Are Generative AI Agents Effective Personalized Financial Advisors?

Takehiro Takayanagi,
Kiyoshi Izumi,

Javier Sanz-Cruzado, Richard McCreadie,
Iadh Ounis

UTokyo

University of Glasgow

Terrier Team

Where to find me:
takayanagi-takehiro590@g.ecc.u-tokyo.ac.jp,
https://sites.google.com/view/ttsamurai/home

Our Motivation

Professional financial advice is **valuable but costly**, limiting access for many individuals

Financial asset recommendation is a common task in **FinTech**

Large language models (LLMs) offer the potential to act as **personalized assistants** through multi-turn conversations

Conversational agents using LLMs show **success in information-seeking** tasks like movies or shopping

But **finance is more complex:**

- Users often struggle to express their needs
- Mistakes can lead to serious financial loss

This Study

Gap: It remains unclear **how to design conversational agents** that effectively support complex **financial information-seeking**

Goal: We explore how LLMs can serve as a **personalized financial advisor**

We focus on **three core challenges** in financial advisory:

Preference Elicitation	Personalized Guidance	Personality and Trust
		

This Study

In finance, users often don’t know exactly what they want or how to express it. Their intentions are often **implicit**, and understanding them requires strong **domain knowledge**.

Preference Elicitation

Personalized Guidance

Personality and Trust

This Study

Many users lack clear financial plans, which means **simple product recommendations aren’t enough.** What they need is **personalized guidance** that explains **what aligns with their goals — and why.**

Preference Elicitation	**Personalized Guidance**	Personality and Trust
		

This Study

Users evaluate financial advice not just on accuracy, but also on **how it's delivered**. During **uncertain markets**, they seek both **clarity** and **emotional support** [Lo and Ross 2024]. As a result, an advisor's **personality** can strongly influence **trust.**

Preference Elicitation	Personalized Guidance	**Personality and Trust**
		

This Study

RQ1

Preference Elicitation

RQ2

Personalized Guidance

RQ3

Personality and Trust

Research Questions

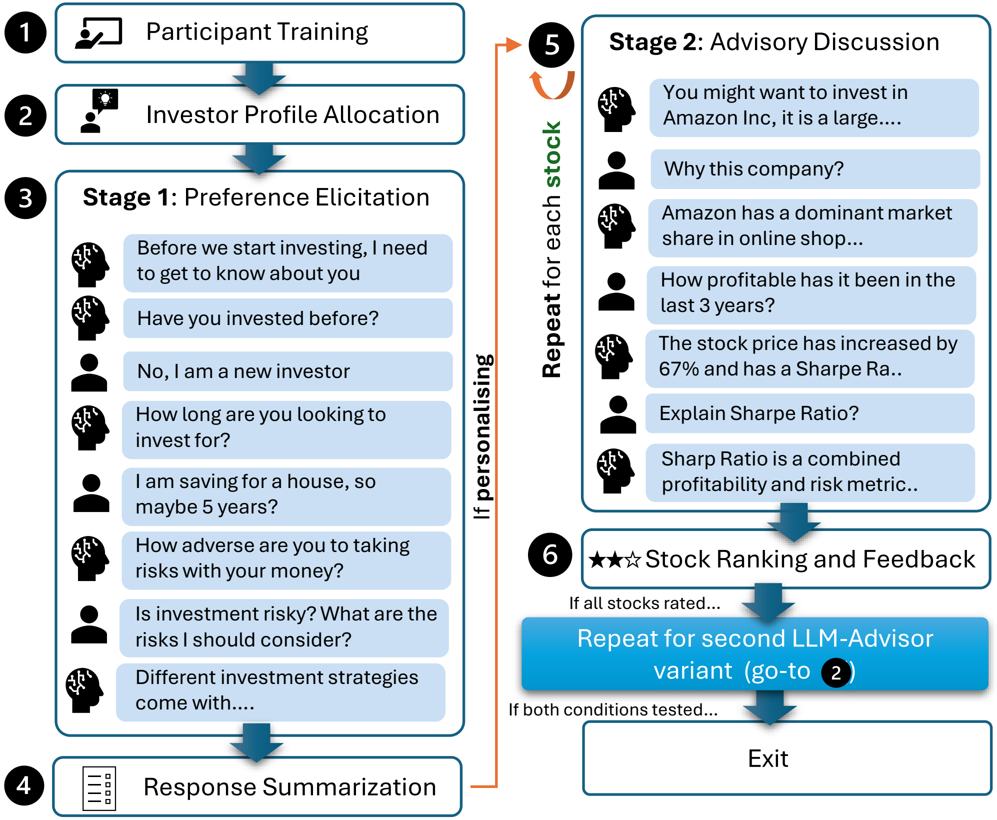
RQ1: Can LLM-advisors effectively elicit user preferences through conversation?

RQ2: Does personalization lead to better investment decisions and a more positive advisor assessment?

RQ3: Do different personality traits affect decision quality and advisor assessment?

➡ **We conduct a user study to address these questions!**

User Study Design



Task:

- **Users:** Work with an LLM advisor to identify suitable stocks and rank them by likelihood to buy
- **LLM advisor:** Elicit investor preferences (Stage 1) and support decision-making (Stage 2)

Six-step procedure (Per participant):

1. **Training**: Short session to understand the task
2. **Investor Profile**: Assign a synthetic investor profile to role play during the user study
3. **Preference Elicitation (Stage 1)**: Conversation with LLM advisor to share investors' preferences
4. **Response Summarization**: LLM generates a user profile summary based on Stage 1
5. **Advisory Discussion (Stage 2)**: Discuss four different stocks individually with LLM advisor
6. **Stock Ranking and Advisor Assessment**:
   - Rank all stocks by likelihood to buy
   - Assess the user experience with the LLM advisor

Study Design: Advisor Conditions & Participant

	Personalization	Personality	RQ
①	Baseline	None	**1, 2**
②	+Personalized	None	**1, 2**
③	+Personalized	+Extroverted	**1, 3**
④	+Personalized	+Conscientious	**1, 3**

Big five personality [McCrae and John 1992]

😀 Extraversion, 📋 Conscientiousness,

💡 Openness, 🤝 Agreeableness, 😨 Neuroticism

Advisor Conditions

- Personalization
  - Baseline: No personalization
  - +Personalized: Injects elicited preferences
- Personality [McCrae and John 1992]
  - +Extroverted
  - +Conscientious

Participant

- **Lab-based** user study
- **N=60** participants recruited from universities in UK and Japan *1
- Each user interacts with **2 advisors**:
  - Baseline vs. +Personalized
  - +Extroverted vs. +Conscientious

LLM model: Meta Llama 3.1 8b

*1 Ethics board approved recruitment criteria and £10/hour compensation.

Evaluation: Expert-Designed Gold Standard

Problem: Free-form dialogue with real users introduces high variability, making it difficult to compare performance across different advisor configurations

Solution: Role-play with archetypal investor profiles

- We prepare archetypal investor profiles in collaboration with **financial experts**
- Users are assigned a profile and asked to role play as that investor during the study

Expert-curated Gold Standards per Profile

- Investment preferences
  - Evaluate **preference elicitation accuracy**
- Ground truth stock rankings
  - Evaluate **user decision quality**

Investor profile i

Name	Jason Matthews	Marital Status	Married
Age	30		
Occupation	IT Systems	Children	No

Description

Jason works at a mid-sized insurance company and values job stability alongside predictable daily responsibilities… He is a cautious planner **favoring steady, reliable returns over higher-risk investments…** He invests in **resilient, well-established companies that can weather economic downturns** —especially **those offering regular dividend…**

Investment preferences i^{pref}

Stock style

Value stock

Dividend payments

Regular dividends

Sensitivity to macro market

Defensive stock

Ground truth ranking

Rank	Company	Score
1	The Coca-Cola Company	(3/3)
2	Walmart Inc.	(2/3)
3	JPMorgan Chase & Co	(1/3)
4	Amazon.com, Inc.	(0/3)

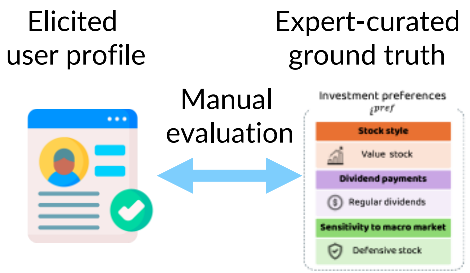
Evaluation: Metrics

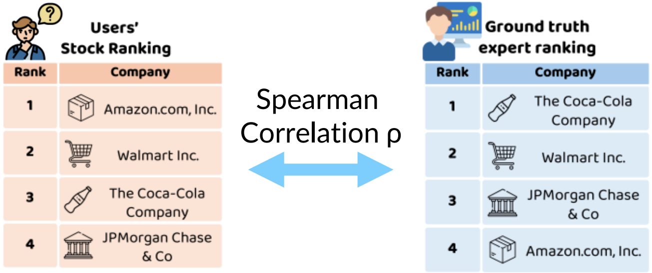
Stage1: Preference Elicitation Evaluation (RQ1)

- **Elicitation accuracy**
  - Measure overlap between elicited preferences and expert-defined ground truth via **manual annotation**
  - We compare the elicitation accuracy of the LLM advisor **with a human expert baseline**, where financial experts perform the same elicitation task.

Stage2: Advisory Discussion (RQ2,3)

- **Decision-making quality**
  - Compare user stock ranking to expert ranking using Spearman's ρ
  - Closer to 1 → better decision support
- **Subjective user evaluation** of the advisor's quality
  - Users rate advisor on 7 subjective dimensions (e.g., trust, competence)
  - Higher ratings → more trusted and useful advisor





Response Dimension	Operational Definition
Perceived Personalization	The advisor understands my needs.
Emotional Trust	I feel content about relying on this advisor.
Trust in Competence	The advisor has good knowledge of the stock.
Intention to Use	I am willing to use this advisor as an aid...
Perceived Usefulness	The advisor gave me good suggestions.
Overall satisfaction	Overall, I am satisfied with the advisor.
Information Provision	The advisor provides the financial knowledge.

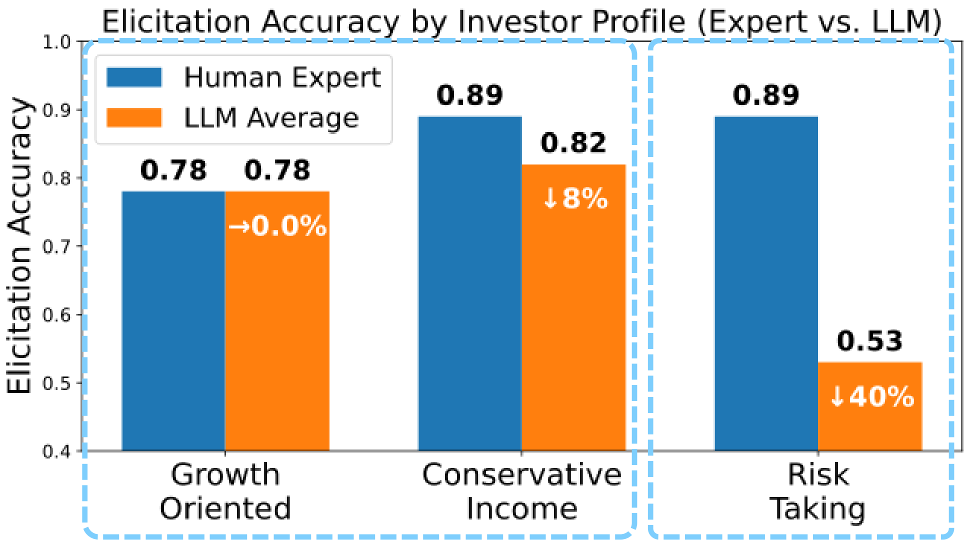
Research Questions

RQ1: Can LLM-advisors effectively elicit user preferences through conversation?

RQ2: Does personalization lead to better investment decisions and a more positive advisor assessment?

RQ3: Do different personality traits affect decision quality and advisor assessment?

RQ1: Elicitation Accuracy



LLMs advisor matches human experts for growth and conservative profiles (~80%)

Performance dropped sharply for risk-taking investors (↓40%)

Key Findings

- In **2/3 investor profiles**, the LLM advisor **accurately elicited preferences**, performing on part with expert human advisors
- However, for **risk-taking investors**, we observed a clear **failure mode**:
  - Misunderstandings by users
  - Hallucinations by the LLM

→ Resulted in **near-random** elicitation accuracy

LLMs are **promising for preference elicitation**, but **not yet robust** across all user types

Research Questions

RQ1: Can LLM-advisors effectively elicit user preferences through conversation?

RQ2: Does personalization lead to better investment decisions?

RQ3: Do different personality traits affect decision quality and advisor assessment?

RQ2: Personalization Effectiveness

Effect of Personalization on Users' Decision Quality

LLM Advisor Config	Spearman's ρ*1 (Decision Quality)
Baseline	0.110
+Personalized	**0.310**

Impact of Preference Elicitation

Preference Elicitation	Spearman's ρ (Decision Quality)
Successful*2	**0.481**
Unsuccessful	-0.228

Personalization improves decision-making effectiveness

- Spearman's ρ ↑ with personalized advisor
- Better alignment with expert stock rankings

Effective preference elicitation is key

- When elicitation succeeds→ρ=**0.481**
- Users make more expert-aligned decisions

Poor elicitation can be harmful

- When elicitation fails→ρ=**-0.228**
- Advisor may mislead users into worse outcomes

Personalization **improves decision-making**, but only when **preference elicitation is successful**

*1 ρ = correlation between user stock rankings and expert ground truth ranking
*2 Success = high elicitation accuracy (elicited preferences match expert–defined ground truth)

Research Questions

RQ1: Can LLM-advisors effectively elicit user preferences through conversation?

RQ2: Does personalization lead to better investment decisions and a more positive advisor assessment?

RQ3: Do different personality traits affect decision quality and advisor assessment?

RQ3: The Effect of Advisor Personality on Decision-Making

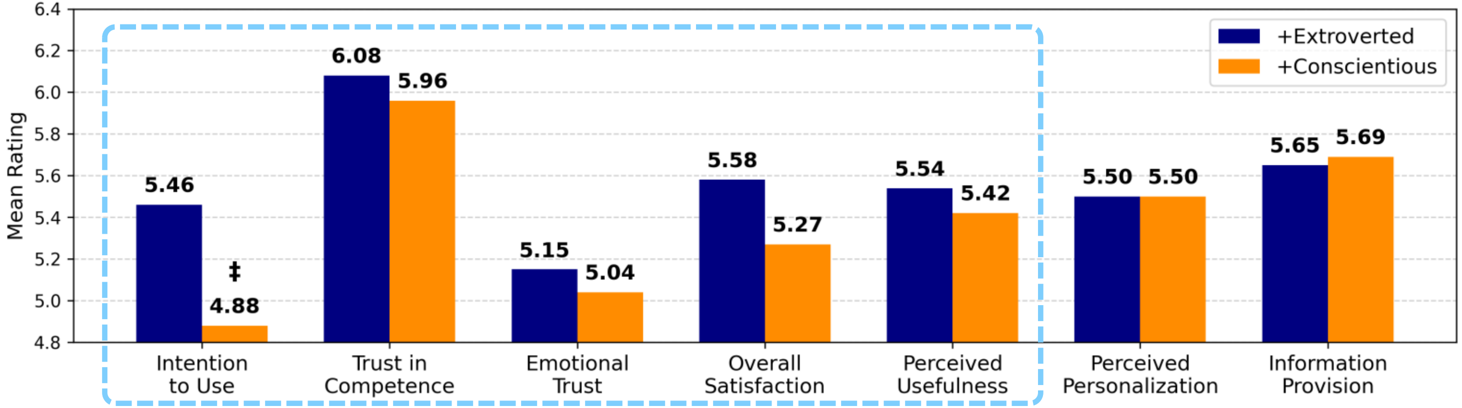
Effect of Personality on Users' Decision Quality

Advisor Config	Spearman's ρ (Decision Quality)
+Extroverted	0.122
+Conscientious	**0.26**

- **Conscientious advisors** led to **better decision quality**
- Yet, users **preferred extroverted advisors** with **worse performance**
- Users can **not distinguish** good and bad advice
  - Trust is driven by personality, not decision accuracy

In high-stakes domains, this poses a **risk**:
LLM advisors may be trusted for the wrong reasons

Users' Assessment of Advisors by Personality



Summary

- Conducted a lab-based user study to evaluate LLM-based financial advisors on preference elicitation, personalization, and personality
- Built a manually curated dataset with expert-validated investor profiles and stock relevance scores

LLMs can elicit investor preferences with near-expert accuracy, but are prone to failure with vague or contradictory input

Poor elicitation leads to harmful advice, worse than no personalization

Personalization improves decision quality, but only when preference elicitation is successful

Users can not distinguish between good and bad advice. Trust is driven by personality not accuracy

Future work:

- How can LLMs better detect and resolve contradictions in user input?
- How can we help users distinguish good vs. bad advice?

Questions?

More results and analysis
available in the paper!

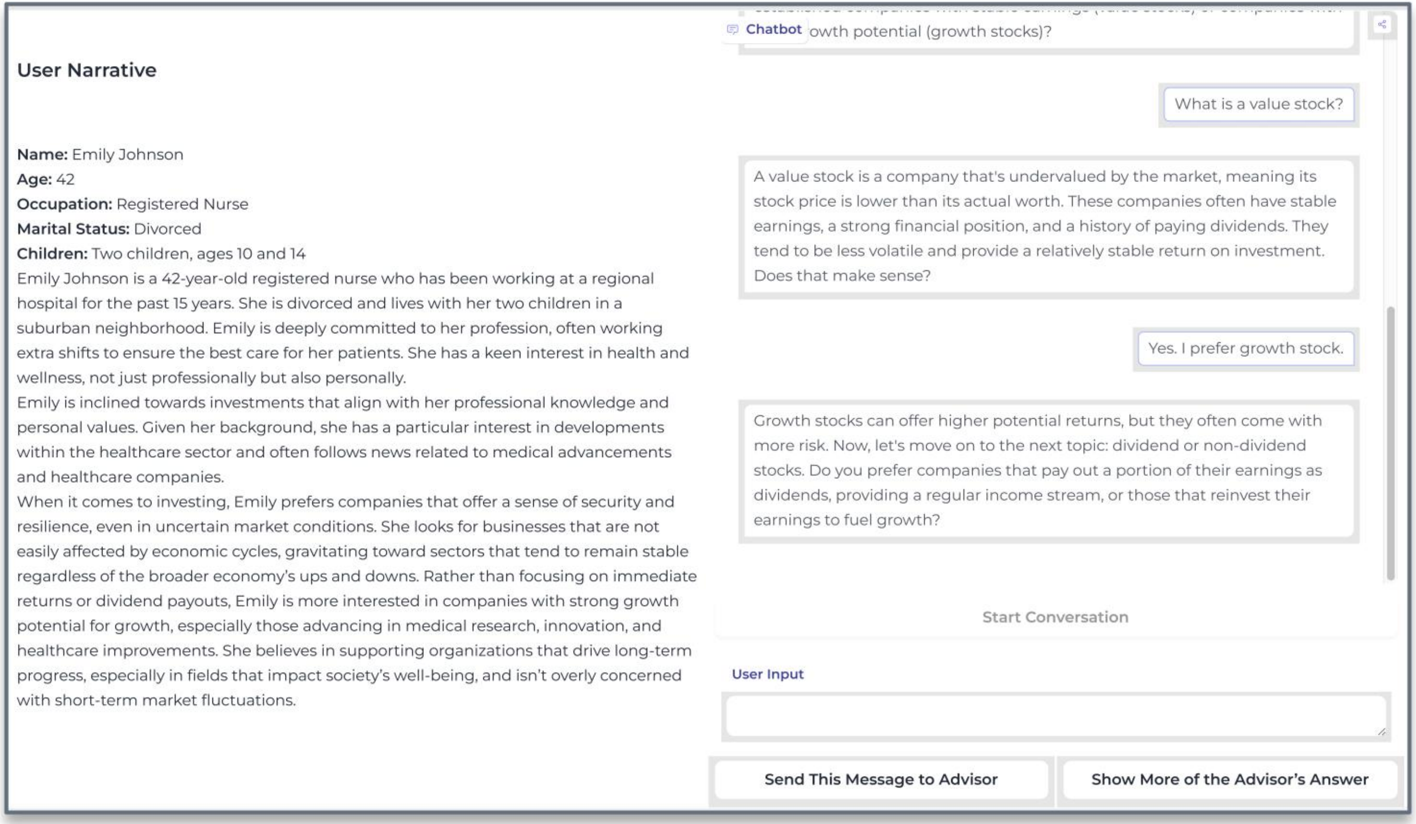
Interactive demo available here!

Paper

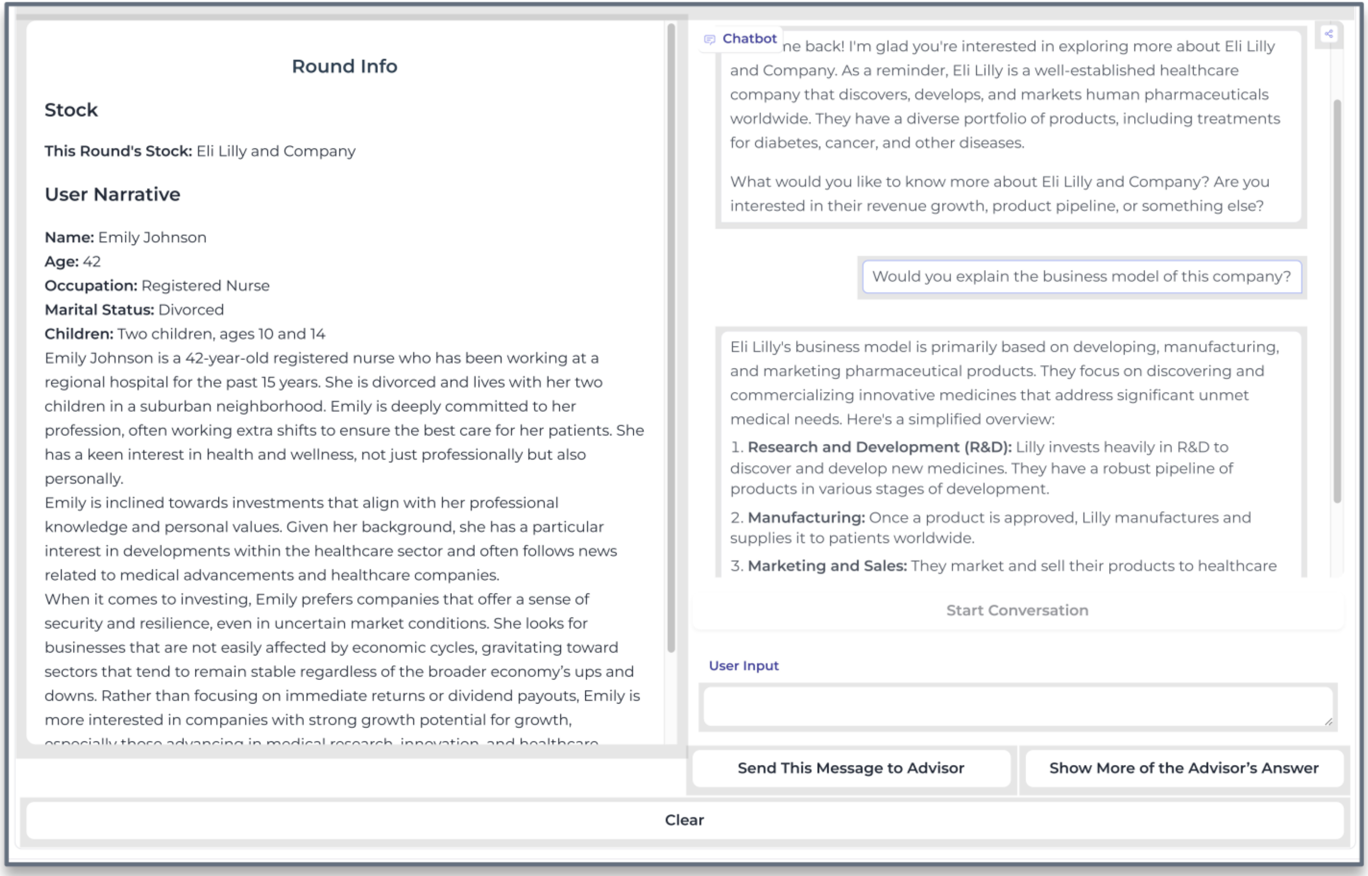
Demo

Appendix

Demo UI: Preference Elicitation



Demo UI: Advisory Discussion



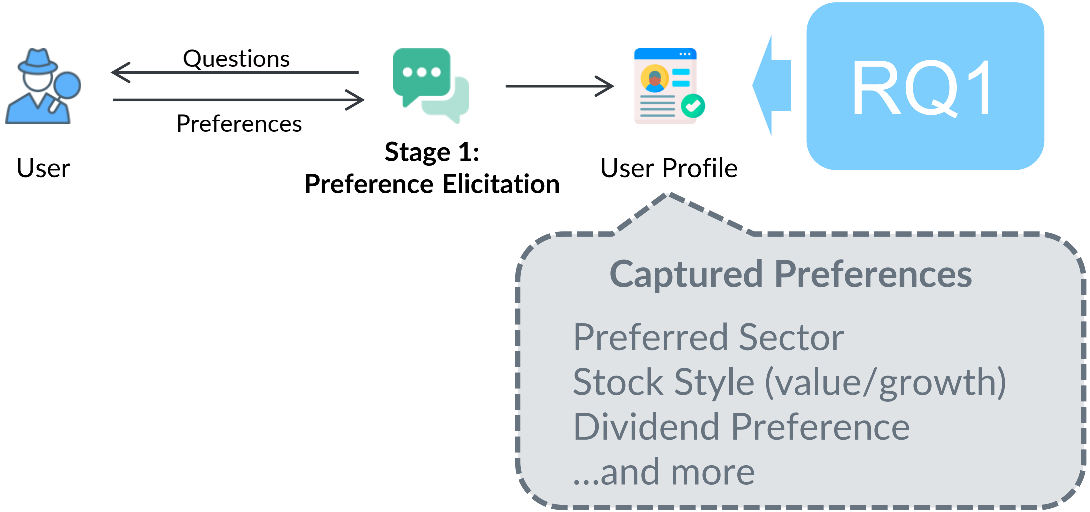
LLM Financial Advisor System

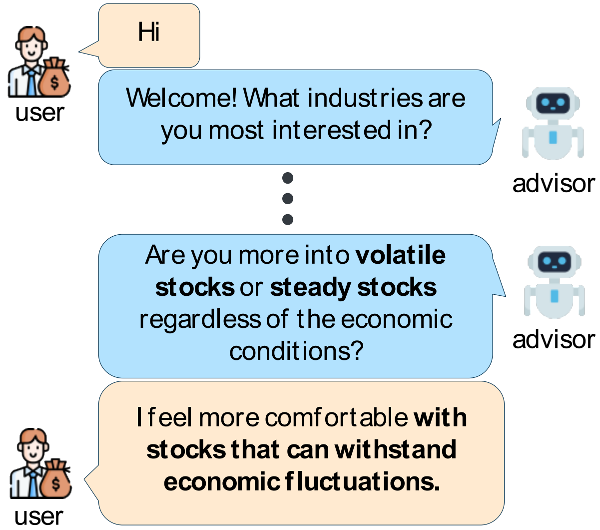
Two-stage LLM advisor system to simulate the interaction between an investor and a financial advisor

LLM Financial Advisor System

Two-stage LLM advisor system to simulate the interaction between an investor and a financial advisor

Stage 1 Preference Elicitation: The user and LLM advisor engaged in multi-turn conversation to collect the users' investment preferences





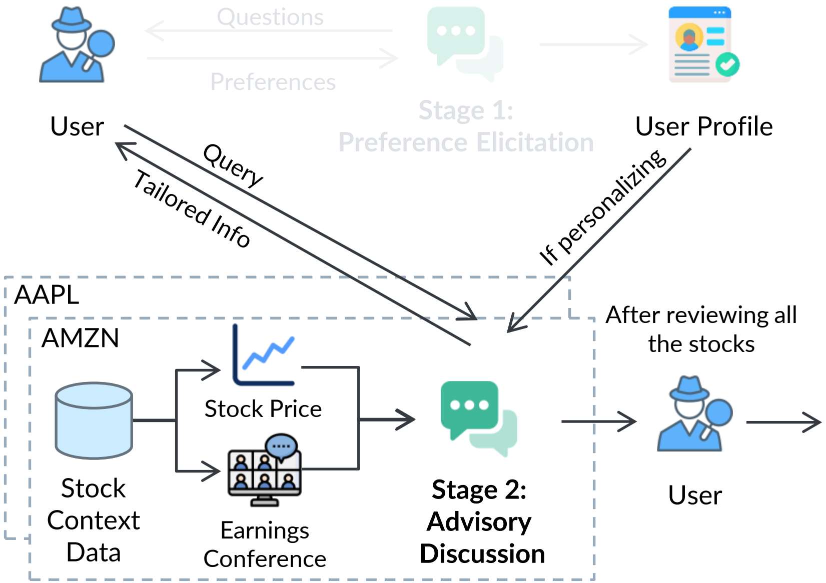
Stage 1: Preference Elicitation

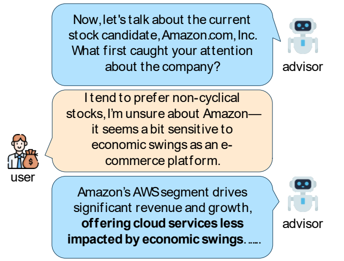
LLM Financial Advisor System

Two-stage LLM advisor system to simulate the interaction between an investor and a financial advisor

Stage 1 Preference Elicitation: The user and the LLM advisor engaged in multi-turn conversation to collect the users' investment preferences

Stage 2 Advisory Discussion: The user and the LLM advisor discuss how well a candidate stock matches users' preferences. This process is **repeated four times**, once for each stock.





Stage 2: Advisory Discussion

RQ2, 3

Stock Ranking

The user **ranks** the four stocks based **on how well they fit their preferences**

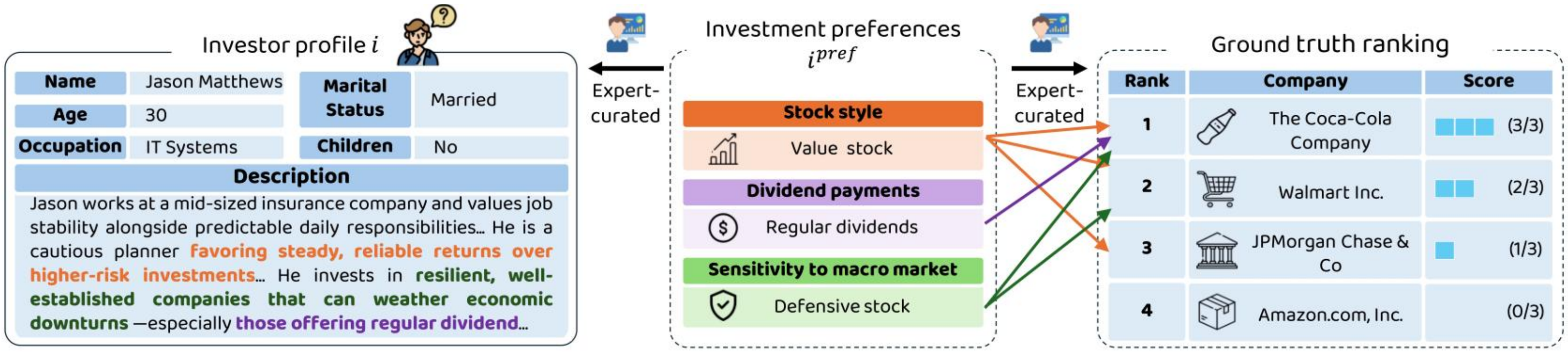
Advisor Assessment

The user rates **their experience** with the advisor on a **7-point Likert scale**

Evaluation: Expert-Designed Gold Standard

- **Problem:** Free-form dialogue with real users introduces high variability, making it difficult to compare performance across different advisor configurations
- **Solution:** Role-play with archetypal investor profiles
  - Users are assigned **expert-designed** profiles and role play during the study
  - Growth-Oriented, Conservative-Income, Risk-Taking
- Each profile includes
  - Expert-curated investment preferences
  - Ground truth stock rankings

This setup provides a **gold standard** for our evaluation



RQ1: Elicitation Accuracy

Elicitation Accuracy: Overlap between ground-truth i^{pref} and elicited preferences i_j^{LLM}

Investor Profile	Expert	LLM-Advisors			
		LLM	+Extr.	+Cons.	Average
Growth-Oriented	0.78	0.76	**0.80**	0.79	$0.78^{\rightarrow 0.0\%}$
Conservative-Income	**0.89**	0.82	0.75	0.87	$0.82^{\downarrow 7.8\%}$
Risk-Taking	**0.89**	0.48	0.60	0.55	$0.53^{\downarrow 40.5\%}$
Average	**0.85**	0.69	0.70	0.73	$0.70^{\downarrow 17.6\%}$

Stage 1 - Comparison of Elicitation Accuracy of an expert vs. different LLM-advisors for each investor profile.
The best advisor is highlighted in bold. Arrows denote percentage increases (↑) or decreases (↓) compared to the expert.

- LLMs advisor matched expert advisors for growth and conservative profiles (~80%)
- Performance dropped sharply for risk-taking investors (↓45%)
- Issues
  - Users misunderstood investment terms
  - LLMs hallucinated or overrode contradictory input

Error Analysis: User Side Error

- These occur when participants **misunderstand financial concepts** or provide **inconsistent information** during preference elicitation.
- 🧠 **Example 1: Concept Confusion**
  A user says they prefer **“non-cyclical stocks”**, which are typically stable across economic cycles (e.g., utilities, consumer staples).
  But later in the conversation, they express interest in the **“consumer discretionary”** sector, which is by definition **cyclical** (sensitive to economic shifts).
- 🔁 *This contradiction introduces noise in the elicited profile and can mislead both the human and LLM advisor.*
- 🧠 **Example 2: Mislabeling Preferences**
  A participant says they like companies “with high future growth potential,” but selects **value stocks** when asked about style preference—indicating confusion between **growth** and **value** investing styles.

Error Analysis: LLM Side Error

- These happen when the **LLM infers or inserts preferences** that were not explicitly stated, often based on prior conversational context.
- 🤖 **Example 1: Preference Hallucination**
  The user explicitly states they're interested in **high-risk, high-reward growth stocks**. *"I'm young and looking for aggressive growth—happy to take some risk."* However, the LLM recommends **conservative, dividend-paying value stocks**, such as large utility companies.
- 🔁 *This mismatch occurred because earlier in the conversation the user said they liked "stable companies" or mentioned "long-term investing," and the LLM inferred a cautious preference.*
- 🤖 **Example 2: Overriding Contradiction**
- When a user hesitates or contradicts themselves ("I think I want growth... or maybe value?"), the LLM might "decide" and finalize a preference for growth—**without explicitly confirming** with the user.

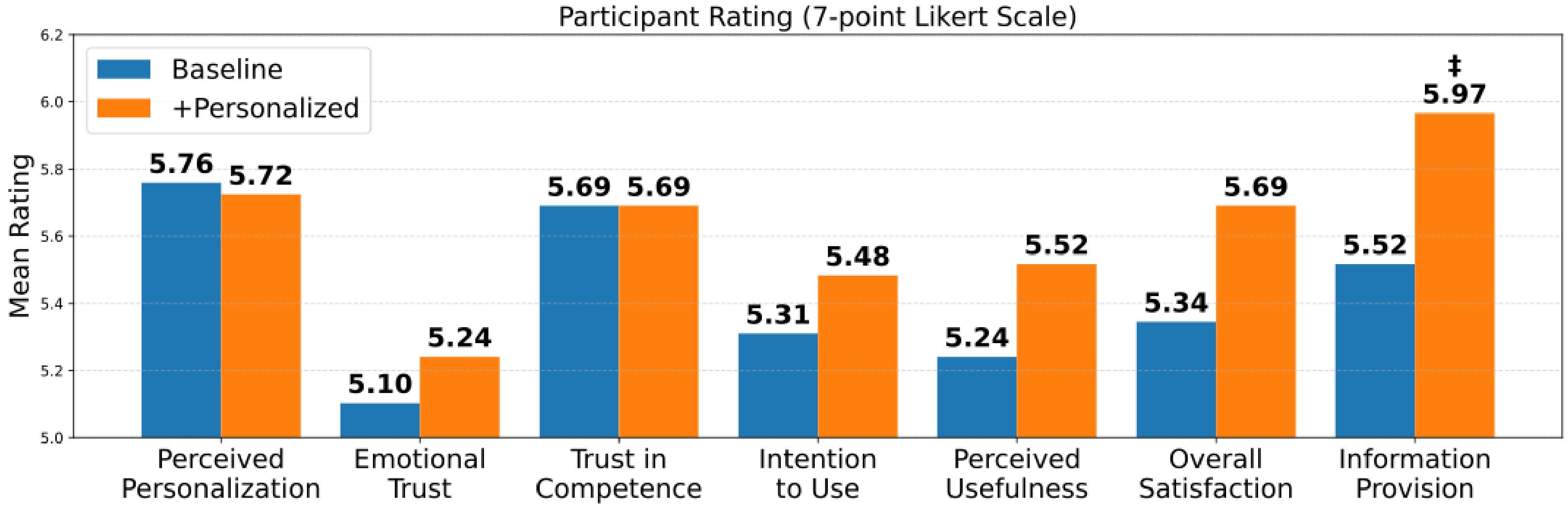
RQ2: Advisor Perception

	(RQ2) Baseline vs. +Personalized					
	All			Successful Elicitation		
Response Dimension	**Baseline**	**+Pers.**	**p**	**Baseline**	**+Pers.**	**p**
Perceived Personalization	5.759	5.724	0.838	5.762	5.905	0.751
Emotional Trust	5.103	5.241	0.446	5.143	5.333	0.537
Trust in Competence	5.690	5.690	0.817	5.810	5.857	0.782
Intention to Use	5.310	5.483	0.505	5.429	5.714	0.166
Perceived Usefulness	5.241	5.517	0.183	5.381	5.810	0.194
Overall Satisfaction	5.345	5.690	0.116	**5.429**	**5.810**	**0.098**†
Information Provision	**5.517**	**5.966**	**0.026**‡	**5.714**	**6.143**	**0.053**†

Participant ratings (7-point Likert scale) for Baseline vs. +Personalized advisors. p-values (Wilcoxon test) show significance for all users and those with successful elicitation (accuracy ≥ 0.5). † = p < 0.1, ‡ = p < 0.05.

- Ratings are mostly similar between Baseline and +Personalized
- Only “Information Provision” showed a significant improvement (‡)
- Even when personalization improved decisions, users couldn’t tell

RQ2: Advisor Perception



- Ratings are mostly similar between Baseline and +Personalized
- Only “Information Provision” showed a significant improvement (‡)
- Even when personalization improved decisions, users couldn’t tell

RQ3: The Effect of Advisor Personality on Decision-Making

Advisor Config		Investor vs. Expert (Spearman's Rho)		
			Preference Elicitation	
Personalization	Personality	All	Successful	Unsuccessful
+Personalized	+Extroverted	0.122	0.243§	-0.286
+Personalized	+Conscientious	0.26	0.365	**-0.025**

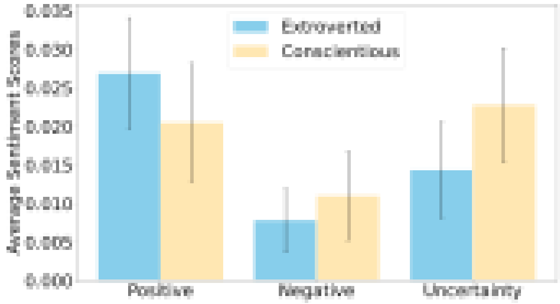
Investor decision-making effectiveness (Spearman's ρ between user and expert rankings). † = p < 0.05 vs. baseline; § = significant difference between successful and unsuccessful elicitation.

- (As seen in RQ2), Poor elicitation can degrade performance
- Conscientious advisor consistently leads to better stock rankings

RQ3: The Effect of Advisor Personality on Decision-Making

	(RQ3) +Conscientious vs. +Extroverted					
	All			Successful Elicitation		
Response Dimension	+Cons.	+Extr.	p	+Cons.	+Extr.	p
Perceived Personalization	5.500	5.500	0.663	5.588	5.706	0.941
Emotional Trust	5.038	5.154	0.600	**4.706**	**5.235**	**0.034**†
Trust in Competence	5.962	6.077	0.558	6.000	6.000	1.000
Intention to Use	**4.885**	**5.462**	**0.005**†	**4.941**	**5.588**	**0.015**†
Perceived Usefulness	5.423	5.538	0.425	5.176	5.118	0.968
Overall Satisfaction	5.269	5.577	0.179	5.118	5.529	0.244
Information Provision	5.692	5.654	0.953	5.588	5.765	0.490

Participant ratings (7-point Likert scale) for
+Extroverted vs. +Conscientious advisors.



Average sentiment scores
by advisor personality

- User preferred the extroverted advisor, despite worse decision performance
  - Emotional Trust, Intention to Use

→Users may trust friendly advisors more, even when their advice is less reliable — a risk in high-stakes domains.