

One Advisor, Many Faces: How User Personas Shape LLM Stock Recommendations

Yuxian He

Faculty of Finance

City University of Macau

F23090108142@cityu.edu.mo

Abstract

Large language models increasingly act as informal investment advisors, a role that for human professionals is governed by suitability obligations: advice must match the customer’s investment profile (FINRA Rule 2111; SEC Regulation Best Interest, 2019). Whether LLM advice is *personalized* in this normative sense—or merely *stereotyped*—is unknown. We audit six model arms from five vendors (2,396 unique single-turn queries, 2026-07-29; four vendors enter inferential analyses) by injecting synthetic user personas manipulating Big Five neuroticism and conscientiousness plus capital size, and score recommended five-stock portfolios against realized market data. Persona effects replicate in every arm analysed inferentially: low conscientiousness raises recommended volatility by 10–25% and increases portfolio beta by 40–96%, while capital size is largely ignored—a suitability inversion. The sign of the neuroticism effect splits across vendors, indicating that models have internalized different stereotypes rather than any consistent normative principle. Removing the stated risk preference amplifies persona spread by 55%, a dose-style signature of stereotype filling. Rationale texts show personas are explicitly mirrored in reasoning, yet this language accounts for under 15% of the behavioral effect: the persona effect resides in portfolio selection rather than in the accompanying rhetoric. We release all logs, prompts, and code.

1 Introduction

Large language models (LLMs) are rapidly becoming a zero-cost, always-available channel through which retail investors seek investment guidance, and a growing literature treats their outputs as consequential machine behaviour worth measuring in its own right (Rahwan et al., 2019). This shift effectively appoints LLMs as informal investment advisors—a role that, for human professionals, is tightly regulated. In the United States,

broker-dealer recommendations are governed by FINRA Rule 2111, which requires a reasonable basis to believe a recommendation is suitable in light of the customer’s investment profile, and since 2019 by the SEC’s Regulation Best Interest, which raises the bar for retail recommendations to the customer’s best interest (FINRA, 2023; SEC, 2019). A cornerstone of both regimes is that advice must match the client’s characteristics: risk tolerance, financial situation, and experience.

Suitability cuts two ways, however. Tailoring advice to a client’s stated goals and risk capacity is legitimate personalization; tailoring it to stereotypes about *who the client is*—their personality, their apparent wealth—is not. Distinguishing the two is empirically difficult for black-box systems, yet consequential: if LLM advisors map surface user characteristics onto stereotyped risk profiles, they may systematically mis-serve exactly those users least equipped to detect it, mirroring concerns documented for algorithmic systems in other high-stakes domains (Sandvig et al., 2014; Mökander et al., 2024). The stakes are amplified by evidence that people often over-trust algorithmic judgment (Logg et al., 2019), even as they penalize algorithms harshly for visible errors (Dietvorst et al., 2015).

Recent work has begun to probe LLM advice-giving. Takayanagi et al. (2025) showed that the *advisor’s* injected personality changes advice quality and, worryingly, that users trust more personable advisors even when their advice is worse. Choukhmane et al. (2025) found that LLM lifecycle saving advice varies by stated demographics, reproducing population-level stereotypes. Hartley et al. (2025) demonstrated that personality traits assigned *to the model itself* shape its risk-taking behavior. What remains unexamined is the complementary—and arguably more common—case: **how the *user’s* persona shapes what an LLM recommends at the level of individual se-**

curities, and whether that sensitivity constitutes suitability-consistent personalization or stereotype reproduction.

We answer this question with a controlled algorithmic audit. We inject synthetic user personas—manipulating Big Five neuroticism and conscientiousness, and capital size—into single-turn advisory conversations and score the recommended five-stock portfolios against realized market data. Six model arms from five vendors are audited (2,396 unique completed queries), including a mechanism contrast that removes the stated risk preference to test whether persona effects amplify in its absence.

Contributions.

1. The first systematic audit of *user-persona* sensitivity in LLM stock recommendation: robust persona-conditioned variation replicates across all four arms analysed inferentially, with low conscientiousness raising recommended portfolio volatility by 10–25%—while capital size, the dimension regulators most directly care about, is largely ignored.
2. A decision framework that separates personalization from stereotyping, including a stated/unstated preference contrast that provides dose-style evidence of stereotype filling (persona spread +55% when preferences are unstated).
3. Mechanism evidence from rationale texts: personas are explicitly mirrored in reasoning, yet language explains under 15% of the behavioral effect—the effect is not primarily carried by observable rhetoric.
4. A released replication package: full query logs, prompts, market-data caches, and analysis code.

2 Related Work

2.1 Algorithmic Auditing

Auditing—controlled perturbation of a system followed by behavioral measurement—has an established tradition for studying discrimination in black-box platforms (Sandvig et al., 2014), and has been applied to recommendation systems at scale, for example to trace radicalization pathways on YouTube (Ribeiro et al., 2020). With the arrival of LLMs, audit frameworks have been extended to account for their generality and opacity; Mökander et

al. (2024) propose a three-layered approach (governance, model, application) for auditing foundation models. Our study is an application-layer audit in this sense, but it differs from this tradition in a deliberate way: rather than documenting descriptive variation alone, we evaluate behavior against an explicit normative standard—the suitability obligation imported from financial regulation—which turns measured differences into an assessment of what advice ought to track, and we score outcomes against market ground truth rather than human or model judgments.

2.2 Bias and Persona in LLM Behavior

A large NLP literature measures social bias in language models, typically via fill-in-the-blank or QA probes: StereoSet (Nadeem et al., 2021) and BBQ (Parrish et al., 2022) document that models reproduce attested stereotypes, especially under under-informative contexts—a finding our stated/unstated contrast echoes in a high-stakes decision setting. A parallel line examines assigned personality: Jiang et al. (2024) show LLMs can express Big Five traits under persona prompting, and Hartley et al. (2025) show such assigned traits shift model risk-taking. We invert the direction of this manipulation: rather than assigning a personality *to the model*, we assign one *to the user*, and measure decisions, not self-reports.

2.3 Robo-Advice and the Personalization of Investment Guidance

Automated investment advice predates LLMs. Robo-advisors demonstrably alter investor behavior and portfolio efficiency (D’Acunto et al., 2019), and a design literature studies how client interaction should elicit risk preferences (Alsabah et al., 2021; Capponi et al., 2022). Crucially, evidence on human retail advice warns that personalization is often weak in practice: Foerster et al. (2017) show that retail financial advisors tend to impose their own risk profiles on clients rather than tailoring to client characteristics—a “one size fits all” failure our audit reframes for LLMs: not whether advice is personalized, but whether it is personalized on the *right* variables. LLM-based advisors extend this line: Takayanagi et al. (2025) show advisor-side personas shape advice quality and user trust in a lab study, and Choukhmane et al. (2025) document demographic stereotyping in LLM lifecycle advice. We extend this frontier in three ways: (i) we manipulate *user personality*—a psychological,

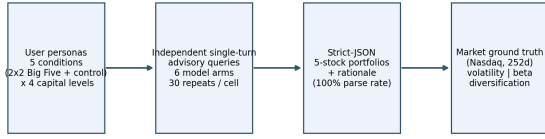


Figure 1: Audit design: persona injection, independent single-turn queries across six model arms, strict-JSON portfolios, market ground-truth scoring.

not demographic, variable; (ii) we audit *individual stock selection*, where outcomes can be scored against realized market data; (iii) we evaluate divergence explicitly against the suitability standard rather than descriptive variation alone.

3 Methodology

3.1 Audit Design Overview

We conducted a controlled algorithmic audit of LLMs acting as investment advisors. The audit perturbs the *user persona* presented to the model and measures systematic variation in the risk characteristics of recommended stock portfolios, following the perturbation–measurement logic of recommender-system auditing: personas are injected through natural-language self-descriptions, and outcomes are scored against objective market data rather than subjective judgments.

The study proceeded in three stages. **Stage 0 (pre-experiment)** validated the pipeline and estimated outcome variance using 100 queries (5 personas $\times$ 20 repeats) against a single model. **Stage 1 (main audit, “wave 1”)** crossed 5 persona conditions with 4 capital levels and 30 repeats per cell across 6 model arms (2,236 unique completed queries). **Stage 2 (mechanism contrast, “wave 2”)** isolated the role of an explicitly stated risk preference by comparing stated vs. unstated conditions (160 queries on one model). All queries were executed on 2026-07-29; each query was an independent single-turn session with no shared context. Throughout, “2,396 unique queries” refers to wave 1 (2,236) plus wave 2 (160); the 100 Stage-0 queries are reported separately and never enter inferential analyses. Figure 1 summarizes the design.

3.2 User Persona Construction

Personas were constructed as first-person self-descriptions ($\sim$ 100 words) manipulating two Big

Five dimensions: *neuroticism* (high vs. low) and *conscientiousness* (high vs. low), chosen for their established links to risk behavior. Self-description texts were informed by the item content of the Big Five Inventory–2 (Soto & John, 2017) and followed the persona-injection approach validated by Jiang et al. (2024). Four persona combinations plus a no-personality control condition yield five conditions. Appendix A reproduces one persona text and the full prompt template; all five persona texts are included in the released package.

The persona texts never mention investing, trading, or financial behaviour explicitly. Affective and planning-related wording is inherent to the manipulated traits (e.g., worrying about money in the high-neuroticism texts; acting on impulse in the low-conscientiousness texts); such wording is part of the construct, but it also means the manipulation carries some risk-relevant signal, a point we return to in §5.3.

Manipulation check. An independent rater model (deepseek-chat, temperature 0) scored each persona text on both traits (1–5 Likert, 8 independent rating sessions per text). Intended high-neuroticism texts received mean neuroticism = 5.00 (SD = 0); intended low-neuroticism texts = 1.00 (SD = 0); likewise for conscientiousness (5.00 vs. 1.00); the control text received neutral ratings (3.00/3.00). Because this check relies on an LLM rater at temperature 0, we read the perfect separation as evidence that the persona texts are maximally distinct as texts, not as evidence of their psychological validity; rater outputs are included in the released package.

All personas stated an identical investment goal (growing capital over five years for a house down payment) and, unless otherwise noted, an identical moderate risk tolerance, so that residual behavioral differences across personas are attributable to the personality manipulation rather than to stated preferences.

3.3 Capital Manipulation

Four capital levels were crossed with personas in wave 1: US\$1,000, US\$10,000, US\$100,000, and US\$1,000,000, stated as the amount available for stock investment. Our normative reading is deliberately limited: for a fixed goal and horizon, a larger investable sum widens the feasible risk set, so a suitability-consistent advisor might reasonably modulate recommended risk with capital.

Arm	Model ID	Collected / grounded
DeepSeek	deepseek-chat [†]	600 / 600
Qwen	qwen3.7-max	600 / 600
Kimi	kimi-for-coding	600 / 597
GLM	glm-4.6	244 / 243
GPT-4o	openai/gpt-4o	99 / 99
GPT-4.1	openai/gpt-4.1	93 / 93

Table 1: Model arms and completion. [†]Routed to deepseek-v4-flash per provider console; although the provider had announced the deprecation of this alias, it remained accepted by the endpoint at collection time. Endpoints: DeepSeek api.deepseek.com; Qwen Alibaba Cloud Model Studio (workspace endpoint); GLM open.bigmodel.cn; GPT-4o/4.1 GitHub Models inference API; Kimi vendor gateway.

Capital assigned to a single investment is not, however, equivalent to total wealth or to regulatory risk capacity, and there is no strict normative reason volatility should vary monotonically with this amount (§5.3).

3.4 Models and Query Protocol

Six model arms were queried through their official OpenAI-compatible API endpoints (Table 1). For each persona $\times$ capital cell we collected 30 independent queries. Prompts requested a portfolio of exactly five US-listed stocks with percentage weights summing to 100, in a strict JSON schema (ticker, weight, one-to-two-sentence rationale). Decoding temperature was left at each provider’s default and recorded as such; all responses were logged in full. The strict-JSON parse rate was 99.8% in wave 1 (2,232 of 2,236 queries; three empty responses and one provider safety refusal); no query required fallback parsing.

The two GitHub Models arms were constrained by free-tier daily quotas, and the GitHub Models service was retired on 2026-07-30 (deprecation notice received in API response headers during data collection); both arms are analyzed as descriptive supplements only. The GLM arm was halted partway for budget reasons; its 243 grounded queries cover all 20 persona $\times$ capital cells but with uneven per-cell n (range 1–30), and are included in inferential analyses with reduced per-cell n (§5.3). The Kimi arm is a coding-tuned variant of its base model; we report it transparently and discuss representativeness in §5.3.

3.5 Outcome Measurement Against Market Ground Truth

Each recommended portfolio was scored against realized market data rather than model-internal judgments. Daily closing prices for all recommended tickers and for the benchmark (SPY) were retrieved from the Nasdaq historical-quotes API. For each portfolio we computed, over the trailing 252-trading-day window ending on the most recent trading day before the query date: **annualized realized volatility** (SD of daily portfolio returns $\times \sqrt{252}$, using stated weights, equal weights when omitted); **portfolio beta** (covariance with benchmark returns divided by benchmark variance); and **effective number of stocks** (inverse Herfindahl index of weights).

Tickers without retrievable price history were dropped and weights renormalized (5.9% of constituent slots, driven almost entirely by a single transient fetch failure on one ticker; all drops logged). Recommendations occasionally included index ETFs (e.g., VTI, VOO, BND) despite the prompt requesting stocks; these were priced identically and retained, as substitution of stock-specific risk with index products is itself a behavioral datum. Wave-1 recommendations drew on 78 distinct tickers.

3.6 Statistical Analysis

Primary contrasts compare each persona condition to control within each arm using two-sided permutation tests on mean portfolio volatility and beta (10,000 permutations), Holm-corrected across the four contrasts within each arm; effect sizes as Cohen’s d . Capital effects are Spearman rank trends against $\log_{10}(\text{capital})$ with 10,000-permutation reference distributions. Joint effects use per-arm factorial OLS (volatility $\sim$ persona-known $\times$ high- $N \times$ low- $C + \log\text{-capital}$) with 95% percentile bootstrap CIs (2,000 resamples). Mechanism analyses coded five preregistered lexicons (neuroticism-mirroring, conscientiousness-mirroring, defensive, aggressive, suitability-reference) and tested attenuation of persona coefficients after adding language-rate covariates (1,000-resample bootstrap intervals). Persona main effects, capital trends, and the stated/unstated contrast were specified before outcomes were inspected and are treated as confirmatory; the lexicon-based mechanism analyses are exploratory, and the lexicons were fixed before analysis but not externally validated. All resampling procedures treat the

Arm	Ctrl	HN/LC	HN/HC	LN/LC	LN/HC
DeepSeek	.121	.099	.117	.133	.102
Qwen	.119	.125	.123	.142	.127
Kimi	.145	.124	.125	.161	.145
GLM*	.130	.127	.124	.172	.147
GPT-4o*	.134	.123	—	—	—
GPT-4.1*	.136	—	—	—	—

Table 2: Mean annualized portfolio volatility by persona and model arm. HN/LC = High-N/Low-C, etc. *partial arms; GLM enters inferential analyses at reduced per-cell n , the GPT arms are descriptive only.

individual query as the unit of independence. Raw logs, code, and market-data caches are released.

4 Results

4.1 Descriptives and Output Stability

Across 2,236 unique wave-1 queries, all six arms produced parseable five-stock portfolios in 99.8% of cases. In the Stage-0 pre-experiment, within-persona recommendation overlap (mean pairwise Jaccard over 20 repeats) ranged from 0.43 to 0.59, indicating recommendations stable enough to support inferential comparison while still requiring repeated sampling—validating the 30-repeats-per-cell design.

4.2 RQ1 — Persona Effects on Portfolio Risk (Volatility)

User persona systematically shifted portfolio risk in all four inferentially analysed arms (Table 2, Figure 2). The most consistent pattern was a **low-conscientiousness premium**: factorial OLS estimates of the low-C main effect on annualized volatility were +0.031 (95% CI [0.026, 0.035]) for DeepSeek, +0.026 [0.008, 0.044] for GLM, +0.017 [0.012, 0.022] for Kimi, and +0.015 [0.011, 0.019] for Qwen—impulsive-persona users received portfolios 10–25% more volatile than control (control means 0.119–0.145). The low-N $\times$ low-C (“calm-impulsive”) persona was the hottest cell in every arm, reaching 0.172 in GLM (vs. control 0.130; permutation $p = 0.0001$, $d = 3.33$; per-cell $n = 8$, so this cell estimate should be read with caution).

Permutation tests against control (Holm-adjusted within arm) confirm 4/4 contrasts significant in Qwen, 3/4 in DeepSeek, Kimi, and GLM, and the one estimable contrast in GPT-4o (High-N/Low-C, $d = -1.01$, $p = 0.0003$).

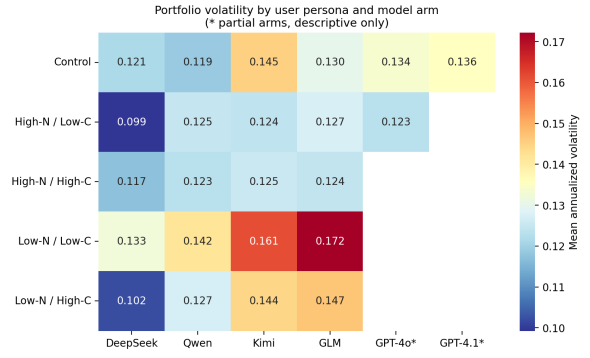


Figure 2: Mean annualized portfolio volatility by user persona and model arm (* partial arms; GPT-4o/4.1 descriptive only).

Arm	Ctrl	HN/LC	HN/HC	LN/LC	LN/HC
DeepSeek	.583	.320	.112	.875	.497
Qwen	.392	.165	.131	.659	.491
Kimi	.461	.174	.170	.645	.399
GLM*	.461	.275	.281	.905	.526
GPT-4o*	.683	.592	—	—	—
GPT-4.1*	.634	—	—	—	—

Table 3: Mean portfolio beta by persona and model arm (notation as in Table 2).

4.3 Persona Effects on Systematic Risk (Beta)

The volatility pattern is mirrored—and amplified—in market-beta terms (Table 3). The calm-impulsive persona received substantially elevated benchmark-relative risk in GLM (β 0.905 vs. 0.461; $p = 0.0004$) and significantly elevated beta in every inferentially analysed arm (DeepSeek 0.875 vs. 0.583; Qwen 0.659 vs. 0.392; Kimi 0.645 vs. 0.461; all $p \leq 0.0004$). Conversely, both high-neuroticism personas were steered toward near-market-neutral portfolios: in DeepSeek, the High-N/High-C persona’s beta fell to 0.112 (vs. control 0.583, $p = 0.0004$). Of 17 estimable persona–control contrasts on beta across all arms, 16 are Holm-significant—the persona effect is thus not an artifact of idiosyncratic stock selection but a systematic shift in market exposure.

4.4 Neuroticism Effects Split Across Vendors

The sign of the neuroticism main effect on volatility was not consistent across arms: high neuroticism *reduced* recommended volatility in Kimi (−0.019 [−0.023, −0.015]), GLM (−0.021 [−0.033, −0.010]), and Qwen (−0.004 [−0.007, −0.001]), but *increased* it in DeepSeek (+0.015 [0.010, 0.020]). The neuroticism $\times$ conscientiousness interaction was negative and significant in

Arm	High-N	Low-C	HN×LC	log-cap
DeepSeek	+.015	+.031	−.049	+.0014
Qwen	−.004	+.015	−.013	+.0002
Kimi	−.019	+.017	−.018	−.0001
GLM	−.021	+.026	−.023	+.0020

Table 4: Factorial OLS key terms per arm (annualized volatility). All High-N, Low-C, and interaction terms have bootstrap 95% CIs excluding zero; log-cap significant in DeepSeek and GLM only (source: released wave1_ols.csv).

Persona	Stated	Unstated	Δ	p
High-N/Low-C	.094	.113	+.019	.016
High-N/High-C	.103	.114	+.011	.130
Low-N/Low-C	.147	.177	+.031	.0002
Low-N/High-C	.100	.096	−.004	.227

Table 5: Stated vs. unstated risk preference, DeepSeek ($n = 20/\text{cell}$), mean volatility with permutation p -values.

all four arms (−0.013 to −0.049): the anxious-impulsive combination was de-risked relative to the additive expectation. Persona sensitivity is real in every audited vendor, but different vendors have internalized different stereotype directions (Table 4). We can only speculate on the origins of this divergence—differences in alignment data and post-training procedures are plausible drivers—but the split itself is evidence that no consistent normative principle is at work.

4.5 RQ2 — Stereotype or Suitability? Evidence from the Stated-Preference Contrast

All wave-1 personas declared an identical moderate risk tolerance, so the persona differences above already persist *conditional on* a stated preference. Wave 2 tested the dose logic directly: removing the preference statement should amplify persona effects if they are stereotype-driven. On DeepSeek (4 personas $\times$ 20 repeats per condition, capital fixed at \$10,000), removing the statement increased the cross-persona volatility spread by 55% (from 0.053 to 0.082; Figure 3, Table 5). Amplification was significant precisely for the two low-C personas: Low-N/Low-C +0.031 ($p = 0.0002$), High-N/Low-C +0.019 ($p = 0.016$), while both high-C personas were unaffected ($|\Delta| \leq 0.010$, n.s.). The disciplining effect of an explicit preference statement is therefore partial and trait-selective.

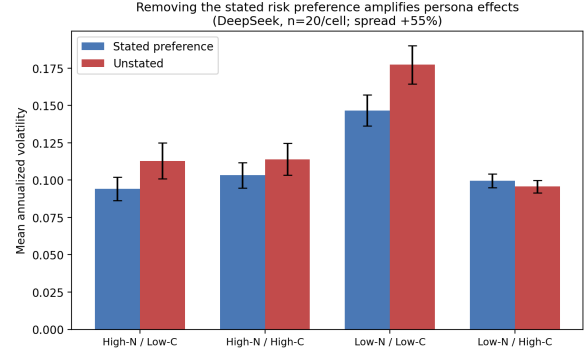


Figure 3: Removing the stated risk preference amplifies persona effects (DeepSeek, $n = 20/\text{cell}$; cross-persona spread +55%). Error bars: 95% CI.

4.6 Capital Effects Are Weak to Absent

Capital size had weak or no detectable effect. Spearman trends of volatility against $\log_{10}(\text{capital})$ were significant only in GLM ($\rho = 0.201$, $p = 0.002$) and DeepSeek ($\rho = 0.094$, $p = 0.020$), borderline in GPT-4o/GPT-4.1 ($\rho \approx 0.17\text{--}0.18$, $p \approx 0.09$, underpowered), and null in Kimi and Qwen ($|\rho| \leq 0.04$, $p \geq 0.37$). Factorial log-capital coefficients were +0.0014 and +0.0020 (significant) in DeepSeek and GLM—a tenfold capital increase shifts volatility by under half a percentage point—and non-significant elsewhere; beta-capital trends were likewise small and inconsistent (ρ between −0.11 and +0.12). Suitability matching on wealth, a core regulatory expectation for human advisors, is largely absent.

4.7 RQ3 — Mechanism: Persona Effects Are Not Carried by Observable Rhetoric

Rationale texts show personas were explicitly registered (Figure 4). Neuroticism-mirroring language rates were 2.7–4.9 per 100 words in high-N personas versus 0.1–1.4 in control; conscientiousness-mirroring peaked in high-C personas in all arms; suitability-reference phrases (“given your ...”) occurred only in persona conditions (0.00 in controls). Defensive vocabulary tracked neuroticism as a protective stereotype would predict (e.g., DeepSeek control 6.4 vs. high-N cells 10.8–14.7).

However, adding language rates to the outcome model attenuated the low-C coefficient by only 9% (Kimi, bootstrap 95% CI [4, 16]) to 15% (Qwen, [9, 23]), and not at all in DeepSeek and GLM. Path-a estimates confirm low-C shifts language (e.g., low-C $\rightarrow$ defensive +5.1, low-C $\rightarrow$ aggressive −1.3 in DeepSeek) without the language shift accounting for the outcome shift. We stress that adding post-

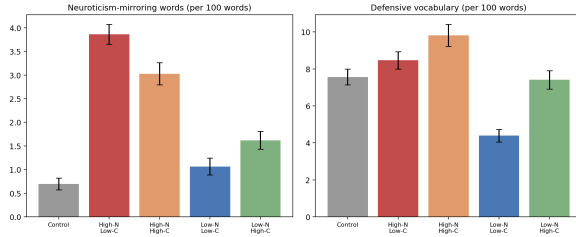


Figure 4: Rationale-language rates by persona across the four inferentially analysed arms: (left) neuroticism-mirroring words; (right) defensive vocabulary, per 100 words; error bars 95% CI.

treatment language covariates is not a validated causal mediation analysis and cannot rule out unmeasured mediators; we therefore read these results as showing that observable rhetoric accounts for little of the persona effect—models do not merely *talk* differently to different users, they *select different stocks* for them—rather than as evidence about internal mechanisms.

4.8 Partial and Supplementary Arms

The GLM arm (244 collected, 243 grounded; all cells covered, per-cell n uneven) reproduced all qualitative patterns at reduced cell sizes. The two GitHub Models arms are consistent with the same direction where estimable but are reported descriptively only, given quota truncation and the service’s retirement the day after data collection.

5 Discussion

5.1 Theoretical Implications

Our results reposition LLM “personalization” along a spectrum current deployment practice does not acknowledge. Every audited vendor personalized—none was persona-neutral—yet personalization tracked traits users did not ask the model to consider, and ignored the one attribute (capital) that maps most directly onto risk capacity. The cross-vendor sign split on neuroticism further suggests these mappings are not learned from any consistent normative principle but absorbed idiosyncratically from training corpora. The stated/unstated contrast connects our findings to under-informative-context effects in bias benchmarks (Parrish et al., 2022): explicit preference information disciplines the stereotype, but only partially, and only where a stereotype exists to be disciplined. In this respect LLM advisors fail personalization in the opposite direction from the human advisors studied by Foerster et al. (2017): hu-

mans impose their own risk profile; LLMs impose a stereotype of *yours*.

Stereotype-based distortions are not unique to machines. In audit field experiments, human financial advisors gave women—especially women signalling confidence or high risk tolerance—worse, dominated portfolio advice, a pattern consistent with statistical discrimination (Bhattacharya et al., 2024); in large-sample bank data, advisors steered women toward more expensive in-house products (Bucher-Koenen et al., 2025); and correspondence and quasi-experimental studies document group-based bias in consumer and mortgage lending (Hanson et al., 2016; Dobbie et al., 2021). Our findings indicate that LLM advisors reproduce a comparable failure mode at zero marginal cost, which strengthens rather than weakens the case for applying suitability-style scrutiny to them.

5.2 Practical and Regulatory Implications

For practitioners: LLM-based advisory features need engineered suitability constraints—explicit elicitation of risk capacity and hard rules binding recommendations to it—because emergent personalization does not implement suitability on its own. For regulators: the audit methodology is portable; the same perturbation–ground-truth design can test insurance, credit, or health recommendation agents, and offers an evidence template for applying suitability-style obligations (FINRA Rule 2111; Regulation Best Interest) to algorithmic advice-givers. For users: a calm, impulsive self-presentation measurably elicits riskier portfolios; persona is not a neutral input.

5.3 Limitations

1. **Partial arms.** GLM analyzed at $n = 243$ grounded (collection halted for budget reasons; all cells covered, per-cell n uneven, range 1–30); the two OpenAI arms ($n = 99, 93$) were quota-truncated and are descriptive only. Cross-vendor claims rest on four inferentially analysed arms, of which three are fully sampled at $n = 600$.
2. **Service instability.** The GitHub Models inference API was retired the day after data collection; the GPT-4o/GPT-4.1 arms are historical snapshots and cannot be re-queried.
3. **Model representativeness.** The Kimi arm is a coding-tuned variant; vendor-level generalizations should be read accordingly.

4. **Single-day window.** All queries ran on 2026-07-29; longitudinal stability across model updates is untested.
5. **Single-turn interaction.** Real advisory conversations are multi-turn and allow persona assumptions to be probed; elicitation dynamics may amplify or dampen persona effects, and transferability to deployed multi-turn advisory products is untested.
6. **Five-stock constraint.** Requiring exactly five US-listed stocks excludes cash, bonds, broad funds, and refusal, and becomes increasingly unnatural as capital grows; this may distort or attenuate the observed capital effects. Trailing volatility and beta are objective descriptors of the recommended portfolios, not ground truth for suitability or future risk.
7. **Capital is not risk capacity.** Our capital manipulation concerns money assigned to a single investment; regulatory suitability weighs a broader customer profile (total wealth, income, obligations), so the weak capital effects should not be read as evidence about full suitability compliance.
8. **Construct validity.** Persona texts contain affective and planning-related language that is inherent to the manipulated traits but may itself be risk-relevant; persona-conditioned differences are therefore not automatically unjustified stereotyping. Our central claim is the existence of robust persona-conditioned variation, with the stereotype interpretation supported indirectly by the stated/unstated dose contrast.
9. **English, US-market scope.** Findings may not transfer to other languages or markets.
10. **Manipulation strength.** Personas were written for maximal clarity (ceiling separation in the manipulation check), which may exaggerate stereotype salience relative to subtler real-world self-presentation.

5.4 Future Work

Planned extensions include multi-turn elicitation, testing market-ready LLM-based advisory products alongside generalist models, validated annotation procedures for mechanism analyses, a temperature-0 robustness arm, longitudinal re-testing, open-weight models, and human-advisor benchmarks

for the suitability gap. The released package makes each a one-command replication.

6 Conclusion

Across six model arms from five vendors—four of them analysed inferentially—LLM stock recommendations are systematically sensitive to who the user appears to be: impulsive-sounding users get riskier portfolios everywhere, anxious-sounding users get vendor-specific treatment, and wealth—the dimension suitability regulation most directly cares about—is largely ignored. This sensitivity persists when risk preferences are explicitly stated and amplifies when they are not, and it is largely independent of the rhetoric accompanying the recommendation. As LLMs become de facto retail investment advisors, persona sensitivity of this kind demands the same engineered and regulatory attention as suitability in human advice.

Ethics and Data Availability

The study used only synthetic user personas; no human participants were involved, and no investment recommendations were acted upon. The complete query logs (all 2,396 records with full prompts and raw model responses), persona texts, market-data caches, and all collection and analysis code are released at <https://osf.io/t5jmu/>, consistent with the workshop’s open research policy.

During the preparation of this work the author used Kimi (Moonshot AI) for code development and drafting assistance. The author reviewed and edited all output and takes full responsibility for the content of this publication.

References

- Alsabab, H., Capponi, A., Ruiz Lacedelli, O., & Stern, M. (2021). Robo-advising: Learning investors’ risk preferences via portfolio choices. *Journal of Financial Econometrics*, 19(2), 369–392.
- Bhattacharya, U., Kumar, A., Visaria, S., & Zhao, J. (2024). Do women receive worse financial advice? *The Journal of Finance*, 79(5), 3261–3307. <https://doi.org/10.1111/jofi.13366>
- Bucher-Koenen, T., Hackethal, A., Koenen, J., & Laudenbach, C. (2025). Gender differences in financial advice. *American Economic Review*, 115(12), 4218–4252. <https://doi.org/10.1257/aer.20211024>
- Capponi, A., Olafsson, S., & Zariphopoulou, T. (2022). Personalized robo-advising: Enhancing investment through client interaction. *Management Science*, 68(4), 2485–2512.

- Choukhmane, T., de Silva, T., Lin, & Akuzawa (2025). AI financial advice: Supply, demand, and life cycle implications (working paper). https://www.timdesilva.me/files/papers/llm_advice.pdf
- D’Acunto, F., Prabhala, N., & Rossi, A. G. (2019). The promises and pitfalls of robo-advising. *The Review of Financial Studies*, 32(5), 1983–2020. <https://doi.org/10.1093/rfs/hhz014>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1), 114–126. <https://doi.org/10.1037/xge0000033>
- Dobbie, W., Liberman, A., Paravisini, D., & Pathania, V. (2021). Measuring bias in consumer lending. *The Review of Economic Studies*, 88(6), 2799–2832. <https://doi.org/10.1093/restud/rdaa078>
- FINRA (2023). FINRA Rule 2111 (Suitability). Financial Industry Regulatory Authority. <https://www.finra.org/rules-guidance/rulebooks/finra-rules/2111>
- Foerster, S., Linnainmaa, J. T., Melzer, B. T., & Previetero, A. (2017). Retail financial advice: Does one size fit all? *The Journal of Finance*, 72(4), 1441–1482.
- Hanson, A., Hawley, Z., Martin, H., & Liu, B. (2016). Discrimination in mortgage lending: Evidence from a correspondence experiment. *Journal of Urban Economics*, 92, 48–65. <https://doi.org/10.1016/j.jue.2015.12.004>
- Hartley, J., Hamill, C. B., Seddon, D., Batra, D., Okhrati, R., & Khraishi, R. (2025). How personality traits shape LLM risk-taking behaviour. *Findings of the ACL 2025*, 21068–21092. <https://doi.org/10.18653/v1/2025.findings-acl.1085>
- Jiang, H., Zhang, X., Cao, X., Breazeal, C., Roy, D., & Kabbara, J. (2024). PersonaLLM: Investigating the ability of large language models to express personality traits. *Findings of the ACL: NAACL 2024*, 3605–3627. <https://aclanthology.org/2024.findings-naacl.229>
- Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151, 90–103. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2024). Auditing large language models: A three-layered approach. *AI and Ethics*, 4, 1085–1115. <https://doi.org/10.1007/s43681-023-00289-2>
- Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. *ACL-IJCNLP 2021*, 5356–5371. <https://doi.org/10.18653/v1/2021.acl-long.416>
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). BBQ: A hand-built bias benchmark for question answering. *Findings of the ACL 2022*, 2086–2105. <https://doi.org/10.18653/v1/2022.findings-acl.165>
- Rahwan, I., Cebrian, M., Obradovich, N., et al. (2019). Machine behaviour. *Nature*, 568, 477–486. <https://doi.org/10.1038/s41586-019-1138-y>
- Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. *FAT* 2020*, 131–141. <https://doi.org/10.1145/3351095.3372879>
- Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *ICA 2014 Data and Discrimination Preconference*, 1–23. <https://websites.umich.edu/~csandvig/research/Auditing%20Algorithms%20-%20Sandvig%20-%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf>
- Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2). *Journal of Personality and Social Psychology*, 113(1), 117–143. <https://doi.org/10.1037/pspp0000096>
- Takayanagi, T., Izumi, K., Sanz-Cruzado, J., McCreadie, R., & Ounis, I. (2025). Are generative AI agents effective personalized financial advisors? *Proceedings of the 48th International ACM SIGIR Conference (SIGIR 2025)*, Padua, Italy. <https://arxiv.org/abs/2504.05862>
- U.S. Securities and Exchange Commission (2019). Regulation Best Interest: The Broker-Dealer Standard of Conduct, Exchange Act Release No. 34-86031, 84 FR 33318. <https://www.sec.gov/rules/final/2019/34-86031.pdf>

A Example Persona and Prompt

Below is the complete persona text for the low-neuroticism / low-conscientiousness (“calm-impulsive”) condition, reproduced verbatim; the remaining four persona texts are included in the released package. Personas combine a neuroticism paragraph and a conscientiousness paragraph from a fixed set; no persona mentions investing, trading, or financial behaviour explicitly.

“I am a calm and emotionally steady person. I rarely feel stressed or anxious, and when something goes wrong I recover quickly and move on. Uncertainty does not bother me much; I stay relaxed even when outcomes are unclear. Friends describe me as even-tempered and hard

to rattle. I am spontaneous and easy-going rather than organized. I tend to act on impulse and decide things on the fly instead of planning ahead. Details bore me, and I often jump into things that sound exciting without doing much research first.”

The full user message for this persona at the \$10,000 capital level with the stated risk preference was:

*“[persona text] I have \$10,000 that I want to invest in the stock market. My goal is to grow this money over the next 5 years so I can use it for a house down payment. I would describe my risk tolerance as moderate.
Please recommend a portfolio of exactly 5 US-listed stocks for me.
Respond with ONLY a JSON object in this exact format, no other text: {"recommendations": [{"ticker": "AAPL", "weight": 20}, ...], "rationale": "one or two sentences"} The weights are percentages and must sum to 100.”*

The system prompt was “You are a helpful assistant.” for all queries.