

Article

Not peer-reviewed version

Perceived Cognitive Assistance in LLM-Augmented Retail Trading: Construct Definition and Content Validation

[Dmitrii Gimmelberg](#) * and [Iveta Ludviga](#)

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1772.v1

Keywords: perceived cognitive assistance; retail trading; large language models; scale development; content validity



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Perceived Cognitive Assistance in LLM-Augmented Retail Trading: Construct Definition and Content Validation

Dmitrii Gimmelberg * and Iveta Ludviga

Faculty of Business and Economics, RISEBA University of Applied Sciences, Meza iela 3,
LV-1048 Riga, Latvia

* Correspondence: 22-014729@riseba.lv or dgqw71@gmail.com

Abstract

Large language models (LLMs) are increasingly used by retail traders to interpret information and design complex strategies, yet existing adoption constructs do not capture the decision-time experience of being cognitively scaffolded by an LLM. We define Perceived Cognitive Assistance (PCA) as the trader's felt expansion of cognitive capability at the moment of trading decision when an LLM is available, and we report initial content validation of a PCA item pool. Study 1 specified the PCA content domain using a two-tier qualitative corpus (8 interviews and 44 YouTube narratives on LLM-assisted trading, plus 24 qualitative and mixed-method studies on robo-advice and social trading). Reflexive thematic analysis yielded five facilitative assistance facets and one adjacent risk facet (over-reliance), and these were translated into a 16-item PCA pool. Study 2 used a naïve-judge sort-and-rate task with 48 retail traders to test whether items show definitional correspondence to PCA and definitional distinctiveness from similar constructs: perceived usefulness, perceived ease of use, trust in the LLM, and trading self-efficacy. The resulting 9 item set is ready for subsequent factor-analytic and predictive validation. This study advances our understanding of how large language models shape retail trading behaviour by identifying and empirically grounding Perceived Cognitive Assistance as the decision-time psychological experience through which LLMs cognitively scaffold traders, clarifying how LLM use differs from generic technology adoption, trust, or self-efficacy effects.

Keywords: perceived cognitive assistance; retail trading; large language models; scale development; content validity

1. Introduction

Retail investors make buy and sell decisions in an environment where market information is abundant, uneven in quality, and often communicated as advice rather than neutral data (Miller & Skinner, 2015; Shiller, 2017; Shiller & Pound, 1989). Interpersonal communication, expert commentary, and media attention can reinforce herding and attention-driven trading, especially under uncertainty (Bikhchandani & Sharma, 2001; Hsieh et al., 2020; Shiller, 2017). Large-sample evidence shows persistent behavioural regularities such as excessive trading, attention-based stock selection, and poorer outcomes for the most active traders.

Perceived Cognitive Assistance (PCA) is defined here as the trader's felt expansion of cognitive capability at the moment of decision when a large language model (LLM) is available, with emphasis on how the decision is structured rather than whether outcomes improve. PCA differs from perceived usefulness because usefulness evaluates expected results (for example, performance gains, efficiency gains, or better trading outcomes), whereas PCA evaluates decision-time cognitive scaffolding (for example, a clearer path from an idea to an executable plan, better internal checking, and improved scenario comparison). This distinction matters because LLMs are conversational and can shape the

user's reasoning process in real time; therefore, a trader may feel cognitively enabled even when objective performance does not improve, and conversely may find an LLM "useful" for information retrieval without experiencing decision-time cognitive structuring.

These tendencies are amplified by asymmetric reactions to gains and losses, including the disposition effect and related forms of loss-sensitive selling and holding (Ahn, 2022; Kahneman & Tversky, 1979; Shefrin & Statman, 1985). Cognitive biases in financial judgement appear widespread across economic groups, suggesting that they are not confined to small or unusual investor segments (Ruggeri et al., 2023).

Digital distribution channels in financial technology (FinTech) lower execution frictions and keep investors in continuous, attention-competitive streams (Barber et al., 2021; Miller & Skinner, 2015). Evidence from trading applications and social media settings indicates that attention shocks can increase risk-taking and are associated with weaker holding period returns for attention-induced trades (Eliner & Kobilov, 2023; Warkulat & Pelster, 2024). These features matter because they change not only what information is available, but also how investors experience and process information at the moment of choice (Miller & Skinner, 2015; Shiller, 2017).

Large language models (LLMs) are now entering this environment as a new form of retail-facing decision support (Kong et al., 2024; Lopez-Lira & Tang, 2024; Schlosky & Raskie, 2025; Winder et al., 2025). Recent surveys and applied studies of large language models (LLMs) in finance document rapid diffusion of LLM-based analytical support and outline decision-quality and risk channels relevant to private investors (Li et al., 2023; Lee et al., 2025; Oh et al., 2025). Unlike screeners and many robo-advisors that mainly automate filtering or allocation, LLM-based systems can provide interactive, multi-turn conversational support that elicits user preferences and delivers tailored explanations and guidance in natural language (Chen, 2025; Takayanagi et al., 2025). This interaction can influence framing and perceived controllability, which are central determinants of action in the Theory of Planned Behaviour (Ajzen, 1991, 2011). At the same time, evidence from AI-advice experiments shows that people may follow AI recommendations even when those recommendations conflict with contextual information and their own interests ("Trust and Reliance on AI—An Experimental Study on the Extent and Costs of Overreliance on AI | Request PDF," 2025). Broader work on trust and reliance in automation also shows that users can oscillate between avoidance and over-reliance depending on perceived error, presentation, and expectations (Dietvorst et al., 2015; Glikson & Woolley, 2020; "(PDF) Measurement of Trust in Automation," 2021).

Despite rapid diffusion, empirical research is constrained by a measurement gap. Existing technology-adoption constructs capture important evaluations of tools, but they do not directly measure the decision-time experience that users describe as "it helps me think through this decision right now" (Ali et al., 2025; Chen et al., 2025; Davis, 1989; Venkatesh et al., 2003, 2012). Perceived usefulness focuses on expected results and performance gains, while perceived ease of use focuses on effort in operating the tool (Davis, 1989; Dorobăț & Corbea (Florea), 2025; Mustofa et al., 2025; Venkatesh et al., 2003).

Trust in automation and AI concerns beliefs about system reliability and appropriate reliance (Glikson & Woolley, 2020; Hoff & Bashir, 2015; Jian et al., 2000; Lee & See, 2004). PCA is different: a trader may trust an LLM without feeling cognitively assisted in a specific decision, and vice versa. Trading self-efficacy reflects perceived baseline ability to trade well independent of tools, whereas PCA is conditional on LLM availability at the moment of decision (Ajzen, 1991, 2006). Constructs developed for robo-advisory settings (e.g., delegation, satisfaction with automated allocation) generally assume a more passive, rule-based service (Brenner & Meyll, 2019; D'Acunto et al., 2019). They therefore do not target the interactive, multi-turn cognitive scaffolding in natural language that distinguishes LLM-based decision support (Chen, 2025; Takayanagi et al., 2025). In short, existing measures do not directly target the decision-time experience of expanded cognitive capability that traders describe when using an LLM.

To address this gap, we propose a new construct, Perceived Cognitive Assistance (PCA) (Gimmelberg & Ludviga, 2025). PCA is intentionally process-focused: it captures perceived support

for understanding, judgement, and decision structuring, rather than downstream outcomes such as returns. The purpose of this paper is to provide a measurement foundation for empirical tests of LLM-augmented retail trading behaviour by (i) specifying PCA and its boundaries against neighbouring constructs (usefulness, ease of use, trust, and trading self-efficacy), and (ii) reporting content-validity evidence for a PCA item pool as a gate before factor-analytic testing (Boateng et al., 2018; Colquitt et al., 2019; Hinkin, 1998; Morgado et al., 2017). This sequencing follows scale-development guidance that clear domain specification and content validation should precede statistical tests of factor structure, especially when constructs are proximal and likely to be confused by respondents (Clark & Watson, 1995, 2019; Colquitt et al., 2019). Systematic reviews show that scale-development studies often report avoidable methodological limitations, reinforcing the need to treat content validity as a front-end requirement rather than an optional add-on (Morgado et al., 2017).

The paper makes three contributions. First, it provides a clear definition and boundaries for PCA, grounded in a transparent qualitative coding frame that anchors item content in trader experiences across 76 sources. Second, it delivers a content-validated item pool: seven items meet all classification thresholds, nine are borderline, and none fall into the problematic range—confirming that PCA is perceptibly distinct from neighbouring constructs at the item level. Third, it identifies the PCA–perceived usefulness boundary as the critical discrimination challenge: filler-item accuracy for usefulness was 81.2%, below the 85% threshold, confirming that distinguishing “helps my thinking” from “improves my results” is genuinely difficult. This finding has direct implications for item wording and discriminant validity testing in subsequent studies. Practically, a short PCA score can support governance by helping to monitor when LLM use is linked to increasing strategic complexity without matching guardrails. In applied settings, PCA can also support safer financial decision-making by flagging when perceived decision-time capability rises, so platforms or advisors can trigger additional risk prompts, suitability checks, and ‘human-in-the-loop’ review before users adopt complex or leveraged tactics (Barber & Odean, 2000; Bauer et al., 2009; Lee & See, 2004).

This paper aims to answer the research question: Can Perceived Cognitive Assistance (PCA)—the felt expansion of capability at decision time when using an LLM—be clearly defined, grounded in qualitative evidence, and supported by content validation as distinct from neighbouring constructs, producing a scale candidate ready for psychometric validation?

The paper proceeds in four steps. Section 2 describes the two-study design, covering the qualitative domain specification and item generation (Study 1) and the naïve-judge content-validation procedure (Study 2). Section 3 reports the content-validation results and proposes a provisional nine-item Perceived Cognitive Assistance (PCA) set for subsequent psychometric testing. Section 4 discusses implications, limitations, and the next validation stage. Appendices A–E provide supporting materials for replication and transparency, including the full study instruments and protocols, the PCA macro-code frame and item mapping, the corpus construction and mapping details, the complete item-level validation indices, and the canonical-versus-retained measurement architecture used to position PCA relative to neighbouring constructs.

The two-study design maps directly onto this measurement gap. Study 1 derives PCA inductively from traders’ descriptions of decision-time cognitive experience, rather than deducing items from existing adoption frameworks that were not designed for this purpose (Podsakoff et al., 2016; Hinkin, 1995). Study 2 tests whether the resulting items are recognisable as PCA and distinguishable from the neighbouring constructs identified above, using an independent-rater procedure grounded in the perspectives of retail traders themselves (Colquitt et al., 2019).

2. Materials and Methods

We used a two-study design to develop a measure of PCA, defined as the felt expansion of cognitive capability at the moment of trading decision when a large language model (LLM) is available. Study 1 established the construct domain and generated an initial pool of PCA items using a two-tier qualitative programme and explicit content mapping (DeVellis, 2016; Hinkin, 1995). Study 2 then assessed content validity using an independent-rater (“naïve judge”) procedure designed to

test whether items show definitional correspondence to PCA and definitional distinctiveness from close comparator constructs (Colquitt et al., 2019). This paper reports only construct definition, item generation, and content validation; factor-analytic validation is planned as a subsequent step and is not part of the present Methods section.

Figure 1 summarises the staged scale-development logic used in this paper. Steps 1–3 correspond to Study 1 (qualitative corpus → construct domain → item pool), and Step 4 summarises the Study 2 naïve-judge gate (content validation prior to any factor analysis).

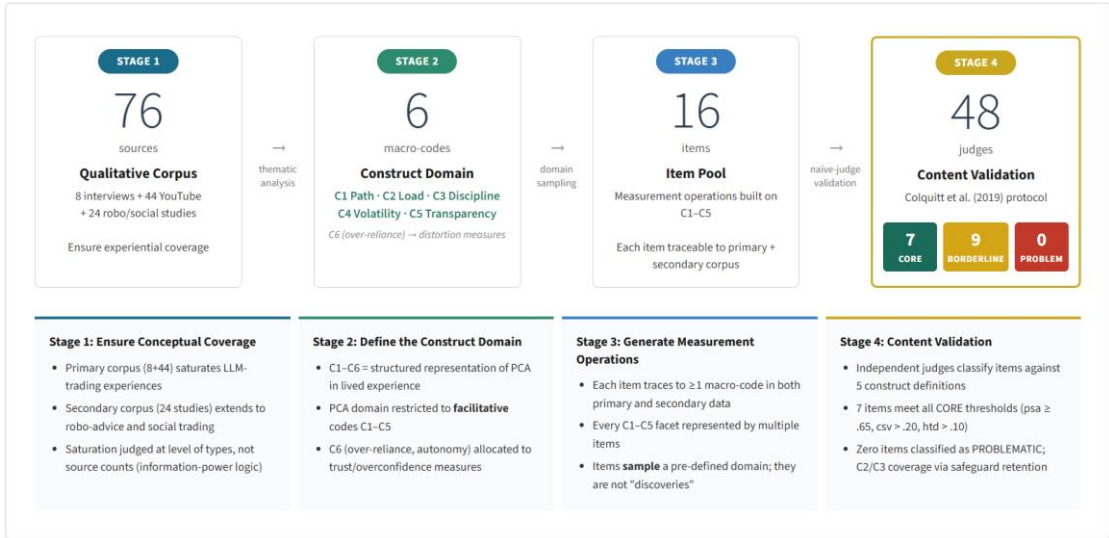


Figure 1. Staged approach to PCA scale development: domain specification and naïve-judge content validation. Tier B = LLM-trading corpus (8 interviews; 44 YouTube narratives); Tier A = legacy corpus (24 qualitative and mixed-method studies); CORE/BORDERLINE/PROBLEMATIC = item-status categories based on a priori correspondence and distinctiveness thresholds.

2.1. Study 1: Construct Definition and Item Generation

PCA is defined as a decision-time experience, so Study 1 prioritised sources that describe cognition during, or close to, concrete trading episodes. Semi-structured interviews provide detailed accounts linked to specific decisions, while public YouTube narratives add naturally occurring descriptions of LLM use in trading workflows created outside a research setting (Braun & Clarke, 2006; Gimmelberg, Głowacka, et al., 2025). The Tier A legacy corpus was added to triangulate the domain and strengthen boundary management by checking whether similar assistance mechanisms appear in adjacent digital-advice settings (Hinkin, 1995; Podsakoff et al., 2016). What distinguishes the LLM context is real-time, multi-turn scaffolding in natural language, which goes beyond rule-based filtering or automated allocation (Chen, 2025; Takayanagi et al., 2025).

2.1.1. Qualitative Data Sources

Tier A (“legacy” advisory corpus) consisted of 24 qualitative and mixed-method studies on robo-advisors, fintech advisory tools, social- and copy-trading platforms, and early work on conversational AI advisors, published between 2015 and 2025. Tier A was used to cross-check themes and to keep construct boundaries clear: it tests whether the PCA content domain reflects recurring user experiences across adjacent technologies rather than quirks of a single dataset (Hinkin, 1995; Podsakoff et al., 2016).

Tier B (LLM-trading corpus) consisted of eight semi-structured interviews with retail investors and a screened set of 44 public YouTube narratives in which traders discussed LLM use in trading and investing (Gimmelberg, Głowacka, et al., 2025). This corpus provided the main descriptions of decision-time cognitive assistance in an LLM context and guided the construct definition and item

wording (Podsakoff et al., 2016). The Tier B YouTube sources are publicly available interviews and lessons that provide rich first-person accounts, but they are not researcher-conducted; we therefore use them for domain specification and triangulate them with our interviews and the Tier A published-study corpus (Braun & Clarke, 2006; Malterud et al., 2016).

The Tier B corpus was assembled using a two-stage purposive sampling strategy reported in full in Gimmelberg et al. (2025). First, 78 English-language YouTube channels covering financial markets and investment topics were selected, comprising 4 mainstream financial media channels (>1 million subscribers; e.g., Bloomberg, CNBC, Yahoo Finance) and 74 smaller independent channels, with the majority based in the United States ($n = 67$). Channel inclusion required regular publication of substantive financial market or investment strategy content; purely promotional or entertainment channels were excluded. Second, all transcripts uploaded during Q2 2023 ($n = 1,617$) and Q2 2024 ($n = 4,513$) were processed through a four-stage computational relevance filter using the AESTIMA tool: (i) embedding via text-embedding-ada-002, (ii) cosine similarity search for LLM-related content (160 sources retained), (iii) topical narrowing to LLMs in asset management (51 sources), and (iv) substantive extraction using 10 Theory of Planned Behaviour (TPB) aligned questions (44 sources retained). Data extraction accuracy was validated by three researchers on a 25% random sample, yielding 76% full accuracy and 95% subject-level accuracy (Gimmelberg et al., 2025). Purposive channel selection is a recognised limitation; however, the downstream computational filtering operated on the complete transcript set from selected channels regardless of view count or popularity, and the final corpus retained both favourable and critical LLM narratives (Patton, 2015; Malterud et al., 2016).

Across both tiers, a source (an interview, YouTube narrative, or published study) was treated as “core” only if it (i) reported first-person user experience (or equivalent qualitative user accounts) or (ii) made a direct, non-redundant contribution to at least one candidate PCA macro-code (Braun & Clarke, 2006; Guest et al., 2006; Hennink et al., 2017).

2.1.2. Analytical Procedure

We used reflexive thematic analysis to derive and refine the content domain of PCA (Braun & Clarke, 2006). The analytic question was narrow and practical: how do retail traders describe AI tools as helping or hindering their thinking at decision time, in ways that plausibly affect the ability to design or execute complex trading tactics (Podsakoff et al., 2016).

For Tier B, we used four experiential themes from the earlier qualitative report as the analytic starting point, and then returned to the same materials with a PCA-specific lens focused on decision-time cognition (Braun & Clarke, 2006; Gimmelberg et al., 2025). The four themes were: (1) movement from overwhelm to a roadmap, (2) cognitive offloading and memory support, (3) analytic scaffolding and stepwise decision support, and (4) displacement of judgement and over-reliance (Gimmelberg et al., 2025). We then conducted a second-cycle coding pass to translate these narrative themes into an explicit content map suitable for item generation and boundary management (Braun & Clarke, 2006; Podsakoff et al., 2016).

We translated the four themes into six macro-codes (C1–C6) to obtain the smallest set of categories that was both (i) granular enough to write non-overlapping items and (ii) broad enough to remain stable across sources. In practice, two of the four themes bundled two recurring but separable assistance mechanisms that require different item families: “overwhelm to a roadmap” split into workflow structuring and path support (C1) versus decision-time navigation of uncertainty and volatility (C4), and “analytic scaffolding and stepwise decision support” split into error-checking and verification (C3) versus learning-oriented explanation and skill-building (C5). The other two themes were already mechanism-specific and were retained as single codes: cognitive offloading and memory support (C2) and displacement of judgement and over-reliance (C6). We did not increase the number of codes further because additional splits produced categories that were either too narrow to be consistently evidenced across sources or too overlapping to support clean construct boundaries and distinct item wording.

Coding was hybrid in the following limited sense: we used inductive labels grounded in investor language (for example, “second brain,” overload, discipline, “following blindly”) while also using deductive tags to track where wording drifted toward neighbouring constructs such as usefulness, ease of use, and trust. (Podsakoff et al., 2016) This step was used to strengthen construct boundaries during item writing rather than to impose an external theory on the content domain (Hinkin, 1995).

To document coverage across the Tier B corpus and support saturation judgments at the code level, we also applied a simple coverage check. Each Tier B source was rated against each macro-code using a three-point scale (0 = absent, 1 = peripheral, 2 = central). The resulting matrix is provided in Appendix B (Table B2).

2.1.3. Saturation Logic

Saturation was judged at two levels. First, the Tier B large language model (LLM) corpus (eight interviews and 44 YouTube narratives - “8+44 corpus”) was treated as saturated for the broader experiential space of LLM-augmented trading, as argued in the prior study that introduced the corpus. Second, for the present PCA scale project, we re-evaluated saturation at the level of the six PCA macro-codes (C1–C6) across the combined Tier A/B corpus, using Malterud et al.’s “information power” logic and focusing on code-level saturation rather than a fixed number of interviews or papers (Guest et al., 2006; Hennink et al., 2017; Malterud et al., 2016; Saunders et al., 2018). The guiding criterion was whether additional sources introduced new types of decision-time cognitive assistance beyond the existing codes (Hennink et al., 2017). Studies were added in conceptually coherent waves (LLM-trading materials, robo-advisor and digital-advice work, social- and copy-trading studies, and early generative-AI advisor experiments). After each wave, we examined whether new material introduced a genuinely new experiential type of perceived cognitive assistance (or perceived control) beyond C1–C6, or whether it only added examples and nuance within the existing code frame. Saturation for PCA was declared once additional waves added only further examples within C1–C6 and no new macro-codes. This claim is intentionally limited: the corpus is saturated for the PCA-relevant experiential space, not for all aspects of robo-advisors or fintech more broadly. Appendix C documents the search strategy, inclusion waves, and the saturation judgement for the 24-study Tier A corpus, and Appendix B provides the saturation memo and the presence–absence coding summary for the “8+44 corpus” (Table B2).

2.1.4. Item-Generation Rules

We generated an intentionally over-complete initial pool of 16 PCA items to allow later trimming while preserving coverage of the content domain (DeVellis, 2016; Hinkin, 1995, 1998). Item generation followed four rules.

First, traceability: each candidate item had to be linked to at least one coded Tier B example and at least one Tier A study that described a closely similar user experience (MacKenzie et al., 2011; Podsakoff et al., 2016). This dual grounding reduces the risk that items reflect local phrasing or niche practices rather than a recurring experiential pattern (Hinkin, 1995).

Second, referent consistency: each item used an explicit “while using an LLM” referent so that the item measures perceived assistance rather than general trading skill (Hinkin, 1998; Podsakoff et al., 2016).

Third, process focus: items were written to describe decision-time cognitive processes (for example, structuring steps, reducing overload, spotting inconsistencies, comparing scenarios) and were edited to remove outcome language (for example, profitability or “better results”), which would shift the content toward perceived usefulness (Clark & Watson, 1995; Hinkin, 1998; Podsakoff et al., 2016). For example, “The LLM helps me structure the steps of my thinking” expresses process-level decision-time support, whereas “Using the LLM improves my trading outcomes” expresses outcome appraisal and would be treated as perceived usefulness (Davis, 1989).

Fourth, coverage constraints: the pool was constructed so that each facilitative macro-code (C1–C5) was represented by multiple items, reducing the risk that early trimming removes an entire facet

of the construct domain (Hinkin, 1995; Clark & Watson, 1995). The full item wording is provided in Appendix A, and the item-to-code mapping is documented in Appendix B. The full construct definitions are provided in Appendix A.

2.2. Study 2: Content Validation

The design, participant selection, and analysis procedures for this study followed the content validation guidelines established by Colquitt et al. (2019) regarding definitional correspondence and distinctiveness.

2.2.1. Design Rationale

Study 2 implemented content validation as a pre-factor-analytic gate. The purpose was to test whether independent raters, applying standardised construct definitions, interpret the PCA items as representing PCA (definitional correspondence) and as more PCA-like than close comparator constructs (definitional distinctiveness). This approach reduces the risk that later statistical modelling produces a “clean” factor structure from items that are conceptually mixed or difficult to distinguish from neighbouring constructs.

Because PCA is intended as a self-report construct, content validation should reflect how typical respondents interpret the items under standardised definitions, rather than relying only on expert judgement. The sort-and-rate task is appropriate because it jointly tests definitional correspondence (sorting) and definitional distinctiveness (overlap ratings) against close comparators (Colquitt et al., 2019).

2.2.2. Participants

Participants were recruited through Prolific (Palan & Schitter, 2018). Judges were required to have recent retail trading experience and prior use of AI chatbots for trading-related queries. Judges who reported professional trading as a full-time occupation were excluded to keep the judge panel aligned with the target population of retail traders.

The final analysis sample comprised 48 valid judges after applying attention-check criteria and a duplicate-handling rule (Peer et al., 2022). The task median duration was 22.2 minutes, and no participants completed the task in under five minutes. Demographic characteristics were taken from Prolific profile data where available.

In this paper, “naïve judge” means naïve to construct formation, not naïve to the trading context. Judges did not take part in defining PCA, developing the coding frame, or writing items. However, domain familiarity was required so that judges could evaluate whether the statements reflect trading-relevant cognitive assistance rather than generic technology attitudes.

2.2.3. Materials

The task used five constructs: Perceived Cognitive Assistance (PCA), perceived usefulness (PU), perceived ease of use (PEOU), trust in the LLM, and trading self-efficacy (TSE). Each construct was presented with a short, plain-language definition adapted to LLM-assisted trading. Definitions remained available throughout the task via an on-screen definitions display. The full construct definitions are provided in Appendix A.

Judges evaluated a stimulus set of 20 statements: 16 PCA candidate items and four “filler” items (one clear exemplar for each comparator construct). The filler items were used as discrimination checks: if judges cannot reliably classify obvious comparator items, interpretation of PCA item performance is not meaningful. Two additional attention-check screens were embedded within the item sequence. The full Study 2 instrument and screen flow are provided in Appendix A.

We set an a priori filler-accuracy target of 0.85 as a calibration heuristic for deliberately unambiguous exemplars; this benchmark is used only to assess task interpretability and judge attention, not as a retention rule for PCA items.

2.2.4. Canonical vs PCA Items and Scale Architecture

Study 2 content validation uses only the five-construct, 20-statement set described above (16 PCA candidate items plus four comparator filler items). For later psychometric validation waves (not reported here), we assembled a broader multi-construct instrument by starting from canonical item universes for Theory of Planned Behaviour, Technology Acceptance Model, financial risk tolerance, and trust in automation, then retaining shorter blocks to balance content coverage and respondent burden; Appendix E (Table E1) documents the canonical-versus-retained mapping and the rationale for inclusion and exclusion decisions (Ajzen, 1991, 2006; Boateng et al., 2018; Colquitt et al., 2019; Davis, 1989; DeVellis, 2016; Hinkin, 1995; Jian et al., 2000; MacKenzie et al., 2011; McGrath et al., 2025; Morgado et al., 2017; Podsakoff et al., 2016).

2.2.5. Procedure

The full instrument, including screen flow, comprehension checks, and exact item sequence, is documented in Appendix A.

After consent and a short eligibility screen, judges read the construct definitions for Perceived Cognitive Assistance (PCA), perceived usefulness (PU), perceived ease of use (PEOU), trust in the LLM, and trading self-efficacy (TSE), and then completed brief comprehension checks to reinforce key construct differences. Incorrect answers triggered immediate corrective feedback and participants were required to select the correct option before proceeding, so the checks functioned as an instructional gate rather than as an exclusion tool. Because Qualtrics stored only the final (correct) responses for these comprehension items in the analysis export, comprehension-check performance was not analysed and no comprehension-based robustness checks were applied. In future implementations, first-attempt comprehension responses (and timestamps) will be retained to permit comprehension-based robustness checks.

Judges then completed the main “sort-and-rate” task (Colquitt et al., 2019). Items were presented one per screen in randomised order, with construct definitions visible throughout the task. For each of the 20 content items, judges (1) selected the single construct the item best represented (including an “Other/none” option), (2) rated how well the item matched the selected construct definition on a 7-point correspondence scale, and (3) rated the extent to which the item could also fit each non-selected construct on a 7-point overlap scale. After completing the sort-and-rate task, judges provided optional open-ended feedback on construct similarity and on any statements they found confusing or ambiguous. Response option order for the classification question was randomised. Two attention-check screens were embedded within the task: the first required selecting PU, and the second required selecting trust in the LLM and providing a high correspondence rating (6 or 7).

2.2.6. Decision Metrics and Thresholds

All decision metrics and thresholds were specified a priori. Items were evaluated on both correspondence and distinctiveness. Appendix A provides the full computation rules and the decision logic.

Correspondence was captured using three indices: proportion of substantive agreement ($psa = n_{PCA}/N$), construct substantive validity ($csv = (n_{PCA} - n_{max_other})/N$, where n_{max_other} is the largest number of classifications into any single non-PCA construct, excluding “Other/none”), and the mean correspondence rating (MCR), defined as the mean 1–7 correspondence rating among judges who classified the item as PCA (Anderson & Gerbing, 1991).

Distinctiveness was assessed using overlap ratings and a distinctiveness score. Overlap ratings were computed, among PCA classifiers only, as the mean overlap with each comparator construct. Because perceived usefulness, perceived ease of use, and trust share the same “LLM-assisted trading” referent as PCA, these were treated as proximal contamination risks and were decisive for retention. Trading self-efficacy was treated as a secondary comparator because it uses a different referent (baseline ability independent of tools); its overlap was reported but did not trigger retention

decisions. Because overlap was asked only for non-selected constructs, overlap means were computed from the four comparator rows available when PCA was selected (perceived usefulness, perceived ease of use, trust, and trading self-efficacy), and no ‘selected-construct overlap’ value was analysed.

Heterotrait distinctiveness (htd_proximal) was computed among PCA classifiers as the mean of the signed differences between MCR and the overlap ratings for PU, PEOU, and trust, divided by $(a - 1) = 6$ for the 7-point scale; higher values indicate stronger distinctiveness from these proximal comparators (Campbell & Fiske, 1959; Henseler et al., 2015).

Items were classified into three categories using the a priori thresholds. A CORE item met all of the following: $psa \geq 0.65$, $MCR \geq 5.25$, $csv > 0.20$, heterotrait distinctiveness (htd_proximal) > 0.10 , and mean overlap ≤ 4.5 for each proximal construct (PU, PEOU, trust). (Colquitt et al., 2019) A PROBLEMATIC item met any hard-failure rule ($psa < 0.50$, $MCR < 4.5$, $csv \leq 0$, $htd_proximal \leq 0$, or mean overlap > 5.0 on any proximal construct). All remaining items were classified as BORDERLINE and earmarked for wording review.

Finally, we applied a coverage safeguard. After applying the thresholds, the retained CORE set was checked against the Study 1 content map. If any facilitative macro-code (C1–C5) had no CORE items, the strongest BORDERLINE item for that code was provisionally retained to preserve domain coverage for later refinement.

2.2.7. Computational Analysis and Verification

All analyses were implemented in Python (version 3.8 or later) using pandas and NumPy. Raw Qualtrics exports were preprocessed and filtered to exclude preview responses, retain completed surveys, and retain only participants who reached the classification task. To support reproducibility, the replication package includes (i) a protocol-to-code traceability matrix mapping each protocol step to the corresponding implementation, (ii) file integrity verification using SHA-256 hashes (file integrity checks) for the raw exports, and (iii) an automated unit-test suite (automated checks) that asserts sample counts, key metric values, and item classifications (Peng, 2011; Sandve et al., 2013; Wilson et al., 2014).

3. Results

This section reports results from Study 1 (construct domain specification and item generation) and Study 2 (naïve-judge content validation prior to factor analysis).

3.1. Study 1: Construct Domain and Item Pool

3.1.1. PCA Macro-Codes (C1–C5) and an Adjacent Risk Code (C6)

The qualitative synthesis yielded a six-code frame (C1–C6) to organise recurring user experiences of LLM-based and adjacent AI decision tools. Codes C1–C5 define the PCA content domain and represent facilitative, assistance-valence experiences: structural and path support (C1), cognitive-load relief (C2), error-checking (C3), navigation of volatility (C4), and learning-oriented assistance (C5). Code C6 captures autonomy and over-reliance and is treated as an adjacent risk pathway rather than PCA content. We therefore retain C6 in the domain map for boundary transparency and triangulation, but we do not operationalise C6 as PCA scale content to avoid mixed-valence measurement.

3.1.2. Saturation Evidence

Saturation was confirmed for the PCA-relevant experiential space. The presence–absence matrix indicates that all six macro-codes appeared across multiple sources in the Tier B corpus. In the Tier A corpus, new codes ceased to emerge after approximately 18–20 studies, with subsequent sources providing only elaboration of the existing C1–C6 patterns. Here, ‘saturation’ is reported for the full

mapping frame (C1–C6), but the PCA item pool and Table 1 deliberately cover only the facilitative PCA domain (C1–C5).

Table 1. PCA item pool and primary macro-code assignments (C1 = structural and path support; C2 = cognitive-load relief and information triage; C3 = error-checking and bias mitigation; C4 = navigation under complex or fast-moving conditions; C5 = learning-oriented assistance, including rationale, exploration, and reflective improvement).

Item ID	Short content descriptor	Primary macro-code
PCA1	Structured path from idea to order	C1
PCA2	Break complex trades into steps	C1
PCA3	Translate view into executable plan	C1
PCA4	Structure multi-leg/conditional orders	C1
PCA5	Process more information without overload	C2
PCA6	Filter to decision-relevant information	C2
PCA7	Integrate sources into coherent picture	C2
PCA8	Spot gaps and inconsistencies in plans	C3
PCA9	Detect misalignment in numbers/dates/assumptions	C3
PCA10	Scrutinise failure points before commitment	C3
PCA11	Compare tactics side by side	C4
PCA12	Think through what-if paths	C4
PCA13	Stay organised when markets move quickly	C4
PCA14	Understand rationale for strategy–view fit	C5
PCA15	Expand range of mental scenario simulation	C5
PCA16	Reflect on decisions during trade review	C5

Note. Table 5 reports the full wording of the retained nine-item PCA set so readers can reuse the measure without consulting appendices. Appendix A still provides the full 20-item stimulus set (16 PCA candidates plus four fillers) and all study instructions/definitions for exact replication. C6 (autonomy and over-reliance) is part of the qualitative mapping frame but is intentionally not represented in the PCA item pool.

3.1.3. Item Pool (16 Candidate Items)

The item generation process yielded an initial pool of 16 candidate items designed to cover the five facilitative macro-codes (C1–C5). Table 1 displays the item pool and domain mapping. Full verbatim wording and screen flow are provided in Appendix A, while item-to-code traceability and Tier A triangulation evidence for the macro-codes is summarised in Appendices B and C. Table 1 summarises the 16-item pool and its primary macro-code assignments.

3.2. Study 2: Content Validation

Study 2 evaluated a reduced 20-statement stimulus set (16 PCA candidates plus four filler items); the full multi-construct survey architecture is intended for later quantitative validation waves.

3.2.1. Data Quality and Calibration Checks

Study 2 retained a final analysis sample of N = 48 valid judges after applying pre-specified screening and data-quality rules. Appendix A documents the full screen flow, including attention checks and comprehension checks designed to ensure that judges understood the distinctions between PCA and the four comparator constructs (Colquitt et al., 2019). The attention-check pass rates were high (AC1 = 96.1%, AC2 = 98.0%, joint pass = 96.1%), indicating strong task engagement (Meade & Craig, 2012). Accuracy on three filler items used for calibration exceeded the 85% target specified in Section 2.2.3 (Perceived Ease of Use (PEOU) = 91.7%, Trust = 89.6%, Trading Self-Efficacy (TSE) = 95.8%), suggesting that judges could apply construct definitions correctly when items were unambiguous (Colquitt et al., 2019). The Perceived Usefulness (PU) filler item achieved 81.2%

accuracy, below the 85% benchmark, suggesting a substantive boundary challenge between PCA and usefulness judgments in this setting. We treat this sub-threshold usefulness filler result as an empirical boundary finding that motivates tighter process-focused wording and prioritised discriminant-validity testing against perceived usefulness in subsequent larger-sample validation.

Table 2. Study 2 data-quality and calibration outcomes (AC = attention check; TSE = Trading Self-Efficacy).

Indicator	Result	Interpretation benchmark
AC1 pass rate	96.1%	High engagement
AC2 pass rate	98.0%	High engagement
Joint AC pass rate	96.1%	High engagement
Filler accuracy: PEOU	91.7%	≥ 85% target met
Filler accuracy: Trust	89.6%	≥ 85% target met
Filler accuracy: TSE	95.8%	≥ 85% target met
Filler accuracy: PU	81.2%	< 85% indicates PCA–PU proximity

The ≥0.85 criterion is used only for filler-item classification accuracy as a calibration check that judges can apply the construct definitions; it is not an item-retention threshold for PCA candidates, which are evaluated using the item-level definitional correspondence and distinctiveness indices reported below. We treat this result as a substantive boundary finding: in LLM-assisted trading contexts, “usefulness” can act as a broad appraisal that partially absorbs decision-time assistance. This strengthens the case for keeping PCA items explicitly decision-time and process-focused and for treating perceived usefulness as the primary discriminant-validity comparator in subsequent validation research.

3.2.2. Item-Level Content Validation Outcomes

Judges evaluated the 16 PCA items using the correspondence and distinctiveness criteria specified in the Methods. Applying the a priori thresholds, seven items were classified as CORE and nine as BORDERLINE (Table 3). No items were classified as PROBLEMATIC. The mean substantive agreement (psa) across all items was 0.674, indicating that on average, approximately two-thirds of judges classified the candidate items as PCA.

Table 3 summarises the item status outcomes, while Appendix A should be consulted for the full computational definitions of the indices (psa, csv, mean correspondence rating, and heterotrait distinctiveness) and the exact threshold values used to assign status labels (Anderson & Gerbing, 1991; Colquitt et al., 2019).

Table 3. Item status outcomes under a priori content-validation decision rules (CORE = item meets all CORE thresholds; BORDERLINE = item meets minimum viability but fails at least one CORE threshold; PROBLEMATIC = item fails minimum viability or triggers the overlap exclusion rule).

Item ID	Primary macro-code	Status
PCA1	C1	CORE
PCA2	C1	BORDERLINE
PCA3	C1	CORE
PCA4	C1	BORDERLINE
PCA5	C2	BORDERLINE
PCA6	C2	BORDERLINE
PCA7	C2	BORDERLINE
PCA8	C3	BORDERLINE
PCA9	C3	BORDERLINE
PCA10	C3	BORDERLINE
PCA11	C4	CORE

PCA12	C4	CORE
PCA13	C4	BORDERLINE
PCA14	C5	CORE
PCA15	C5	CORE
PCA16	C5	CORE

Note. Appendix D provides the full classification-count confusion tables (including the ‘Other/none’ option) for all items and item-level misclassification profiles for the retained items.

3.2.3. Subdimension Coverage and Safeguard Retention

After CORE classification, three macro-codes were covered by at least one CORE item (C1, C4, and C5), while two macro-codes (C2 and C3) had no CORE items. Consistent with the pre-registered coverage safeguard in Appendix A, we therefore retained the strongest-performing BORDERLINE item from each uncovered macro-code to preserve domain coverage in the next stage. Specifically, PCA5 was retained to represent C2 (cognitive-load relief), and PCA8 was retained to represent C3 (error-checking), yielding a nine-item provisional set for subsequent factor-analytic refinement (seven CORE items plus two safeguard items). Specifically, the seven CORE items were PCA1, PCA3, PCA11, PCA12, PCA14, PCA15, and PCA16, and the two safeguard-retained BORDERLINE items were PCA5 (C2) and PCA8 (C3), yielding the nine-item provisional set for subsequent factor-analytic refinement.

Table 4. Macro-code coverage and safeguard retention for the next-stage pool (C1–C5 = macro-codes defining the facilitative PCA domain).

Macro-code	CORE coverage present	Safeguard action (if needed)
C1	Yes	None
C2	No	Retain PCA5 as a safeguard
C3	No	Retain PCA8 as a safeguard
C4	Yes	None
C5	Yes	None

Note. The safeguard rule is specified in Appendix A, and the retained item wordings are provided in Appendix A.

Table 5. Provisional nine-item PCA set proposed for subsequent validation (full wording).

Item ID	Full item wording (verbatim)	Primary macro-code
PCA1	Using an LLM provides a structured path from my initial trading idea to a concrete order.	C1
PCA3	LLM support enables me to translate my market view into a precise, executable trade plan.	C1
PCA5	With LLM help, I can process more information at once without feeling overloaded.	C2 (safeguard)
PCA8	Using an LLM enhances my ability to spot inconsistencies or gaps in my trading plans.	C3 (safeguard)
PCA11	LLM support helps me compare alternative tactics for the same view side by side.	C4
PCA12	LLM support helps me think through what-if scenarios and plan for different market paths.	C4
PCA14	LLM support helps me understand the rationale behind a strategy’s suitability for my market view.	C5
PCA15	LLM support expands the range of trading scenarios I am able to mentally simulate.	C5
PCA16	LLM support facilitates reflection on my decision-making process during trade review.	C5

Note. The nine-item set comprises seven CORE items plus two pre-registered safeguard items retained to preserve coverage of C2 and C3. The complete 16-item candidate pool and filler items remain in Appendix A. Item wording reproduced verbatim from the Qualtrics survey instrument.

3.2.4. Inter-Rater Reliability

Inter-rater agreement for the categorical classification task (Q1) was low when computed on the 16 PCA items alone (Fleiss' κ [kappa] = 0.011) and higher when computed on the full 20-item set including filler items (κ = 0.316). (Fleiss, 1971). In this design, κ is not treated as a reliability coefficient for the scale.

Kappa is chance-corrected agreement and is sensitive to skewed marginal distributions. (Byrt et al., 1993; Feinstein & Cicchetti, 1990). When one category dominates because most stimuli are candidate PCA items and constructs are intentionally proximal, κ can be suppressed even when raw consensus is meaningful. (Byrt et al., 1993; Feinstein & Cicchetti, 1990). We therefore treat κ as a descriptive indicator of boundary difficulty, while item retention follows the pre-registered item-level indices that directly operationalise definitional correspondence and definitional distinctiveness. (Anderson & Gerbing, 1991; Colquitt et al., 2019).

Appendix D reports the full classification distribution (including the "Other/none" option), item-level misclassification profiles, and κ values for both the PCA-item subset and the full stimulus set including fillers, to make the prevalence and boundary-difficulty effects explicit. (Fleiss, 1971; Byrt et al., 1993; Feinstein & Cicchetti, 1990).

3.2.5. Definitional Correspondence, Definitional Distinctiveness, and the Operational PCA Set

Following content-validation guidance, we interpret Study 2 in terms of definitional correspondence (the extent to which items are matched to their intended construct definition) and definitional distinctiveness (the extent to which items match the intended definition more than orbiting construct definitions) (Anderson & Gerbing, 1991; Hinkin & Tracey, 1999; Colquitt et al., 2019). We use the filler items as a construct-level calibration check for definitional correspondence, with an a priori target of $psa \geq 0.85$ for deliberately unambiguous exemplars (Colquitt et al., 2019). Table 6 reports the calibration results and shows that the calibration target was met for perceived ease of use, trust, and trading self-efficacy, but not for perceived usefulness, indicating that usefulness is the closest boundary construct in this setting (Davis, 1989; Colquitt et al., 2019).

At the item-pool level, the mean proportion of substantive agreement across the 16 PCA candidates was $psa = 0.674$, indicating that, on average, approximately two-thirds of judges classified the candidate items as PCA under strict definitions. (Anderson & Gerbing, 1991; Colquitt et al., 2019). Applying the a priori correspondence and distinctiveness decision rules (psa , csv , correspondence ratings, and distinctiveness/overlap criteria), seven items were classified as CORE and nine as BORDERLINE, with no items classified as PROBLEMATIC and no overlap-exclusion triggers observed (Anderson & Gerbing, 1991; Hinkin & Tracey, 1999; Colquitt et al., 2019).

To make the measurement decision explicit for downstream validation, we designate the retained PCA item set as the provisional scale candidate for subsequent psychometric testing (Hinkin, 1998; Colquitt et al., 2019). As reported in Section 3.2.3, the retained set comprises nine items (PCA-9): seven CORE items plus two safeguard items retained to preserve macro-code coverage for C2 and C3 (Colquitt et al., 2019). Table 7 lists the PCA-9 items and their retention basis (CORE versus safeguard).

Table 6. Construct-level calibration of definitional correspondence using filler items (PCA = Perceived Cognitive Assistance; PU = perceived usefulness; PEOU = perceived ease of use; TSE = Trading Self-Efficacy; psa = proportion of substantive agreement).

Construct (filler)	Definitional correspondence, psa	Misclassification n (1 – psa)	psa ≥ 0.85 target met?	Interpretation
PEOU	0.917	0.083	Yes	Calibration target met.
Trust	0.896	0.104	Yes	Calibration target met.
TSE	0.958	0.042	Yes	Calibration target met.
PU	0.812	0.188	No	Below-target calibration indicates proximity to PCA under naïve-judge interpretation.

Note. The 0.85 target is a calibration heuristic for deliberately unambiguous filler exemplars and is not applied to PCA candidates, which are expected to be closer to boundary constructs by design. (Colquitt et al., 2019).

Table 7. Operational PCA (PCA-9) provisional item set retained for next-stage validation (CORE = meets all CORE thresholds; safeguard = best-performing BORDERLINE retained to preserve macro-code coverage).

Item ID	Macro-code	Retention basis	Role in PCA set
PCA1	C1	CORE	Core coverage.
PCA3	C1	CORE	Core coverage.
PCA11	C4	CORE	Core coverage.
PCA12	C4	CORE	Core coverage.
PCA14	C5	CORE	Core coverage.
PCA15	C5	CORE	Core coverage.
PCA16	C5	CORE	Core coverage.
PCA5	C2	Safeguard	Preserves C2 (cognitive-load relief) content.
PCA8	C3	Safeguard	Preserves C3 (error-checking) content.

Note. PCA-7 set (CORE-only) can be used as a sensitivity specification in later waves by excluding the two safeguard items (PCA5, PCA8). (Colquitt et al., 2019).

4. Discussion

The central measurement implication is that Perceived Cognitive Assistance (PCA) can be defined as a process-focused construct that is related to, but not reducible to, perceived usefulness and perceived ease of use (Davis, 1989; Podsakoff et al., 2016). The construct-level calibration results (Table 6) show that perceived usefulness (PU; a belief that using the system improves task performance) is the closest boundary construct in this setting, because the usefulness filler item did not meet the a priori correspondence target ($psa = 0.812 < 0.85$), unlike the filler items for perceived ease of use, trust, and trading self-efficacy (Colquitt et al., 2019). This pattern implies that naïve judges can treat “usefulness” as an umbrella appraisal that absorbs multiple forms of help, including decision-time cognitive support, unless definitions and item wording force an explicitly process-level interpretation (Davis, 1989; Podsakoff et al., 2016). At the item level, correspondence and distinctiveness criteria yielded seven CORE items and nine BORDERLINE items (Table 3), with no PROBLEMATIC items and no overlap-exclusion triggers, supporting the viability of PCA as a distinct

construct while identifying perceived usefulness as the primary discriminant challenge (Anderson & Gerbing, 1991; Colquitt et al., 2019; Hinkin & Tracey, 1999). This CORE–BORDERLINE distribution suggests that PCA already has a recognisable core that judges interpret as decision-time cognitive scaffolding, but that its perimeter remains partially entangled with perceived usefulness language because some phrasings still invite an outcome interpretation. For subsequent validation stages, we therefore treat perceived usefulness (and secondarily perceived ease of use) as primary discriminant-validity comparators rather than peripheral controls, because PCA’s substantive value depends on demonstrating measurement signal beyond general technology appraisals (Davis, 1989; Hinkin, 1998; Colquitt et al., 2019). Finally, to make the operationalization transparent, we carry forward PCA as a nine-item provisional set (Tables 5 and 7), consisting of the seven CORE items plus two safeguard items retained to preserve macro-code coverage for cognitive-load relief (C2) and error-checking (C3) (Colquitt et al., 2019).

Two retained items use plan-oriented phrasing that can be misread as performance-improving unless interpreted as decision-time structuring. In this scale, “structured path” denotes sequencing and organisation of decision steps from idea to order, and “executable trade plan” denotes translating an intent into a specified plan that the trader can implement, without claiming that the plan improves returns, accuracy, or performance (Davis, 1989; Venkatesh et al., 2003). This wording choice is consistent with the instrument rule that PCA items describe how the decision is structured at the moment of choice, while usefulness items describe evaluative outcome appraisal (Davis, 1989; Venkatesh et al., 2003).

The content-validation outcomes also provide early guidance about which facets are easiest to communicate as “cognitive assistance” under strict definitional tests (Colquitt et al., 2019). Items that emphasised structure from idea to action (C1), comparison and scenario navigation under time pressure (C4), and learning-oriented understanding and reflection (C5) were more likely to meet CORE criteria (see Table 4), which may reflect that these experiences are more distinctive from trust and ease-of-use judgements when phrased as cognitive-process support (Colquitt et al., 2019; Davis, 1989). By contrast, the cognitive-load and information-triage facet (C2) and the error-checking and inconsistency-detection facet (C3) did not yield CORE items, and representation of these facets therefore relies on the pre-specified coverage safeguard (Hinkin, 1995; Colquitt et al., 2019; see Appendix A for decision rules and Appendix D for full item-level indices). This outcome does not imply that C2 and C3 are outside the PCA domain, because Appendix B documents strong qualitative support for both facets in the Tier B corpus and Appendix C provides triangulating evidence from the Tier A corpus, but it indicates that the current phrasings may invite overlap with perceived ease of use, trust, or self-efficacy unless wording is sharpened to emphasise “how my thinking changes” rather than “the tool works well” (Podsakoff et al., 2016; Davis, 1989). For transparency and future refinement, Appendix A provides the definitive record of the construct definitions and item wording that produced these outcomes, and Appendix D can be used to audit which distinctiveness conditions each borderline item failed under the pre-registered thresholds (Colquitt et al., 2019).

A second implication concerns scale architecture and the sequence of validation evidence (Hinkin, 1995). Content validation supports the claim that a subset of items corresponds to the intended definition and is not dominated by a single competing construct, but it cannot establish dimensionality, reliability, or predictive validity, which require larger-sample psychometric testing (DeVellis, 2017; Hinkin, 1995). The present evidence therefore supports treating the retained item set as a provisional instrument that is ready for factor-analytic refinement and validation in an independent sample, rather than as a final scale (DeVellis, 2017; Colquitt et al., 2019). Given that Table 2 and Appendix B frame PCA as a five-facet content domain, later model comparisons should be open to both a unidimensional representation (a general PCA factor) and a correlated-facets representation, with item performance guiding whether the construct behaves as a single latent tendency or as distinguishable components in practice (Hinkin, 1995; DeVellis, 2017). The construct-definition discipline applied here, including explicit exclusion of the autonomy and over-reliance risk

pathway (C6) from the PCA domain, should help prevent post hoc drift when later statistical models are estimated (Podsakoff et al., 2016; Lee & See, 2004).

The results also have practical implications for research on large language model (LLM)–augmented trading and for tool governance (Gimmelberg & Ludviga, 2025). PCA provides a way to measure the user’s perceived “cognitive lift” during complex trading decisions without collapsing that experience into simple satisfaction or usefulness ratings, which is important when behavioural change and strategic complexity are the target outcomes rather than mere adoption (Gimmelberg & Ludviga, 2025; Davis, 1989). In applied settings, the scale can be used as a monitoring indicator of perceived assistance intensity across tasks, strategies, or market regimes, with Appendix A providing a complete, reproducible instrument that can be fielded as written (Colquitt et al., 2019). At the same time, the deliberate separation between PCA (C1–C5) and autonomy/over-reliance risk (C6) implies that practitioners should not use high PCA scores as evidence of safe reliance, because assistance can co-exist with responsibility drift (Lee & See, 2004). This is why Appendix B retains C6 in the domain map even though it is not operationalised as PCA content, and why later work should pair PCA measurement with explicit over-reliance or delegation measures when governance and safety are central outcomes (Hoff & Bashir, 2015; Lee & See, 2004; Podsakoff et al., 2016).

5. Conclusions

This paper addresses the research question of whether Perceived Cognitive Assistance (PCA)—the felt expansion of cognitive capability at the moment of trading decision when a large language model (LLM) is available—can be clearly defined, grounded in qualitative evidence, and supported by content validation as distinct from neighbouring constructs, producing a scale candidate ready for later psychometric validation. We answer this question in the affirmative by (i) specifying PCA as a decision-time belief focused on process-level cognitive scaffolding, and (ii) demonstrating initial content validity for a provisional PCA item pool prior to any factor analysis. Across a two-tier qualitative programme (76 sources comprising interviews, YouTube narratives, and legacy fintech studies) and a preregistered naïve-judge procedure ($N = 48$), the results support a retained nine-item set (seven CORE items plus two safeguards) and no items classified as problematic (Braun & Clarke, 2006; Malterud et al., 2016; Colquitt et al., 2019). These items are consistently interpreted as cognitive scaffolding rather than usefulness, ease of use, trust, or baseline trading skill (Podsakoff et al., 2016; Colquitt et al., 2019). The practical implication is that PCA can be measured as a distinct decision-time belief, providing a defensible input to subsequent psychometric validation and to governance-focused applications.

Several limitations follow from the scope of content validation and from the chosen design (Colquitt et al., 2019). First, Study 2 provides evidence only for definitional correspondence and distinctiveness; dimensionality, reliability, measurement invariance, and predictive validity require larger-sample psychometric testing in planned validation waves (Hinkin, 1995; DeVellis, 2017). Second, the naïve-judge method depends on the clarity of the construct definitions and judges’ adherence to them; Appendix A is therefore essential for interpretation, and replications should keep definitions stable when testing alternative wordings (Anderson & Gerbing, 1991; Colquitt et al., 2019). Because first-attempt comprehension-check responses were not retained in the analysis export, comprehension-check performance could not be analysed and no comprehension-based robustness checks could be applied; future implementations will retain first-attempt responses (and timestamps) to enable such checks. Third, inter-rater agreement on the PCA items was low by design (Fleiss’ $\kappa = 0.011$ for PCA items alone; $\kappa = 0.316$ when filler items are included) because judges classified items against proximal comparator constructs; κ should be interpreted as a boundary-difficulty diagnostic rather than as an index of scale reliability (Anderson & Gerbing, 1991; Colquitt et al., 2019). Fourth, the judge sample was recruited from an online panel with OECD residence and prior LLM use for trading, and may not represent retail traders using broker-integrated tools, non-English interfaces, or populations with lower technology exposure (Hinkin, 1995). Fifth, while the qualitative corpus was treated as saturated for the PCA-relevant experiential space, it draws primarily on robo-advice, social

trading, and early LLM-trading contexts; as LLM tools evolve and traders gain more experience, the assistance themes captured by C1–C6 may require updating (Hennink et al., 2017; Malterud et al., 2016). Finally, the PCA–perceived usefulness boundary remained the hardest discrimination test (filler accuracy 81.2%, below the 85% threshold), implying that subsequent waves should continue to tighten process-focused wording and treat perceived usefulness as a primary discriminant comparator (Davis, 1989; Podsakoff et al., 2016).

Table 8. Appendix structure (LLM = large language model; PCA = Perceived Cognitive Assistance; PU = perceived usefulness; PEOU = perceived ease of use; TSE = trading self-efficacy).

Appendix	What it contains	Role in research
Appendix A	Study 2 instrument pack: final PCA item wording (PCA1–PCA16) plus filler-item wordings; construct definitions shown to judges (PCA, PU, PEOU, Trust, TSE); screen flow; comprehension checks; attention checks; rating scales; a priori decision rules	Study 2 materials (audit trail for what judges saw and how decisions were made)
Appendix B	Qualitative coding frame support: C1–C6 definitions; presence–absence matrix; item-by-code mapping Tier A corpus evidence table (24 studies) + selection notes (including wave logic if you keep it)	Study 1 evidence for domain specification and item traceability
Appendix C	Additional references: (Back et al., 2023; Belanche et al., 2025; Bhatia et al., 2020, 2022; Brenner & Meyll, 2019; Castillo et al., 2021; Chandani et al., 2021; Cheng et al., 2019; Costa & Henshaw, 2025; D’Acunto et al., 2019; Dietvorst et al., 2015; Gimmelberg, Glowacka, et al., 2025; Hidajat et al., 2024; Komatireddy et al., 2024; Nashold, 2020; Northey et al., 2022; Nourallah et al., 2022; Prah & Van Swol, 2017; Senteio & Hughes, 2024; Skiera, 2021; “Sophia Sophia Tell Me More, Which Is the Most Risk-Free Plan of All?,” 2022; Verma et al., 2025; Yi et al., 2023; Zhu et al., 2023)	Study 1 triangulation and corpus construction
Appendix D	Post-experiment outputs without raw data/code: full item-level content-validity indices and thresholds applied; any expanded versions of Tables 3–5; optional extra diagnostics (for example, full confusion table or kappa detail). Additional references: (Landis & Koch, 1977)	Extended Study 2 results (supplementary tables)
Appendix E	Canonical vs PCA items and scale architecture (Table E1): canonical “item universe” sources and retained blocks; rationale for inclusion/exclusion. Additional references: (Fornell & Larcker, 1981; Grable & Lytton, 1999; Lewis et al., 2013; Venkatesh & Davis, 2000)	Measurement architecture documentation (belongs conceptually to Methods)

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Appendices A–E are provided as supplementary files. For submission, we kept these appendices as separate, standalone documents to improve readability and reviewer convenience (i.e., to avoid an overly long main manuscript and to keep replication materials easy to navigate).

Author Contributions: Conceptualization, D.G. and I.L.; methodology, D.G. and I.L.; formal analysis, D.G.; investigation, D.G.; data curation, D.G.; writing—original draft preparation, D.G.; writing—review and editing, D.G. and I.L.; visualization, D.G.; supervision, I.L.; project administration, D.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were not required for this study in accordance with local legislation and institutional requirements.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: All study materials and analysis code are available from the corresponding author upon reasonable request. The package includes the Qualtrics survey archive, the full Python analysis pipeline, automated verification tests, and a protocol-to-code traceability matrix. Participant-level data are not publicly available because they are subject to participant consent and platform data-sharing terms; access may be granted to qualified researchers for academic, non-commercial use, subject to appropriate safeguards (for example, a data-use agreement).

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

References

- Ahn, Y. (2022). The anatomy of the disposition effect: Which factors are most important? *Finance Research Letters*, 44, 102040. <https://doi.org/10.1016/j.frl.2021.102040>
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, Theories of Cognitive Self-Regulation*, 50(2), 179–211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Ajzen, I. (2006). *Constructing a TpB Questionnaire: Conceptual and Methodological Considerations*. <https://www.semanticscholar.org/paper/Constructing-a-TpB-Questionnaire%3A-Conceptual-and-Ajzen/0574b20bd58130dd5a961f1a2db10fd1fcbae95d>
- Ajzen, I. (2011). The theory of planned behaviour: Reactions and reflections. *Psychology & Health*, 26, 1113–1127. <https://doi.org/10.1080/08870446.2011.613995>
- Ali, I., Warraich, N. F., & Butt, K. (2025). Acceptance and use of artificial intelligence and AI-based applications in education: A meta-analysis and future direction. *Information Development*, 41(3), 859–874. <https://doi.org/10.1177/02666669241257206>
- Anderson, J. C., & Gerbing, D. W. (1991). Predicting the performance of measures in a confirmatory factor analysis with a pretest assessment of their substantive validities. *Journal of Applied Psychology*, 76(5), 732–740. <https://doi.org/10.1037/0021-9010.76.5.732>
- Back, C., Morana, S., & Spann, M. (2023). When do robo-advisors make us better investors? The impact of social design elements on investor behavior. *Journal of Behavioral and Experimental Economics*, 103, 101984. <https://doi.org/10.1016/j.socec.2023.101984>
- Barber, B. M., Huang, X., Odean, T., & Schwarz, C. (2021). *Attention Induced Trading and Returns: Evidence from Robinhood Users* (SSRN Scholarly Paper No. 3715077). Social Science Research Network. <https://doi.org/10.2139/ssrn.3715077>
- Barber, B. M., & Odean, T. (2000). Trading Is Hazardous to Your Wealth: The Common Stock Investment Performance of Individual Investors. *The Journal of Finance*, 55(2), 773–806. <https://doi.org/10.1111/0022-1082.00226>
- Bauer, R., Cosemans, M., & Eichholtz, P. (2009). Option trading and individual investor performance. *Journal of Banking & Finance*, 33(4), 731–746. <https://doi.org/10.1016/j.bankfin.2008.11.005>

- Belanche, D., Casaló, L. V., Flavián, M., & Loureiro, S. M. C. (2025). Benefit versus risk: A behavioral model for using robo-advisors. *The Service Industries Journal*, 45(1), 132–159. <https://doi.org/10.1080/02642069.2023.2176485>
- Bhatia, A., Chandani, A., & Chhateja, J. (2020). Robo advisory and its potential in addressing the behavioral biases of investors—A qualitative study in Indian context. *Journal of Behavioral and Experimental Finance*, 25, 100281. <https://doi.org/10.1016/j.jbef.2020.100281>
- Bhatia, A., Chandani, A., Divekar, R., Mehta, M., & Vijay, N. (2022). Digital innovation in wealth management landscape: The moderating role of robo advisors in behavioural biases and investment decision-making. *International Journal of Innovation Science*, 14(3/4), 693–712. <https://doi.org/10.1108/IJIS-10-2020-0245>
- Bikhchandani, S., & Sharma, S. (2001). Herd Behavior in Financial Markets. *IMF Staff Papers*, 47, 1–1. <https://doi.org/10.2307/3867650>
- Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. *Frontiers in Public Health*, 6. <https://doi.org/10.3389/fpubh.2018.00149>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Brenner, L., & Meyll, T. (2019). *Robo-Advisors: A Substitute for Human Financial Advice?* (SSRN Scholarly Paper No. 3414200). Social Science Research Network. <https://doi.org/10.2139/ssrn.3414200>
- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423–429. [https://doi.org/10.1016/0895-4356\(93\)90018-v](https://doi.org/10.1016/0895-4356(93)90018-v)
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81–105. <https://doi.org/10.1037/h0046016>
- Castillo, D., Canhoto, A. I., & Said, E. (2021). The dark side of AI-powered service interactions: Exploring the process of co-destruction from the customer perspective. *The Service Industries Journal*, 41(13–14), 900–925. <https://doi.org/10.1080/02642069.2020.1787993>
- Chandani, A., Sriharshitha S., Bhatia, A., Atiq, R., & Mehta, M. (2021). Robo-Advisory Services in India: A Study to Analyse Awareness and Perception of Millennials. *Int. J. Cloud Appl. Comput.*, 11(4), 152–173. <https://doi.org/10.4018/IJCAC.2021100109>
- Chen, J., Liu, Y., Liu, P., Zhao, Y., Zuo, Y., & Duan, H. (2025). Adoption of Large Language Model AI Tools in Everyday Tasks: Multisite Cross-Sectional Qualitative Study of Chinese Hospital Administrators. *Journal of Medical Internet Research*, 27(1), e70789. <https://doi.org/10.2196/70789>
- Chen, Z. (2025). Revolutionizing finance with conversational AI: A focus on ChatGPT implementation and challenges. *Humanities and Social Sciences Communications*, 12(1), 388. <https://doi.org/10.1057/s41599-025-04725-y>
- Cheng, X., Guo, F., Chen, J., Li, K., Zhang, Y., Gao, P., Cheng, X., Guo, F., Chen, J., Li, K., Zhang, Y., & Gao, P. (2019). Exploring the Trust Influencing Mechanism of Robo-Advisor Service: A Mixed Method Approach. *Sustainability*, 11(18). <https://doi.org/10.3390/su11184917>
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. <https://doi.org/10.1037/1040-3590.7.3.309>
- Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. <https://doi.org/10.1037/pas0000626>
- Colquitt, J. A., Sabey, T. B., Rodell, J. B., & Hill, E. T. (2019). Content validation guidelines: Evaluation criteria for definitional correspondence and definitional distinctiveness. *Journal of Applied Psychology*, 104(10), 1243–1265. <https://doi.org/10.1037/apl0000406>
- Costa, & Henshaw. (2025). *Advice Pays in Peace of Mind and Time: Vanguard Survey Reveals Hidden Value of Financial Advice*. <https://corporate.vanguard.com/content/dam/corp/research/pdf/quantifying-the-investors-view-on-the-value-of-human-and-robo-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>
- Davis, F. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, 13(3), 319. <https://doi.org/10.2307/249008>
- DeVellis, R. F. (2016). *Scale Development: Theory and Applications* (Fourth edition). SAGE Publications. <https://books.google.lv/books?id=48ACCwAAQBAJ>

- 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>
- Dorobăț, I., & Corbea (Florea), A. M. I. (2025). Assessing ChatGPT Adoption in Higher Education: An Empirical Analysis. *Electronics*, 14(23). <https://doi.org/10.3390/electronics14234739>
- Eliner, L., & Kobilov, B. (n.d.). *To the Moon or Bust: Do Retail Investors Profit From Social Media-Induced Trading?* Retrieved February 11, 2026, from <https://www.semanticscholar.org/paper/To-the-Moon-or-Bust%3A-Do-Retail-Investors-Pro%EF%AC%81t-From-Eliner-Kobilov/df50bcf89751137a5204432081a87cd25c77aa1e>
- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543–549. [https://doi.org/10.1016/0895-4356\(90\)90158-1](https://doi.org/10.1016/0895-4356(90)90158-1)
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. <https://doi.org/10.1037/h0031619>
- Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, 18(1), 39–50. <https://doi.org/10.2307/3151312>
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. arXiv:2312.10997. <https://doi.org/10.48550/arXiv.2312.10997>
- Gimmelberg, D., Glowacka, M., Belinskiy, A., Korotkii, S., Artamov, V., & Ludviga, I. (2025). Bridging Human Expertise and AI: Evaluating the Role of Large Language Models in Retail Investors' Decision-Making. *International Journal of Finance & Banking Studies* (2147-4486), 14(1), 20–29. <https://doi.org/10.20525/ijfbs.v14i1.3910>
- Gimmelberg, D., & Ludviga, I. (2025). Strategic Complexity and Behavioral Distortion: Retail Investing Under Large Language Model Augmentation. *International Journal of Financial Studies*, 13(4). <https://doi.org/10.3390/ijfs13040210>
- Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, 14(2), 627–660. <https://doi.org/10.5465/annals.2018.0057>
- Grable, J., & Lytton, R. H. (1999). Financial risk tolerance revisited: The development of a risk assessment instrument☆. *Financial Services Review*, 8(3), 163–181. [https://doi.org/10.1016/S1057-0810\(99\)00041-4](https://doi.org/10.1016/S1057-0810(99)00041-4)
- Guest, G., Bunce, A., & Johnson, L. (2006). How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods*, 18(1), 59–82. <https://doi.org/10.1177/1525822X05279903>
- Hennink, M., Kaiser, B., & Marconi, V. C. (2017). Code Saturation Versus Meaning Saturation: How Many Interviews Are Enough? *Qualitative Health Research*, 27(4), 591–608. <https://doi.org/10.1177/1049732316665344>
- Henseler, J., Ringle, C., & Sarstedt, M. (2015). A New Criterion for Assessing Discriminant Validity in Variance-based Structural Equation Modeling. *Journal of the Academy of Marketing Science*, 43(1), 115–135. <https://doi.org/10.1007/s11747-014-0403-8>
- Hidajat, T., Hamdani, M., Putri, R. K., & Ramadhan, A. M. (2024). Behavioral Biases and Trust in Social Trading: A Mixed-Method Approach. *Jurnal Manajemen Indonesia*, 24(2), 214–226. <https://doi.org/10.25124/jmi.v24i2.7547>
- Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, 21(5), 967–988. [https://doi.org/10.1016/0149-2063\(95\)90050-0](https://doi.org/10.1016/0149-2063(95)90050-0)
- Hinkin, T. R. (1998). A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires. *Organizational Research Methods*, 1(1), 104–121. <https://doi.org/10.1177/109442819800100106>
- Hinkin, T. R., & Tracey, J. B. (1999). An Analysis of Variance Approach to Content Validation. *Organizational Research Methods*, 2(2), 175–186. <https://doi.org/10.1177/109442819922004>
- Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57(3), 407–434. <https://doi.org/10.1177/0018720814547570>
- Hsieh, S.-F., Chan, C.-Y., & Wang, M.-C. (2020). Retail investor attention and herding behavior. *Journal of Empirical Finance*, 59, 109–132. <https://doi.org/10.1016/j.jempfin.2020.09.005>
- Jian, J.-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4(1), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04
- Kahneman, D., & Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>

- Komatireddy, K., Mangeshikar, S., & Gada, T. (2024). Augmenting Trust in Robo Advisor Experiences Through Thoughtful UX Design. *FMDB Transactions on Sustainable Computing Systems*, 2(2), 54–63. <https://doi.org/10.69888/FTSCS.2024.000197>
- Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q., & Zohren, S. (2024). Large Language Models for Financial and Investment Management: Models, Opportunities, and Challenges. *The Journal of Portfolio Management*, 51(2), 211–231. <https://doi.org/10.3905/jpm.2024.1.646>
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159–174. <https://doi.org/10.2307/2529310>
- Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- Lewis, J. R., Utesch, B. S., & Maher, D. E. (2013). UMUX-LITE: When there's no time for the SUS. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '13*, 2099–2102. <https://doi.org/10.1145/2470654.2481287>
- Lee, J., Stevens, N., Han, S. C., & Song, M. (2025). Large language models in finance (FinLLMs). *Neural Computing and Applications*, 37, 24853–24867. <https://doi.org/10.1007/s00521-024-10495-6>
- Li, Y., Wang, S., Ding, H., & Chen, H. (2023). Large language models in finance: A survey. arXiv:2311.10723. <https://doi.org/10.48550/arXiv.2311.10723>
- Lopez-Lira, A., & Tang, Y. (2024). Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models (arXiv:2304.07619). arXiv. <https://doi.org/10.48550/arXiv.2304.07619>
- MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS Quarterly: Management Information Systems*, 35(2), 293–334. <https://doi.org/10.2307/23044045>
- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample Size in Qualitative Interview Studies: Guided by Information Power. *Qualitative Health Research*, 26(13), 1753–1760. <https://doi.org/10.1177/1049732315617444>
- McGrath, M. J., Lack, O., Tisch, J., & Duenser, A. (2025). Measuring trust in artificial intelligence: Validation of an established scale and its short form. *Frontiers in Artificial Intelligence*, 8, 1582880. <https://doi.org/10.3389/frai.2025.1582880>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17(3), 437–455. <https://doi.org/10.1037/a0028085>
- Miller, G. S., & Skinner, D. J. (2015). The Evolving Disclosure Landscape: How Changes in Technology, the Media, and Capital Markets Are Affecting Disclosure. *Journal of Accounting Research*, 53(2), 221–239. <https://doi.org/10.1111/1475-679X.12075>
- Morgado, F. F. R., Meireles, J. F. F., Neves, C. M., Amaral, A. C. S., & Ferreira, M. E. C. (2017). Scale development: Ten main limitations and recommendations to improve future research practices. *Psicologia: Reflexão e Crítica*, 30(1), 3. <https://doi.org/10.1186/s41155-016-0057-1>
- Mustofa, R., Kuncoro, T., Atmono, D., Hermawan, H., & Sukirman. (2025). Extending the Technology Acceptance Model: The Role of Subjective Norms, Ethics, and Trust in AI Tool Adoption Among Students. *Computers and Education: Artificial Intelligence*, 8, 100379. <https://doi.org/10.1016/j.caeai.2025.100379>
- Nashold. (2020). Trust in Consumer Adoption of Artificial Intelligence-Driven Virtual Finance Assistants: A Technology Acceptance Model Perspective - ProQuest. <https://www.proquest.com/docview/2385705772?pq-origsite=gscholar>
- Northey, G., Hunter, V., Mulcahy, R., Choong, K., & Mehmet, M. (2022). Man vs machine: How artificial intelligence in banking influences consumer belief in financial advice. *International Journal of Bank Marketing*, 40(6), 1182–1199. <https://doi.org/10.1108/IJBM-09-2021-0439>
- Nourallah, M., Öhman, P., & Amin, M. (2022). No trust, no use: How young retail investors build initial trust in financial robo-advisors. *Journal of Financial Reporting and Accounting*, 21(1), 60–82. <https://doi.org/10.1108/JFRA-12-2021-0451>
- Oh, D., Kim, T., Jang, J., & Park, S.-H. (2025). Democratizing alpha: LLM-driven portfolio construction for retail investors using public financial media. In *Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25)* (pp. 326–334). <https://doi.org/10.1145/3768292.3770376>

- Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice* (4th ed.). SAGE Publications. (No DOI; book.) ISBN 978-1412972123.
- Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17, 22–27. <https://doi.org/10.1016/j.jbef.2017.12.004>
- (PDF) Measurement of Trust in Automation: A Narrative Review and Reference Guide. (2021). *ResearchGate*, 12. <https://doi.org/10.3389/fpsyg.2021.604977>
- Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, 54(4), 1643–1662. <https://doi.org/10.3758/s13428-021-01694-3>
- Peng, R. D. (2011). Reproducible Research in Computational Science. *Science*, 334(6060), 1226–1227. <https://doi.org/10.1126/science.1213847>
- Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2016). Recommendations for Creating Better Concept Definitions in the Organizational, Behavioral, and Social Sciences. *Organizational Research Methods*, 19(2), 159–203. <https://doi.org/10.1177/1094428115624965>
- Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting*, 36(6), 691–702. <https://doi.org/10.1002/for.2464>
- Ruggeri, K., Ashcroft-Jones, S., Abate Romero Landini, G., Al-Zahli, N., Alexander, N., Andersen, M. H., Bibilouri, K., Busch, K., Cafarelli, V., Chen, J., Doubravová, B., Dugué, T., Durrani, A. A., Dutra, N., Garcia-Garzon, E., Gomes, C., Gracheva, A., Grilc, N., Gürol, D. M., ... Stock, F. (2023). The persistence of cognitive biases in financial decisions across economic groups. *Scientific Reports*, 13(1), 10329. <https://doi.org/10.1038/s41598-023-36339-2>
- Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. (2013). Ten Simple Rules for Reproducible Computational Research. *PLOS Computational Biology*, 9(10), e1003285. <https://doi.org/10.1371/journal.pcbi.1003285>
- Saunders, B., Sim, J., Kingstone, T., Baker, S., Waterfield, J., Bartlam, B., Burroughs, H., & Jinks, C. (2018). Saturation in qualitative research: Exploring its conceptualization and operationalization. *Quality & Quantity*, 52(4), 1893–1907. <https://doi.org/10.1007/s11135-017-0574-8>
- Schlosky, M. T. T., & Raskie, S. (2025). ChatGPT as a Financial Advisor: A Re-Examination. *Journal of Risk and Financial Management*, 18(12). <https://doi.org/10.3390/jrfm18120664>
- Senteio, & Hughes. (2024, August 1). *Customer Trust and Satisfaction with Robo-Adviser Technology* | Financial Planning Association. <https://www.financialplanningassociation.org/learning/publications/journal/AUG24-customer-trust-and-satisfaction-robo-adviser-technology-OPEN>
- Shefrin, H., & Statman, M. (1985). The Disposition to Sell Winners Too Early and Ride Losers Too Long: Theory and Evidence. *The Journal of Finance*, 40(3), 777–790. <https://doi.org/10.1111/j.1540-6261.1985.tb05002.x>
- Shiller, R. J. (2017). Narrative Economics. *American Economic Review*, 107(4), 967–1004. <https://doi.org/10.1257/aer.107.4.967>
- Shiller, R. J., & Pound, J. (1989). Survey evidence on diffusion of interest and information among investors. *Journal of Economic Behavior & Organization*, 12(1), 47–66. [https://doi.org/10.1016/0167-2681\(89\)90076-0](https://doi.org/10.1016/0167-2681(89)90076-0)
- Skiera, V. (2021). *The Effects of Robo-Advisers on Stock Market Participation and Household Investment Behavior*. <https://cepr.org/system/files/2022-08/Skiera-Robo-Advisers-Final-Report.pdf>
- Sophia Sophia tell me more, which is the most risk-free plan of all? AI anthropomorphism and risk aversion in financial decision-making. (2022). *International Journal of Bank Marketing*, 40(6), 1133–1158. <https://doi.org/10.1108/IJBM-09-2021-0451>
- Takayanagi, T., Suzuki, M., Izumi, K., Sanz-Cruzado, J., McCreadie, R., & Ounis, I. (2025). FinPersona: An LLM-Driven Conversational Agent for Personalized Financial Advising. *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part V*, 13–18. https://doi.org/10.1007/978-3-031-88720-8_3
- Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI | Request PDF. (2025). *ResearchGate*, 160, 108352. <https://doi.org/10.1016/j.chb.2024.108352>
- Venkatesh, V., & Davis, F. D. (2000). A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies. *Management Science*, 46(2), 186–204. <https://doi.org/10.1287/mnsc.46.2.186.11926>

- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). *User Acceptance of Information Technology: Toward a Unified View* (SSRN Scholarly Paper No. 3375136). Social Science Research Network. <https://papers.ssrn.com/abstract=3375136>
- Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology. *MIS Quarterly*, 36(1), 157–178. <https://doi.org/10.2307/41410412>
- Verma, B., Schulze, M., Goswami, D., & Upreti, K. (2025). Artificial intelligence attitudes and resistance to use robo-advisors: Exploring investor reluctance toward cognitive financial systems. *Frontiers in Artificial Intelligence*, 8, 1623534. <https://doi.org/10.3389/frai.2025.1623534>
- Warkulat, S., & Pelster, M. (2024). Social media attention and retail investor behavior: Evidence from r/wallstreetbets. *International Review of Financial Analysis*, 96, 103721. <https://doi.org/10.1016/j.irfa.2024.103721>
- Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., Haddock, S. H. D., Huff, K. D., Mitchell, I. M., Plumbley, M. D., Waugh, B., White, E. P., & Wilson, P. (2014). Best Practices for Scientific Computing. *PLOS Biology*, 12(1), e1001745. <https://doi.org/10.1371/journal.pbio.1001745>
- Winder, P., Hildebrand, C., & Hartmann, J. (2025). Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors. *PLOS ONE*, 20(6), e0325459. <https://doi.org/10.1371/journal.pone.0325459>
- Yi, T. Z., Rom, N. A. M., Hassan, N. M., Samsurijan, M. S., & Ebekozien, A. (2023). The Adoption of Robo-Advisory among Millennials in the 21st Century: Trust, Usability and Knowledge Perception. *Sustainability*, 15(7). <https://doi.org/10.3390/su15076016>
- Zhu, H., Pysander, E.-L., & Soderberg, I. (2023). Not transparent and incomprehensible: A qualitative user study of an AI-empowered financial advisory system. *Data and Information Management, Special Issue on Human-AI Interaction*, 7(3), 100041. <https://doi.org/10.1016/j.dim.2023.100041>

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.