted percentile over 1849 scored predictions. R^2 is the leaderboard ranking metric; α, β, and MSE are visible only to your team.

R^2	SLOPE (BETA)	INTERCEPT (ALPHA)	MSE	SCORED PREDICTIONS
0.072	0.26	0.20	0.077	1849

Predicted vs. realized

Each point is one scored prediction. The dashed line is perfect calibration; the solid line is your fitted regression — its R^2 is what ranks you.



Search ticker…

EVENT DATE ▾	TICKER	FISCAL PERIOD	PREDICTED	ACTUAL	ERROR	MEDIAN |ERROR|	SURPRISE	CAR1	STATUS
Nov 14	TOMZ	Q3 2025	0.60	0.20	-0.40	0.36	-1.22%	-7.6%	Scored
Nov 14	DFLI	Q3 2025	0.18	0.54	+0.36	0.31	+45.66%	-0.2%	Scored
Nov 14	BEEM	Q3 2025	0.57	0.63	+0.06	0.17	+4.95%	+1.6%	Scored
Nov 14	QUBT	Q3 2025	0.88	0.86	-0.02	0.06	+0.57%	+9.5%	Scored

Conclusions and Next Steps

- First true OOS test of Optimized Agentic AI on a benchmark of broad interest to practice and academics
  - Strong emphasis on interpretability (can read the reasoning)
  - SOTA optimization & models → significant increase in the R^2. But still < 20%
- Benchmark launched in partnership with Optiver on Monday
  - 200+ people have joined so far, 40+ active submissions making predictions
  - Join/follow at explainingmarkets.ai
- **Live project** – plan to make updates every earnings season
  - Extend the context beyond the current call (e.g., contemporaneous calls, own history, industry or agentically-defined peers)