

Discussion:  
Retirement Pension Portfolio Choice with AI  
Assistance: An Experimental Study

by Hongseok Choi, Jeongbin Kim, Matthew Kovach, Kyu-Min  
Lee, **Euncheol Shin**, and Hector Tzavellas

Duk Gyoo Kim

2026 KEA Law and Econ

# Quick Recap

## Research Question:

- ▶ How do LLMs recommend financial decisions?
- ▶ How do investors react to these recommendations?

## Experimental Design:

- ▶ 400 individuals with DC pension plans.
- ▶ Allocate endowment across 11 pseudo pension products.
- ▶ Within-subject design (Initial Choice → AI Recommendation → Revision).

## Key Findings:

- ▶ Participants adjusted portfolios marginally (distance reduced by ~21%) towards the AI recommendation.
- ▶ GPT-4 Turbo suggested aggressive (high risk) portfolios, while GPT-4o suggested conservative (low risk) ones.
- ▶ Neither model significantly improved the Sharpe ratio.

## Things I Like (1/2): Design & Relevance

### Timely Policy Implications

- ▶ The study highlights the “Black Box” risk: different LLMs imply different fiduciary stances (Aggressive vs. Conservative) without explicit user prompting.
- ▶ Effectively questions liability: If AI increases risk without improving efficiency (as Turbo did), who is responsible?

### Realistic Experimental Setting

- ▶ Payoffs are tied to a simulation of ‘55–Age’ year returns, aligning incentives with the long-term nature of pensions.
- ▶ Using 11 actual market products mimics the real choice architecture faced by pension holders.

## Things I Like (2/2): Behavioral Identification

### Clean Identification of “Algorithm Aversion”

- ▶ The use of Euclidean distance (before vs. after recommendation) provides a precise metric for trust and anchoring.
- ▶ Finding that users move ~21% regardless of the model suggests a “nudge” mechanism rather than genuine information updating.

### The Null Result on “Explainability”

- ▶ Providing text-based rationale did not significantly increase adoption or trust.
- ▶ This challenges the assumption that simply adding “explanations” (XAI) is sufficient to cure hesitancy or liability concerns in financial advising.

## Things I Want to Know More (1/3)

### Prompt Sensitivity vs. Intrinsic Bias

- ▶ The divergence between Turbo and Omni is interesting.
- ▶ Is this a fundamental model difference, or an artifact of how we parsed the prompt?
- ▶ Robustness checks with distinct prompts (e.g., “Maximize Sharpe Ratio” vs. “Minimize Volatility”) to disentangle instruction-following from model bias.

### Content Analysis of the Rationale

- ▶ The regression shows the *presence* of rationale didn't matter.
- ▶ What about the *quality*? Did Turbo justify high risk by downplaying volatility or emphasizing returns?
- ▶ NLP analysis of the generated text might reveal why users found the explanations unpersuasive.

## Things I Want to Know More (2/3)

### Interaction with Financial Literacy

- ▶ Did low-literacy users rely *more* on the AI (substitution) or *less* (aversion)?
- ▶ Interaction terms in the regression ( $\text{Literacy} \times \text{Treatment}$ ) to see if AI acts as an equalizer for less skilled investors.

### Welfare Analysis beyond Efficiency

- ▶ The paper focuses on the Sharpe ratio (Efficiency).
- ▶ Based on elicited risk preferences, you may want to check how AI suggestions are aligned to the ideal direction of change.
- ▶ Did the AI move users closer to their personal utility maximum?

## Things I Want to Know More (3/3)

### Does AI Eliminate “Wrong” Choices?

- ▶ The menu contains **strictly dominated** products. For example, P8 dominates P9 (higher returns with the same risk), and P11 dominates P10 (higher returns and lower risk).
- ▶ Define  $Share_{Dominated}$  as the % of wealth allocated to P9 and P10. Does AI successfully reduce allocations to these choices?
- ▶ If AI reduces  $Share_{Dominated}$  to zero, it provides clear “sanitation” value.