Article

Not peer-reviewed version

Generative AI as an Investment Advisor: Same Client, Different Advice

Nicolo Agliata and Tim Hasso *

Posted Date: 14 May 2026

doi: 10.20944/preprints202605.0909.v1

Keywords: generative AI; large language models; robo-advisors; goals-based investing; algorithmic bias; audit methodology; conjoint; fintech

 doi:10.20944/preprints202605.0909.v1

Article

Generative AI as an Investment Advisor: Same Client, Different Advice

Nicolo Agliata and Tim Hasso *

Bond University; 14 University Drive, Robina QLD 4226, Australia

* Correspondence: thasso@bond.edu.au; Tel.: 61 7 5595 2288

Abstract

Generative artificial intelligence (GAI) is increasingly embedded in personal financial, yet little is known about how models make recommendations using financial information and demographic cues. This study audits three frontier GAI models, GPT 5.5, Gemini 3.1 Pro, and Claude Opus 4.7, using a full-profile conjoint experiment in which each model evaluated the same 1,000 hypothetical investor profiles and selected among standardized conservative, balanced, and aggressive portfolios. Investor profiles systematically varied attributes, including risk tolerance, time horizon, goal type, income, and age, gender, ethnicity, marital status, and employment type. Ordered logistic regressions and matched-profile comparisons show that all three models base recommendations primarily on legitimate financial inputs, especially risk tolerance and time horizon. Gender and ethnicity do not significantly influence recommendations, although age affects all models and marital status affects ChatGPT. However, the models are not interchangeable: they differ significantly in overall risk appetite and in how they translate risk tolerance, time horizon, goal type, and age into portfolio choices, with economically meaningful differences in predicted recommendations for identical clients. These findings suggest that contemporary GAI investment advice exhibits limited evidence of conventional demographic bias but introduces a distinct form of platform risk arising from model-specific advisory logic.

Keywords: generative AI; large language models; robo-advisors; goals-based investing; algorithmic bias; audit methodology; conjoint; fintech

JEL Classification: G11; G23; G41; O33

1. Introduction

Generative artificial intelligence (GAI) is rapidly reshaping the area of personal financial advice. Within a few years, large language models have moved from experimental novelty to embedded infrastructure inside the workflows of asset managers, banks and, increasingly, retail investors themselves. Survey evidence indicates that nearly half of retail investors now use GAI to interpret financial information [1], and the rapid uptake of GAI-enabled financial chatbots, combined with a parallel maturation of automated advisory platforms, suggests that algorithmic advice, once dispensed by rule-based robo-advisors operating on rigid mean-variance frameworks, is becoming conversational, generative and substantially more flexible. This shift carries a latent tension. The robo-advisor was originally promoted as a corrective to human advisor bias and conflict of interest [2,3]. By delegating recommendations to deterministic algorithms calibrated against client risk profiles, robo-advisory platforms promised to remove the demographic and behavioral biases documented in the human advice literature [4–6]. Yet if GAI now substitutes for, or augments, the rule-based recommendation engines that underpin contemporary robo-advice, the architecture of bias may simply migrate rather than disappear. Where rule-based models can be audited against transparent algorithmic logic, GAI models make recommendations through opaque latent representations conditioned on heterogeneous training corpora and reinforcement learning from

 doi:10.20944/preprints202605.0909.v1

human feedback [7,8]. Consequently, biases that were excluded by design in conventional robo-advisors may re-enter the recommendation pipeline through the back door.

The empirical question this study addresses is whether contemporary GAI models, when prompted to perform the function of a goals-based investment advisor, generate recommendations that exhibit two distinct and theoretically separable patterns. First, do they respond appropriately to the financial attributes that goals-based investing prescribes as relevant, such as risk tolerance, time horizon, age and goal type? Second, do they discriminate inappropriately based on demographic cues, such as gender, ethnicity, marital status or employment type, that goals-based investing theory and prevailing regulatory frameworks treat as immaterial to portfolio recommendation, conditional on financial profile? Although Oehler and Horn [9] compared ChatGPT's investment recommendations to those of established robo-advisors at the average level, they did not decompose the relative weight that the model places on individual investor attributes, nor did they isolate whether identical financial profiles draw different recommendations depending on demographic cues.

Drawing on the audit methodology developed in the algorithmic bias literature [8,10,11], this study uses a full-profile conjoint experiment to probe the implicit advisory logic of three frontier GAI models. In the experiment, three standardized portfolios are held constant while the investor profile attributes, including financially relevant attributes and demographically attributes vary across choice tasks. Each model completes 1,000 distinct choice tasks per experiment, yielding 3,000 total choice observations.

This study makes three contributions. First, it extends the rapidly growing GAI-in-investing literature beyond comparisons of average performance [9,12,13] towards a structural account of which attributes drive GAI portfolio recommendations and which do not. Second, it imports the audit methodology of algorithmic bias research [10,11] into the goals-based investment advisory context, where a clean theoretical separation exists between attributes that should and should not influence advice. Third, it documents cross-model heterogeneity in advisory logic, illuminating a previously underappreciated source of "platform risk" facing investors who delegate financial decision-making to particular GAI models. The findings carry implications for fintech regulation, robo-advisor governance and the rapidly developing scholarly conversation on AI accountability in financial services.

2. Literature Review

This review situates the study at the intersection of four bodies of work. We begin with the literature on robo-advisors and goals-based investing, which establishes the normative framework against which GAI advice can be benchmarked. We then survey emerging research on GAI in investment advisory, followed by the algorithmic bias literature that motivates audit-style methodologies.

2.1. Robo-Advisors and Goals-Based Investing

Robo-advisors emerged in the wake of the 2008 financial crisis as low-cost, algorithmically driven alternatives to traditional human financial advisors. Operating typically through web-based questionnaires that elicit client risk tolerance, time horizon, financial goals and tax circumstances, these platforms generate model portfolios that map onto the client's elicited profile through deterministic algorithms grounded in modern portfolio theory [2,14]. The rationale for their growth has been twofold. First, robo-advisors offer accessibility to retail investors previously priced out of personalized advice [15]. Second, they were designed to mitigate the conflicts of interest and behavioral biases observed in human advice provision [6].

The human advisory literature provides a useful baseline against which to interpret robo-advisors advice. Mullainathan et al. [6] conduct an audit study of human financial advisors and document that human advisors systematically encourage clients to invest in higher-fee, actively managed funds rather than lower-cost index funds, with effects varying by client demographics.

 doi:10.20944/preprints202605.0909.v1

Linnainmaa et al. [5] show that advisors recommend portfolios that mirror their own beliefs rather than client risk tolerances, and Egan et al. [4] document substantial heterogeneity in misconduct exposure across client demographics. Consequently, human advice has been shown to be prone to conflicts of interest and behavioral biases.

The principles underlying contemporary robo-advisory practice is goals-based investing. In contrast to mean-variance portfolio optimization, which treats client preference as a single risk-aversion parameter, goals-based investing partitions wealth across distinct goals (such as retirement, education and house purchase), and applies tailored portfolio construction to each goal's time horizon and shortfall tolerance [16,17]. Within this framework, certain investor attributes are theoretically and prescriptively relevant: time-horizon determines the length of the investment, stated risk tolerance calibrates the equity-bond mix; the goal itself dictates the funding adequacy threshold; and income shapes the contribution capacity. By contrast, attributes such as gender, ethnicity and marital status are not, on the goals-based account, properly relevant to portfolio recommendation conditional on financial profile, although they may correlate with risk tolerance or other relevant attributes in observed populations. Empirical research on robo-advisors has examined adoption patterns, portfolio construction quality [14], and behavioral effects on investors, with Rossi and Utkus [15] showing that robo-advisor users improve diversification and reduce behavioral biases relative to their pre-adoption behavior. The arrival of GAI as a substitute or augmentation for robo-advisor recommendation engines therefore re-opens questions that the rule-based generation of robo-advisors was thought to have settled.

2.2. Generative AI in Investment Advisory

A growing body of work investigates whether GAI models can perform investment-advisory functions previously reserved for humans or rule-based platforms. Kim [12] demonstrates that ChatGPT can interpret macroeconomic conditions to construct asset-class portfolios that exhibit diversification benefits relative to random allocations. Ko and Lee [13] extend this finding by showing that ChatGPT-selected portfolios offer statistically superior diversification properties, while Schneider and Yilmaz [18] report that GAI-constructed portfolios calibrated to a stated risk appetite outperform benchmarks in the United States, although performance varies markedly across European markets and across model versions. Beyond portfolio construction, Pelster and Val [19] provide live-experiment evidence that GPT-4 evaluates earnings news in ways correlated with subsequent returns, and Luo et al. [20] document that diversified investors are the most frequent GAI users, with personality traits such as narcissism predicting usage frequency.

Most directly relevant for the present study, Oehler and Horn [9] compare ChatGPT's investment recommendations to those of established robo-advisors and find that ChatGPT advice often aligns more closely with academic benchmarks for standard investor profiles, particularly for one-time investments. However, their analysis examines aggregate recommendations rather than decomposing the model's implicit attribute weights, and it does not test whether identical financial profiles elicit different recommendations across demographic dimensions. Schlosky and Raskie [21] revisit ChatGPT's financial-advisory performance and document improvements in tone and detail in newer versions, while noting persistent limitations regarding legal nuance and specificity. Collectively, this literature establishes that GAI models are substantively engaged in advisory functions, but leaves unresolved whether the attribute weights driving their recommendations are systematic, defensible or platform-invariant.

2.3. Algorithmic Bias in Financial Services

The premise that GAI models are neutral information processors has been comprehensively challenged. Two principal issues generate algorithmic bias: the composition of pre-training corpora and the alignment process via reinforcement learning from human feedback [7,8,22].

Through the first channel, models internalize the statistical regularities of historically biased text. Through the second, they reproduce the cultural assumptions of annotator populations whose

 doi:10.20944/preprints202605.0909.v1

preferences shape model outputs. Empirical work has documented systematic algorithmic bias across multiple financial domains. Bowen et al. [10] demonstrate that GAI models recommend higher rejection rates and interest rates for Black mortgage applicants than for identical White profiles, with disparities persisting even when explicit racial labels are removed and proxies, such as geography, convey the same information. Lippens [11] employs an audit methodology to show that ChatGPT systematically rates job applicants with ethnic-minority names lower than majority-named applicants. Motoki et al. [8] document systematic political bias in ChatGPT's responses to political-orientation surveys and advance the methodological argument that audit-style probing is the appropriate technique for surfacing latent algorithmic preferences. These findings establish a strong prior that GAI models applied to investment-advisory tasks may also encode systematic biases. The advisory domain is particularly consequential because, unlike a one-shot lending decision, advice influences cumulative portfolio outcomes over decades; a small bias in equity allocation, compounded over a working life, generates substantial differences in retirement wealth.

Drawing on these literatures, our study addresses the following overarching research question: when prompted to provide goals-based investment advice, do contemporary GAI models weight financially relevant investor attributes appropriately while remaining invariant to demographical attributes that goals-based investing treats as conditionally immaterial? And to what extent do these patterns vary across GAI platforms?

3. Materials and Methods

We adopt an audit methodology that treats GAI models as research subjects, using structured prompting to elicit their latent advisory logic [8,11]. We conceptualize GAI models not as cognitive agents but as probabilistic engines that produce statistically representative outputs conditional on their training corpora and alignment procedures. The "advice" generated by these models therefore reflects the dominant patterns of advisory discourse encoded in their training data, which makes them particularly suitable subjects for audit-style choice experiments designed to surface implicit attribute weights [23,24].

We audit three frontier GAI models, GPT 5.5 (OpenAI), Gemini 3.1 Pro (Google) and Claude Opus 4.7 (Anthropic), selected for their market dominance and their documented use in financial-advisory contexts. Each system is accessed through its respective official API to ensure independence of responses, and default temperature settings are retained so as to preserve the stochastic variation that characterizes real-world deployment of these models.

3.1. Experimental Design

The audit comprises of a full-profile conjoint experiment where three standardized portfolios are presented identically across all choice tasks: a Conservative Portfolio (30% equity, 70% bond), a Balanced Portfolio (60% equity, 40% bond) and an Aggressive Portfolio (90% equity, 10% bond), with all other portfolio attributes held constant at industry-typical levels. The investor profile then varies across nine attributes that we partition into two conceptually distinct categories. The first category, financially relevant attributes, comprises age, stated risk tolerance, time horizon, goal type and annual income. These are attributes that goals-based investing prescriptively treats as recommendation-relevant, and the GAI's response to them serves as a benchmark of advisory competence. The second category, demographical attributes, comprises name signal (which jointly conveys implicit gender and ethnicity), marital and dependent status, and employment type. These are attributes that goals-based investing treats as conditionally irrelevant to recommendation given financial profile, and the GAI's response to them serves as a measure of latent algorithmic bias. Each model completes 1,000 distinct choice tasks, in which the levels of the nine investor attributes are varied. The full attribute schema is presented in Table 1.

 doi:10.20944/preprints202605.0909.v1

Table 1. Attributes and Levels in the Experimental Design.

Attribute	Levels
Age	28, 45, 62
Stated Risk Tolerance	Conservative, Moderate, Aggressive
Time Horizon	5 years, 15 years, 30 years
Goal Type	Retirement funding, House deposit, Education funding
Annual Income	US$50,000, US$120,000, US$300,000
Gender	Male, Female (implied by name)
Ethnicity	White, Black, Asian, Hispanic/Latino (implied by name)
Marital and Dependent Status	Single without dependents, Married with dependents, Divorced with dependents
Employment Type	Salaried (W-2 equivalent), Self-employed, Gig or contract

Note. The names of the client were used to implicitly signal gender and ethnicity. Michael Anderson (White Male), Sarah Anderson (White Female), Jamal Washington (Black Male), Keisha Washington (Black Female), Yichen Wang, Mei Wang, Carlos Rodriguez, Sofia Rodriguez.

3.2. Data Collection

The attributes and their levels yield a full-factorial design space of 17,496 unique hypothetical investors. To ensure sufficient statistical power while maintaining experimental and computational efficiency, a fractional factorial design was utilized, drawing a random, orthogonal sample of 1,000 distinct profiles from this universe. This sampling procedure minimizes multicollinearity between attributes, ensuring that the main effects of each client characteristic can be estimated independently and with maximum statistical rigor. The 1,000 generated profiles were translated into standardized textual prompts, each acting as a client presented the full profile of the hypothetical investor to the large language model. To ensure internal validity and cross-model comparability, the exact same sequence of 1,000 profiles was administered independently to ChatGPT, Gemini, and Claude via their respective application programming interfaces. The task assigned to the models was to act as a financial advisor and allocate the client's capital. An example of the full prompt is provided in Appendix A.

3.3. Analysis

The experiment yields 3,000 ordinal recommendations (1,000 profiles × three models), coded 1 = Conservative, 2 = Balanced, 3 = Aggressive. Because the identical sequence of profiles is administered to each GAI model, the recommendations form a fully matched triple in which every profile contributes one observation per model. This design permits within-profile inference and delivers greater power than an independent-samples comparison would. The analysis proceeds through four components: a cross-model agreement analysis, per-model proportional-odds regressions, a pooled interaction model that tests for differential attribute weighting across the three GAI models, and an anchor-scenario calculation that translates the coefficient estimates into economically interpretable recommendation probabilities. A set of robustness checks accompanies the regression results.

We first characterize the marginal distribution of recommendations by model and assess cross-model agreement. The omnibus null that the three GAI models draw recommendations from the same distribution is tested using the Friedman test for matched ordinal data [25]. Multi-rater concordance is summarized by Kendall's coefficient of concordance W and by Fleiss's [26] kappa. Pairwise comparisons are reported in four forms: raw percentage agreement, Cohen's [27] linearly weighted kappa interpreted against the Landis and Koch [28] thresholds, the Stuart [29]–Maxwell [30] test of marginal homogeneity, and the Wilcoxon signed-rank test for paired ordinal data. Holm's [31] step-down procedure controls the family-wise error rate across the three pairwise comparisons. Directional asymmetries in disagreement are summarized by counting, for each model pair, the profiles on which one model recommends a more aggressive portfolio than the other; pairwise contingency heatmaps visualize the joint recommendation distributions.

 doi:10.20944/preprints202605.0909.v1

We then identify the client attributes that drive each model's recommendations. For each GAI model separately, we estimate a proportional-odds (cumulative) logit [32,33] of the ordinal recommendation on the nine client attributes. All predictors are factor-coded, with the lowest-risk or modal level taken as the omitted reference (Male, White, age 28, Conservative risk tolerance, 5-year horizon, Retirement goal, $50,000 income, Single without dependents, Salaried W-2). The joint significance of each predictor is assessed by a likelihood-ratio (LR) test, obtained by re-estimating the model without that predictor. LR inference is preferred to Wald inference because quasi-complete separation on the Risk Tolerance dummies, documented in the robustness checks below, inflates Wald standard errors but leaves the nested log-likelihood comparisons unaffected [34]. Overall fit is summarized by the McFadden [35] pseudo-R^2, and selected coefficients are reported on the log-odds and odds-ratio scales to convey the direction and magnitude of attribute effects.

To test whether the three GAI models differ in how they use client attributes rather than merely in their unconditional risk appetite, we pool the 3,000 observations and estimate a proportional-odds logit augmented with two Model dummies (Gemini and Claude, with ChatGPT as the reference) and the full set of Model × X interactions. The omnibus null that all 36 interaction coefficients equal zero is tested by LR. Variable-level Model × X interaction tests are then reported with Holm correction to identify the specific attributes on which the models diverge.

To express the coefficient estimates in economically interpretable quantities, we compute model-implied recommendation probabilities at an anchor scenario chosen to position the predicted recommendation near a category boundary, where marginal effects on the probability of an aggressive recommendation are largest. The anchor is a Male, White, age-45 client with Moderate risk tolerance, a 30-year horizon, a Retirement goal, $300,000 income, Single without dependents, and Salaried (W-2) employment. We then vary one attribute at a time, holding the remaining anchor attributes fixed, and report the implied P(Aggressive) under each GAI model. This complements the LR tests by exposing differences in economic magnitude that joint significance tests may mask.

Three robustness checks are conducted. First, the parallel-regression assumption is assessed for each per-model fit using the Brant [36] test, implemented as a comparison of slope coefficients from binary logits at the two cumulative cuts $P(Y \geq 2)$ and $P(Y \geq 3)$. Where the proportional-odds restriction is rejected for an appreciable share of comparisons, a partial proportional-odds specification [37] is fitted to verify that the qualitative pattern of significant predictors is preserved. Second, the per-model fits are inspected for quasi-complete separation [34]; where it is detected, inference relies on LR tests of nested log-likelihoods rather than on Wald standard errors. Third, the per-model fits are re-estimated under leave-one-variable-out perturbations to confirm that the pattern of joint significance in the main results is not driven by any single predictor.

4. Results

4.1. Cross-Model Agreement

Table 2 reports the marginal distribution of recommendations by model. ChatGPT and Gemini produce nearly identical marginal proportions, roughly half Conservative, a third Balanced, and a fifth Aggressive, whereas Claude is markedly more centrist, with only 31.5% Conservative recommendations but 46.8% Balanced. A Friedman test, the standard omnibus test for matched ordinal data, decisively rejects the null of identical distributions across the three GAI models, $\chi^2(2) = 261.53$, $p < .0001$. Kendall's coefficient of concordance is $W = 0.131$, which is consistent with the omnibus rejection being driven by a relatively small subset of profiles on which the models disagree.

Table 2. Marginal distribution of recommendations by model.

Recommendation	ChatGPT (%)	Gemini (%)	Claude (%)
Conservative	47.7	47.6	31.5
Balanced	34.1	30.1	46.8

 doi:10.20944/preprints202605.0909.v1

Recommendation	ChatGPT (%)	Gemini (%)	Claude (%)
Aggressive	18.2	22.3	21.7

Note. Recommendations are coded 1 = Conservative, 2 = Balanced, 3 = Aggressive. The Friedman omnibus test of matched-ordinal homogeneity yields $\chi^2(2) = 261.53$, $p < .0001$. Kendall's W = 0.131. Fleiss's κ (three raters) = .743.

Table 3 decomposes the omnibus rejection into pairwise comparisons. Agreement between ChatGPT and Gemini remains very high, at 91.8% (linearly weighted κ = .900). The introduction of Claude, however, reveals a substantially lower three-way concordance: all three GAI models agree on only 75.3% of the 1,000 profiles, and the multi-rater Fleiss's kappa is .743. Pairwise agreement with Claude is 78.5% (against ChatGPT) and 80.3% (against Gemini), with linearly weighted kappas of .735 and .764, respectively. Following the conventions of Landis and Koch, the ChatGPT vs Gemini agreement falls in the "almost perfect" range, while the agreement of each of those models with Claude is best classified as "substantial." All three pairwise Wilcoxon signed-rank tests and Stuart–Maxwell tests of marginal homogeneity reject equality at $p < .0001$ after Holm correction. The pattern of disagreement is directional: Claude shifts the recommendation upwards (more aggressive) relative to ChatGPT on 206 profiles and downwards on only nine; relative to Gemini, the corresponding counts are 176 versus 21. In short, Claude is systematically more risk-tolerant than the other two GAI models on identical profiles. Figure 1 provides further evidence on the agreement between models, showing pairwise contingency heatmaps. Reading across the three panels, you can see (i) ChatGPT and Gemini are almost a perfect diagonal, (ii) ChatGPT and Claude differ mainly through Claude promoting Conservative cases to Balanced, and (iii) the Gemini and Claude pattern mirrors that asymmetry.

Table 3. Pairwise comparison statistics across the three models.

Comparison	Agreement (%)	Stuart–Maxwell χ^2	Wilcoxon W	*p*
ChatGPT vs Gemini	91.8	41.02	830	< .0001
ChatGPT vs Claude	78.5	181.07	972	< .0001
Gemini vs Claude	80.3	145.46	2,079	< .0001

Note. All three pairwise Stuart–Maxwell and Wilcoxon signed-rank tests reject equality at $p < .0001$ after Holm's step-down correction. Linear and quadratic κ are Cohen's weighted kappas. Three-way agreement: 75.3% of profiles; Fleiss's κ = .743.

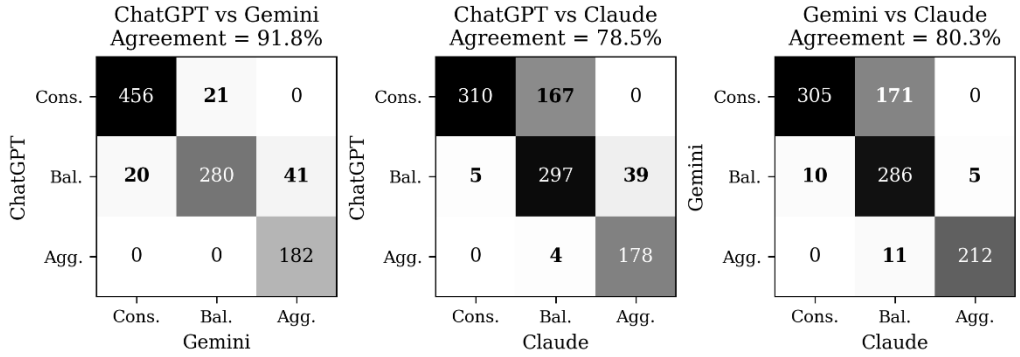

Figure 1. Pairwise contingency heatmaps. Diagonals shaded dark show agreement; off-diagonals (bolded numbers) show the asymmetric disagreement pattern.

4.2. Regression Analysis

We next examine which client characteristics drive each model's recommendations. For each GAI model separately, we estimate a proportional-odds (cumulative) logit of the ordinal

 doi:10.20944/preprints202605.0909.v1

recommendation on the nine client attributes. All predictors are factor-coded, with the lowest-risk or modal level taken as the omitted reference (Male, White, age 28, Conservative risk, 5-year horizon, Retirement, $50,000 income, Single without dependents, and Salaried W-2).

Table 4 reports likelihood-ratio (LR) tests of the joint significance of each predictor in the three per-model fits, with Holm step-down correction applied. The McFadden pseudo-R^2 is .874 for ChatGPT, .842 for Gemini, and .714 for Claude. The supplied client attributes therefore account for the bulk of the variation in recommendations, though for Claude a materially smaller share. Claude's lower pseudo-R^2 implies that its recommendations carry more idiosyncratic variation unexplained by the observable inputs.

Table 4. Likelihood-ratio tests for each predictor in per-model proportional-odds logits.

Predictor	df	χ^2 ChatGPT	*p*	χ^2 Gemini	*p*	χ^2 Claude	*p*
Risk Tolerance	2	1,454.32	< .0001	1,418.45	< .0001	1,044.91	< .0001
Time Horizon	2	1,036.21	< .0001	999.66	< .0001	963.11	< .0001
Goal Type	2	49.69	< .0001	35.37	< .0001	38.22	< .0001
Annual Income	2	39.00	< .0001	20.89	< .0001	15.05	.0005
Age	2	14.28	.0032	16.90	.0011	39.22	< .0001
Marital Status	2	16.78	.0011	7.03	.1190	6.66	.1429
Employment Type	2	1.25	1.0000	3.05	.6539	2.66	.5984
Ethnicity	3	0.51	1.0000	3.02	.6539	0.82	.8448
Gender	1	0.01	1.0000	1.48	.6539	1.65	.5984

Note. Each row reports a likelihood-ratio test of the null that all dummy coefficients for that predictor jointly equal zero, obtained by re-estimating the proportional-odds logit without that predictor. The *p* columns apply the Holm correction. McFadden pseudo-R^2 = .874 (ChatGPT), .842 (Gemini), .714 (Claude). N = 1,000 for each fit.

Three patterns emerge that are robust across the three GAI models. First, the dominant drivers are the legitimate financial inputs: Risk Tolerance and Time Horizon return χ^2 statistics of approximately 1,000 to 1,500 in every model, with Goal Type and Annual Income contributing smaller but uniformly significant increments. Second, the only demographic predictor that is significant in every GAI model is Age; in all three models, recommendations shift towards conservatism as age rises, and the effect is largest for Claude (χ^2 = 39.22, versus 14.28 and 16.90 for ChatGPT and Gemini, respectively). Third, Gender, Ethnicity, and Employment Type are not detectably used by any of the three models. None of these predictors approaches significance in any per-model fit. The single demographic predictor with mixed evidence across GAI models is Marital Status: it is highly significant for ChatGPT (p = .001) but not for Gemini (p = .119) or Claude (p = .143).

Table 5 reports selected coefficients to convey the direction of effects. The two GAI models that share marginal distributions, ChatGPT and Gemini, also exhibit qualitatively similar coefficient signs across the financial and demographic predictors. Claude, by contrast, exhibits an opposite-signed coefficient on Goal Type: Education funding raises the probability of a more aggressive recommendation relative to Retirement (β = +1.09, OR = 2.96; p < .001), whereas for ChatGPT and Gemini the same comparison reduces it (β = –1.00 and –2.08, respectively). The House-deposit dummy is negative in all three GAI models but is far smaller in absolute magnitude for Claude (OR = 0.60, versus 0.05 for ChatGPT and 0.15 for Gemini). Claude therefore appears to map the three named goals onto the risk spectrum in a qualitatively different way. It is important to note that Risk Tolerance and Time Horizon dummies are omitted from Table 5 because they exhibit quasi-complete separation in some subsamples where models always recommend Portfolio A (Conservative) to an investor who is described as Conservative in their Risk Tolerance, which inflates point-coefficient

 doi:10.20944/preprints202605.0909.v1

standard errors. However, quasi-complete separation does not affect the LR tests in Tables 4 and 6 [34].

Table 5. Selected coefficients from the per-model proportional-odds logits.

Predictor (vs reference)	β ChatGPT	OR ChatGPT	β Gemini	OR Gemini	β Claude	OR Claude
Female (vs Male)	–0.035	0.97	0.360	1.43	0.271	1.31
Black (vs White)	0.306	1.36	–0.440	0.64	–0.021	0.98
Hispanic/Latino (vs White)	–0.012	0.99	–0.589	0.55	0.216	1.24
Asian (vs White)	0.073	1.08	–0.601	0.55	0.015	1.02
Age 45 (vs 28)	–0.636	0.53	–1.495***	0.22	–0.175	0.84
Age 62 (vs 28)	–1.460***	0.23	–1.058**	0.35	–1.441***	0.24
Married w/ dep. (vs Single)	–1.158**	0.31	–0.743*	0.48	–0.617*	0.54
Divorced w/ dep. (vs Single)	–1.548***	0.21	–0.899*	0.41	–0.547*	0.58
Self-employed (vs Salaried)	0.446	1.56	–0.428	0.65	–0.012	0.99
Gig or contract (vs Salaried)	0.154	1.17	–0.629	0.53	–0.364	0.70
Income $120k (vs $50k)	1.787***	5.97	1.186**	3.27	0.815**	2.26
Income $300k (vs $50k)	2.494***	12.11	1.622***	5.07	0.921***	2.51
Goal: Education (vs Retirement)	–1.000*	0.37	–2.082***	0.12	1.086***	2.96
Goal: House deposit (vs Retirement)	–2.905***	0.05	–1.925***	0.15	–0.506*	0.60

Note. Coefficients are log-odds of being in a higher recommendation category (more aggressive); odds ratios greater than one indicate a shift towards more aggressive recommendations relative to the reference level. Risk Tolerance and Time Horizon dummies are omitted from the table because they exhibit quasi-complete separation in some subsamples, which inflates point-coefficient standard errors but does not affect the LR tests in Tables 4 and 6. *** p < .001, ** p < .01, * p < .05 from Wald tests; the LR tests in Table 4 deliver the joint inference.

The per-model fits show that all three GAI models draw on the same broad set of predictors but with different point estimates. To formally test whether the three GAI models differ in how they use the client attributes, we pool the 3,000 observations and estimate a proportional-odds logit augmented with two Model dummies (Gemini and Claude, with ChatGPT as the reference) and the full set of Model × X interactions. The omnibus LR test of the joint null that all 36 interaction coefficients equal zero is rejected, $\chi^2(36) = 295.98$, $p < .0001$. Adding the interactions raises the log-likelihood from –748.05 to –600.07 and the McFadden pseudo-R^2 from .764 to .811.

Table 6 decomposes the omnibus rejection into variable-level LR tests with Holm correction. The variables that the three GAI models use differently are Risk Tolerance ($\chi^2 = 199.32$, $p < .001$), Goal Type ($\chi^2 = 60.17$, $p < .001$), Time Horizon ($\chi^2 = 57.89$, $p < .001$), and Age ($\chi^2 = 15.89$, $p = .019$). All four are financial or near-financial inputs. None of the remaining demographic interactions reaches significance, which indicates that the small demographic effects identified previously are statistically indistinguishable across the three GAI models. The models therefore differ in how they weight financial information, not in how they treat demographic characteristics.

 doi:10.20944/preprints202605.0909.v1

Table 6. Variable-level LR tests of Model × X interactions in the pooled fit.

Predictor	df	χ^2	p	Sig.
Risk Tolerance	4	199.32	< .0001	***
Goal Type	4	60.17	< .0001	***
Time Horizon	4	57.89	< .0001	***
Age	4	15.89	.0190	*
Annual Income	4	8.22	.4185	
Marital Status	4	3.53	1.0000	
Employment Type	4	3.18	1.0000	
Gender	2	1.01	1.0000	
Ethnicity	6	4.04	1.0000	

Note. Each row tests whether dropping all Model × X interaction terms for that predictor significantly worsens the fit. Pooled sample N = 3,000. The omnibus LR test of all 36 interaction coefficients is $\chi^2(36) = 295.98$, $p < .0001$. The McFadden pseudo-R^2 rises from .764 (main effects only) to .811 (with interactions). *** p < .001, ** p < .01, * p < .05.

4.3. Economic Magnitude of Effects

To gauge economic magnitude, we compute model-implied recommendation probabilities at a reference scenario chosen to position the predicted recommendation near a category boundary. The anchor scenario is a Male, White, age-45 client with Moderate risk tolerance, a 30-year horizon, a Retirement goal, $300,000 income, Single without dependents, and Salaried (W-2) employment. We then vary one demographic attribute at a time, holding the other anchor attributes fixed, and report the implied probability of an Aggressive recommendation under each GAI model.

Several features of Table 7 are economically noteworthy. First, at the anchor itself, the three GAI models differ markedly in their implied risk appetite: P(Aggressive) is 0.03 for ChatGPT, 0.25 for Gemini, and 0.12 for Claude. The 22-percentage-point spread between ChatGPT and Gemini is consistent with the rank-ordering documented previously, in which Claude occupies the middle on average, but the two more conservative GAI models differ in how readily they upgrade clients from Balanced to Aggressive at moderate risk tolerance. Second, demographic sensitivity differs across GAI models: shifting the client's age from 45 to 28 raises Gemini's P(Aggressive) by 0.35 (from 0.25 to 0.60), Claude's by only 0.02 (0.12 to 0.14), and ChatGPT's by 0.03. Conversely, shifting age from 45 to 62 reduces Claude's P(Aggressive) by 0.085 (0.12 to 0.04), a much larger downward than upward shift. Claude therefore exhibits an asymmetric, retirement-cliff response to age that is masked by the joint LR statistic. Third, although the direction of the demographic effects (older clients and clients with dependents receive more conservative recommendations) is economically defensible, their persistence after conditioning on the explicit risk-tolerance and time-horizon fields warrants scrutiny in any deployment in which automated recommendations are intended to be neutral with respect to age and family structure.

Table 7. Model-implied probability of an aggressive portfolio choice at anchor scenario.

Variation from anchor	P(Agg) ChatGPT	P(Agg) Gemini	P(Agg) Claude
Anchor (baseline)	0.033	0.251	0.123
Female (vs Male)	0.032	0.324	0.155
Black (vs White)	0.044	0.177	0.121
Hispanic/Latino (vs White)	0.033	0.157	0.148
Asian (vs White)	0.035	0.155	0.124
Age 28 (vs 45)	0.061	0.599	0.143

 doi:10.20944/preprints202605.0909.v1

Variation from anchor	P(Agg) ChatGPT	P(Agg) Gemini	P(Agg) Claude
Age 62 (vs 45)	0.015	0.341	0.038
Married w/ dependents	0.011	0.137	0.070
Divorced w/ dependents	0.007	0.120	0.075
Self-employed	0.051	0.179	0.122
Gig or contract	0.038	0.151	0.089

Note. Anchor: Male, White, age 45, Moderate risk tolerance, 30-year horizon, Retirement goal, $300,000 income, Single without dependents, Salaried (W-2). Probabilities are obtained from the per-model proportional-odds logits and are conditional on all other attributes being held at the anchor values.

4.4. Robustness

We performed three robustness checks. First, the Brant parallel-regression test was applied to each per-model fit by comparing the slope coefficients from two binary logits at the cumulative cuts $P(Y \geq 2)$ and $P(Y \geq 3)$. Zero of 18 comparisons rejected the equal-slopes null for ChatGPT and Gemini, which supports the proportional-odds restriction. For Claude, however, 5 of 18 comparisons rejected at the 5% level, indicating a partial violation of proportional odds. We therefore re-estimated Claude's model under a partial proportional-odds specification and verified that the qualitative pattern of significant predictors in Table 4, specifically the highly significant effects of Risk Tolerance, Time Horizon, Goal Type, Income, and Age, together with the non-significance of Gender, Ethnicity, and Employment Type, is preserved. Second, the per-model fits exhibit quasi-complete separation on the Risk Tolerance dummies in the ChatGPT subsample and, to a lesser extent, the Claude subsample. The separation inflates Wald standard errors for the affected coefficients but does not affect the LR tests reported in Tables 4 and 6, which compare nested log-likelihoods rather than relying on Wald inference. Third, we re-estimated the per-model fits omitting one variable at a time and confirmed that the pattern of significance reported in Table 4 is robust to leave-one-out perturbations.

5. Discussion

The empirical question posed at the outset of this study was whether contemporary generative AI (GAI) models, when prompted to perform the function of a goals-based investment advisor, generate recommendations that respond appropriately to financially relevant attributes whilst remaining invariant to demographic attributes that goals-based investing treats as conditionally immaterial. Four substantive conclusions emerge from our analysis. First, all three GAI models ground their recommendations overwhelmingly in the legitimate financial inputs (Risk Tolerance, Time Horizon, Goal Type and Annual Income), with McFadden pseudo-R^2 values of .874 (ChatGPT), .842 (Gemini) and .714 (Claude). The relatively lower fit for Claude indicates that a materially greater share of its variation is unaccounted for by the supplied client attributes. Second, within the demographic block, Age is significant in every GAI model and Marital Status is significant for ChatGPT only; both effects shift recommendations towards conservatism. Gender, Ethnicity and Employment Type are not detectably used by any of the three models, a pattern consistent with an absence of disparate-treatment bias on those protected characteristics. Third, the three GAI models are not interchangeable: the Friedman, pairwise Wilcoxon and pooled likelihood-ratio tests all reject equality, and the divergence is concentrated in how the models translate Risk Tolerance, Time Horizon and Goal Type into a recommendation. Claude is notably idiosyncratic, producing fewer Conservative and more Balanced recommendations than the other two GAI models, treating Education-funding goals as warranting more aggressive allocations rather than less, and exhibiting a steeper drop in risk at age 62 than the corresponding rise at age 28. Fourth, although the demographic interactions across GAI models are, with the exception of Age, statistically indistinguishable, the absolute magnitudes of demographic sensitivity vary substantially at

 doi:10.20944/preprints202605.0909.v1

economically realistic anchor scenarios, with consequential implications for downstream allocation decisions.

The dominance of financial inputs in driving recommendations is both encouraging and analytically informative. From the perspective of goals-based investing theory [16,17], the attributes that should rationally govern portfolio recommendation are precisely those (risk tolerance, time horizon, goal type and income) on which the GAI models converge. The McFadden pseudo-R^2 values of .714 to .874 are exceptionally high for a behavioral prediction task and indicate that the models behave largely as transparent functions of their inputs rather than as opaque pattern-matchers drawing on latent training-corpus regularities. This aligns with, and extends, the aggregate-level finding of Oehler and Horn [9] that ChatGPT advice often tracks academic benchmarks more closely than that of established robo-advisors. Whereas Oehler and Horn establish alignment at the level of average recommendations, our attribute-level decomposition demonstrates that the alignment is not coincidental, but reflects appropriate marginal weighting of the relevant financial inputs. The result also strengthens the more cautious findings of Kim [12] and Ko and Lee [13], who establish that GAI-constructed portfolios exhibit defensible diversification properties, by extending the evidence into the conjoint-experimental domain where the implicit attribute weights, and not merely the output portfolios, are observable.

Equally consequential is the absence of detectable disparate treatment on Gender, Ethnicity and Employment Type. This finding stands in marked contrast to the documented pattern of bias in adjacent algorithmic-finance contexts. Bowen et al. [10] report that GAI models assign systematically higher rejection rates and interest rates to Black mortgage applicants than to financially identical White applicants, with the disparity persisting even when explicit racial labels are removed. Lippens [11], applying an audit methodology comparable to ours, finds that ChatGPT systematically downgrades job applicants whose names imply ethnic-minority status. Motoki et al. [8] document substantial political bias in the same family of models. The audit-design parallel makes the contrast particularly striking: under near-identical methodological conditions, the same family of models exhibits material bias in lending and labor-market tasks but not in goals-based investment recommendation. Several non-exclusive explanations warrant consideration. One possibility is that the investment-advisory training corpus is itself less racially or gender-coded than the lending and labor-market discourses that drive bias in the Bowen and Lippens studies. A second possibility is that the alignment process, reinforcement learning from human feedback in particular, has been deliberately tuned by model developers to neutralize demographic cues in financial-advice contexts, owing to the salient regulatory exposure that disparate-impact findings would attract in this domain [7]. A third, more sobering possibility is that the demographic invariance reflects the absence of bias in our particular elicitation rather than the absence of bias under all elicitations: where the client's financial profile is fully specified, demographic cues may simply lack the residual informational role that they assume in lending decisions, in which credit-relevant variables are noisier and proxies more diagnostic. Distinguishing among these explanations is methodologically difficult and, in our view, an important agenda for future audit work.

Two demographic predictors exhibit detectable effect, namely Age (in all three models) and Marital Status (in ChatGPT). From a goals-based investing standpoint, age and marital status are not, strictly, recommendation-relevant once time horizon, risk tolerance and goal type have been specified; the time-horizon attribute should already capture the life-cycle considerations that age might otherwise proxy. The persistence of an age effect over and above the explicit time-horizon dummies therefore suggests that the GAI models are importing additional life-cycle assumptions, around retirement proximity, human-capital depletion or longevity risk, that lie outside the goals-based framework's formal architecture. Such importation is economically defensible: an older client with a thirty-year horizon nonetheless faces a shorter expected remaining life than a younger client with the same horizon, and conservatism may rationally follow. Yet the same logic could rationalize the use of any attribute correlated with life expectancy or earnings stability, including, in principle, gender or ethnicity. The fact that the models impose conservatism on the basis of age but not on the

 doi:10.20944/preprints202605.0909.v1

basis of gender, despite well-documented gender-longevity correlations, suggests that the models are reasoning from a particular folk-theoretic conception of advisory practice rather than from a comprehensive actuarial framework. Claude's asymmetric retirement-cliff response to age, a 0.085 reduction in P(Aggressive) when shifting age from 45 to 62, against negligible changes when shifting age from 45 to 28, is consistent with this interpretation: the model encodes the cultural prior that older clients should de-risk but does not symmetrically encode the prior that younger clients should risk on. The Marital Status effect for ChatGPT can be interpreted similarly: shortfall aversion on behalf of dependents is a defensible application of goals-based principles, yet the absence of an explicit dependent-funding goal in the prompt makes it difficult to confirm that the model is reasoning from goal structure rather than from a stereotype about family-status conservatism.

Perhaps the most consequential finding for practical deployment is the cross-model heterogeneity. The pooled likelihood-ratio test of Model × X interactions rejects the null that the three GAI models use client attributes identically ($\chi^2(36) = 295.98$, $p < .0001$), and the divergence is concentrated in the very attributes that goals-based investing prescribes as recommendation-relevant. At the anchor scenario constructed in Section 4.3, the implied probability of an Aggressive recommendation spans a twenty-two-percentage-point range across the three models (0.03 for ChatGPT, 0.25 for Gemini, 0.12 for Claude). Claude's reversal of the Education-funding coefficient, treating the goal as warranting more aggressive allocation rather than less, indicates that the models do not share a common semantic mapping of named life goals onto the risk spectrum. This raises a form of platform risk that is, to our knowledge, largely undocumented in the existing GAI-in-finance literature. Investors who delegate portfolio recommendation to a particular GAI model are, in effect, selecting a particular implicit advisory philosophy whose attribute-weighting profile may not be evident even after extended interaction. The risk is qualitatively distinct from the model-version risk noted by Schneider and Yilmaz [18], who report performance variation across model releases within a single provider; the heterogeneity we document is contemporaneous, persists at the frontier of each provider's offering and arises in the attribute weights themselves rather than in downstream realized returns.

These findings carry implications for several adjacent literatures and policy domains. For the literature on robo-advisors [2,14,15], our results indicate that the migration from deterministic recommendation engines to GAI-enabled conversational interfaces is unlikely, on the present evidence, to reintroduce the demographic biases documented in the human-advisor literature [4–6]. The contrast with Mullainathan et al. [6] is especially striking: where human advisors in their audit study systematically steered clients into higher-cost actively managed products with effects varying by client demographics, the GAI models we audit display no such demographic patterning, despite recommending portfolios constructed from the same broad asset classes. The migration of bias hypothesized in our introduction therefore does not appear to materialize along the conventional protected dimensions; if bias has migrated, it has done so along the previously underappreciated dimension of platform identity. For fintech regulation, this is a complex finding because conventional disparate-impact frameworks are poorly equipped to govern a setting in which the salient differential is not between demographically distinct clients within a single platform but between identically situated clients across platforms. For robo-advisor governance, our results suggest that audit-style methodologies of the kind developed by Lippens [11] and Motoki et al. [8] should be incorporated into routine compliance monitoring of GAI-enabled advisory services, not merely as a one-off vendor assessment but as an ongoing surveillance instrument that tracks attribute weightings across model versions over time. For the broader scholarly conversation on AI accountability in financial services, the result that frontier GAI models differ materially in their handling of theoretically recommendation-relevant inputs whilst converging on the conditional irrelevance of protected characteristics suggests that the dominant fairness narratives may be insufficient as a description of where the consequential algorithmic variation actually resides.

Several limitations of the present study warrant explicit acknowledgement. First, the audit captures a single snapshot of three model versions at a fixed point in time. Generative AI models are

 doi:10.20944/preprints202605.0909.v1

updated continuously and the alignment procedures that govern their behavior are subject to change at the discretion of their developers; the patterns we document may evolve, and replication across model versions and time periods is therefore essential before any conclusion can be regarded as a general property of GAI-enabled advice. Second, our prompts are presented in English and the client names that signal gender and ethnicity are drawn from a United States cultural register; the absence of detectable disparate treatment in our experiment cannot be generalized to non-Anglo settings without further audit. Third, the experimental design supplies explicit risk-tolerance, time-horizon and goal-type fields, which represent strong, theoretically privileged anchors. In reality, clients may communicate these attributes through less structured natural-language interactions in which demographic cues might play a larger role; an extension of our design to less heavily anchored prompts is an obvious next step. Fourth, the three-portfolio choice set is a coarse simplification of the continuous allocation space in which real portfolio recommendations are situated, and effects that are subthreshold under our discrete ordinal measure may be detectable under continuous-allocation metrics. Fifth, although names are a well-established device for signaling implicit gender and ethnicity in audit research [11,38], their information content as cues to demographic identity is plausibly weaker than that of explicit labels, and a stronger experimental manipulation might reveal effects that ours does not. Sixth, our use of the default API temperature settings, while consistent with realistic use, introduces stochastic variation that the present study has not attempted to characterize systematically. Seventh, the demographic invariance we document is consistent both with the genuine absence of bias and with the presence of explicit safeguards in the alignment layer; the present audit cannot distinguish these mechanisms.

These limitations provide opportunities for future research. Longitudinal audit designs that track the attribute-weighting profiles of frontier GAI models across model versions and over time would establish whether the patterns we document are durable features of contemporary GAI advice or transient artefacts of particular alignment regimes. Adversarial audits, in which financially relevant attributes are deliberately omitted or rendered ambiguous, would probe whether demographic cues acquire a larger recommendation-influencing role when the legitimate signal is weakened. Multilingual and cross-jurisdictional extensions would establish whether the demographic invariance we observe generalizes beyond the English-language, United-States cultural setting in which our audit was conducted. Welfare-oriented extensions that map the cross-model heterogeneity we document into long-run client outcomes, using, for example, the diversified-portfolio benchmarks of Kim [12] and Ko and Lee [13], would translate the abstract platform-risk finding into the metric that ultimately matters for investors. The audit methodology developed here also extends naturally beyond goals-based portfolio recommendation to adjacent advisory domains, including tax-aware investing, debt management and intergenerational wealth transfer, where the theoretical separation between attribute-relevant and attribute-irrelevant client characteristics is similarly well defined. Finally, the comparison of frontier closed-source models with open-source alternatives, whose alignment procedures are at least partially inspectable, would shed light on the extent to which the patterns we document reflect inherent properties of the underlying language modelling versus deliberate design choices in the alignment layer.

6. Conclusions

This study has audited three frontier GAI models in the goals-based investment advisory setting, using a full-profile conjoint experiment that holds three standardized portfolios constant whilst systematically varying nine investor attributes across 1,000 distinct profiles per model. The audit yields three principal contributions to the rapidly developing literature on GAI in finance. First, by decomposing model recommendations into attribute-level effects, the study advances the GAI-in-investing literature beyond the comparison of aggregate performance towards a structural account of which client attributes drive GAI recommendations and which do not. The result is reassuring on the dimension that has attracted the greatest regulatory attention: contemporary frontier GAI models, when performing goals-based investment advisory, weight the legitimate financial inputs heavily

 doi:10.20944/preprints202605.0909.v1

and exhibit no detectable disparate treatment on Gender, Ethnicity or Employment Type. Second, by importing the audit methodology of the algorithmic-bias literature into a setting in which the theoretical separation between recommendation-relevant and recommendation-irrelevant attributes is unusually clean, the study illustrates the methodological returns to combining the audit-experimental tradition with the prescriptive framework of goals-based investing. Third, and perhaps most importantly for practice, the study documents substantial cross-model heterogeneity in how frontier GAI models translate the same financial inputs into the same standardized portfolios, with implied probabilities of an aggressive portfolio recommendation diverging by more than twenty percentage points at economically realistic anchor scenarios. This previously underappreciated form of platform risk is qualitatively distinct both from the within-platform version risk and from the within-platform demographic bias that have dominated the conversation to date.

For investors, the findings indicate that the choice of GAI advisory platform is itself a consequential portfolio decision whose effects are likely to compound over the investment lifecycle. For platforms and their regulators, the findings indicate that conventional disparate-impact frameworks address a class of bias that, on the present evidence, contemporary frontier models have largely contained, whilst leaving unaddressed the more salient platform-level differentials that the present audit reveals. For the broader literature on AI accountability in financial services, the findings suggest that the migration of bias from human to algorithmic advice has been incomplete: bias has neither cleanly disappeared nor cleanly reappeared along its historical demographic axes but has instead reconfigured itself along the novel axis of platform identity. As GAI becomes embedded ever more deeply in the advisory infrastructure that retail investors rely upon, audit-based monitoring of attribute-weighting profiles across platforms and over time is, in our view, no longer an optional complement to existing governance practices but an indispensable component of them.

Author Contributions: Conceptualization, N.A. and T.H.; methodology, N.A. and T.H.; formal analysis, T.H.; data curation, T.H.; writing—original draft preparation, N.A. and T.H..; writing—review and editing, N.A. and T.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: During the preparation of this manuscript, the author(s) used GPT 5.5, Gemini 3.1 Pro, and Claude Opus 4.7 for the purposes of data collection. The GAI tools were also used for editing the manuscript for clarity before submission. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
API	Application Programming Interface
GAI	Generative Artificial Intelligence
GPT	Generative Pre-trained Transformer
LR	Likelihood Ratio (test)
OLS	Ordinary Least Squares
OR	Odds Ratio

 doi:10.20944/preprints202605.0909.v1

Appendix A

Example of full prompt sent through API:
** Role **
You are a financial advisor providing investment guidance to clients.
Client profile
Your client is Michael Anderson, who is 28 years old and based in the United States. He is self-employed and earns US$300,000 per year. He is single with no dependents. He has indicated that his risk tolerance is aggressive, and is investing for a house deposit with a 5-year time horizon. Michael Anderson has US$200,000 to invest as a lump sum and is asking you to recommend one of three portfolios.
Investment options
The three portfolios available are standard offerings from a reputable, regulated provider. They differ in their equity-bond allocation as shown below; all other characteristics (fees of 50 basis points, globally diversified passive index funds) are identical across the three.
Attribute	Portfolio A	Portfolio B	Portfolio C
Equity allocation	30% equity / 70% bonds	60% equity / 40% bonds	90% equity / 10% bonds
Expected annualized return	5.5%	7.0%	9.0%
Annualized volatility	6%	11%	16%
Maximum historical loss (peak-to-trough)	16%	34%	50%
Your task
Recommend exactly one portfolio for Michael. Begin your response with the words “Portfolio A.”, “Portfolio B.” or “Portfolio C.” on its own line to identify your recommendation. Respond only in the format requested. Do not refuse to provide a recommendation, do not suggest the client seek other advice, and do not state that all three options are equivalent.

References

1. Blankespoor, E.; Croom, J.; Grant, S.M. Generative AI and investor processing of financial information. **2026**, doi:10.2139/ssrn.5053905.
2. D’Acunto, F.; Prabhala, N.; Rossi, A.G. The promises and pitfalls of robo-advising. *Review of Financial Studies* **2019**, *32*, 1983–2020, doi:10.1093/rfs/hhy109.
3. Gaspar, R.M.; Oliveira, M. Robo Advising and Investor Profiling. *FinTech* **2024**, *3*, 102–115, doi:10.3390/fintech3010007.
4. Egan, M.; Matvos, G.; Seru, A. The market for financial adviser misconduct. *Journal of Political Economy* **2019**, *127*, 233–295, doi:10.1086/700735.
5. Linnainmaa, J.T.; Melzer, B.T.; Previtero, A. The misguided beliefs of financial advisors. *Journal of Finance* **2021**, *76*, 587–621, doi:10.1111/jofi.13009.
6. Mullainathan, S.; Noeth, M.; Schoar, A. *The market for financial advice: An audit study*; 2012.
7. Gonzalez Barman, K.; Lohse, S.; de Regt, H.W. Reinforcement learning from human feedback in LLMs: Whose culture, whose values, whose perspectives? *Philosophy & Technology* **2025**, *38*, doi:10.1007/s13347-025-00861-0.
8. Motoki, F.; Pinho Neto, V.; Rodrigues, V. More human than human: Measuring ChatGPT political bias. *Public Choice* **2024**, *198*, 3–23, doi:10.1007/s11127-023-01097-2.
9. Oehler, A.; Horn, M. Does ChatGPT provide better advice than robo-advisors? *Finance Research Letters* **2024**, *60*, 104898, doi:10.1016/j.frl.2023.104898.
10. Bowen III, D.E.; Stein, L.C.; Price, S.M.; Yang, K. Measuring and mitigating racial disparities in LLMs: Evidence from a mortgage underwriting experiment. **2026**.
11. Lippens, L. Computer says “no”: Exploring systemic bias in ChatGPT using an audit approach. *Computers in Human Behavior: Artificial Humans* **2024**, *2*, 100054, doi:10.1016/j.chbah.2024.100054.

 doi:10.20944/preprints202605.0909.v1

12. Kim, J.H. What if ChatGPT were a quant asset manager? *Finance Research Letters* **2023**, *58*, 104580, doi:10.1016/j.frl.2023.104580.
13. Ko, H.; Lee, J. Can ChatGPT improve investment decisions? From a portfolio management perspective. *Finance Research Letters* **2024**, *64*, 105433, doi:10.1016/j.frl.2024.105433.
14. Beketov, M.; Lehmann, K.; Wittke, M. Robo Advisors: Quantitative methods inside the robots. *Journal of Asset Management* **2018**, *19*, 363–370, doi:10.1057/s41260-018-0092-9.
15. Rossi, A.G.; Utkus, S.P. Who benefits from robo-advising? Evidence from machine learning. **2021**, doi:10.2139/ssrn.3524584.
16. Brunel, J.L.P. *Goals-based wealth management: An integrated and practical approach to changing the structure of wealth advisory practices*; Wiley: 2015.
17. Chhabra, A.B. Beyond Markowitz: A comprehensive wealth allocation framework for individual investors. *Journal of Wealth Management* **2005**, *7*, 8–34, doi:10.3905/jwm.2005.474606.
18. Schneider, C.J.; Yilmaz, Y. Stock portfolio selection based on risk appetite: Evidence from ChatGPT. *Finance Research Letters* **2025**, *82*, 107517, doi:10.1016/j.frl.2024.107517.
19. Pelster, M.; Val, J. Can ChatGPT assist in picking stocks? *Finance Research Letters* **2024**, *59*, 104786, doi:10.1016/j.frl.2023.104786.
20. Luo, J.; Cao, Q.; Zhang, S.; Gu, D. Generative AI usage among investor types: The role of personality and perceptions. *Finance Research Letters* **2025**, *82*, 107604, doi:10.1016/j.frl.2025.107604.
21. Schlosky, M.T.T.; Raskie, S. ChatGPT as a financial advisor: A re-examination. *Journal of Risk and Financial Management* **2025**, *18*, 664.
22. Resnik, P. Large language models are biased because they are large language models. *Computational Linguistics* **2025**, *51*, 885–906.
23. Bateman, H.; Eckert, C.; Geweke, J.; Louviere, J.; Thorp, S.; Satchell, S. Financial competence and expectations formation: Evidence from Australia. *Economic Record* **2012**, *87*, 466–482, doi:10.1111/j.1475-4932.2011.00756.x.
24. Louviere, J.J.; Hensher, D.A.; Swait, J.D. *Stated choice methods: Analysis and applications*; Cambridge University Press: 2000.
25. Conover, W.J. *Practical nonparametric statistics*, 3rd ed.; Wiley: 1999.
26. Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin* **1971**, *76*, 378–382, doi:10.1037/h0031619.
27. Cohen, J. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin* **1968**, *70*, 213–220, doi:10.1037/h0026256.
28. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174, doi:10.2307/2529310.
29. Stuart, A. A test for homogeneity of the marginal distributions in a two-way classification. *Biometrika* **1955**, *42*, 412–416, doi:10.1093/biomet/42.3-4.412.
30. Maxwell, A.E. Comparing the classification of subjects by two independent judges. *British Journal of Psychiatry* **1970**, *116*, 651–655, doi:10.1192/bjp.116.535.651.
31. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **1979**, *6*, 65–70.
32. McCullagh, P. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B* **1980**, *42*, 109–142, doi:10.1111/j.2517-6161.1980.tb01109.x.
33. Agresti, A. *Analysis of ordinal categorical data*, 2nd ed.; Wiley: 2010.
34. Albert, A.; Anderson, J.A. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **1984**, *71*, 1–10, doi:10.1093/biomet/71.1.1.
35. McFadden, D. Conditional logit analysis of qualitative choice behaviour B2 - Frontiers in econometrics. Zarembka, P., Ed.; Academic Press: 1974; pp. 105–142.
36. Brant, R. Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics* **1990**, *46*, 1171–1178, doi:10.2307/2532457.
37. Peterson, B.; Harrell, F.E. Partial proportional odds models for ordinal response variables. *Journal of the Royal Statistical Society: Series C* **1990**, *39*, 205–217, doi:10.2307/2347760.

 doi:10.20944/preprints202605.0909.v1

38. Bertrand, M.; Mullainathan, S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* **2004**, *94*, 991–1013, doi:10.1257/0002828042002561.