One Advisor for the Whole World?*

Claes Bäckman†

November 19, 2026

Abstract

Households around the world now turn to AI models for financial advice. I pose three financial questions to frontier models in twenty-one countries to understand if that advice differ across countries and languages. AI advice on portfolio allocations varies little across countries, but varies more if local economic conditions warrant. Explanations invoke local financial institutions without changing allocations. Models apply remarkably similar investor-assessment rules across countries, but only selectively incorporate household balance sheets. The findings suggest that AI is globalizing the supply of financial advice and the rules used to assess investors, while selectively adapting recommendations to national financial environments.

*I thank João Cocco, Lena Lieblich, Olga Balakina, Christine Laudenbach, Mikael Paaso, Tobin Hanspal, and Ralf Elsas, as well as seminar participants at Leibniz Institute for Financial Research SAFE, Stockholm Business School and LMU for helpful comments. I gratefully acknowledge research support from the Leibniz Institute for Financial Research SAFE. This project was supported by Behavioral Finance e.V. The experimental design, regression specifications, primary hypotheses, and inference rules are pre-registered at the Open Science Framework (OSF) prior to the main data collection, accessible at https://osf.io/2ch5z. The interested reader can explore the results, the prompts, and differences across models at https://claesbackman.com/Explorables/FinancialAdvice_AI_Digital/explorer.html.

†Leibniz Institute for Financial Research SAFE. Email: claes.backman@gmail.com

1 Introduction

Retail investors are increasingly turning to artificial intelligence (AI) for financial advice, through chatbots built on large language models (LLMs) (OECD, 2026). The share of U.S. adults who use AI tools specifically for financial advice roughly doubled between 2024 and 2025 (Financial Health Network, 2025), and over half of U.K. adults now use such tools to help manage their money (Lloyds Banking Group, 2025). This represents a fundamental shift in the nature of financial advice for households all around the world, from national markets in which advice reflects local norms and institutions to a global market where advice comes from a handful of foundation models trained overwhelmingly on U.S.-English text. This shift raises several questions about the new global financial advisor: What financial advice do AI models give in different countries and do they tailor the advice to local circumstances?

To answer these questions, I use a pre-registered experiment where thirteen frontier AI models from six vendors answer three financial problems posed in twenty-one countries. I ask the questions in the local language, in English, and with local context added, and also vary the characteristics of the investor asking for advice to identify the characteristics that the models respond to. The findings speak to the supply side of financial advice from AI by identifying how models respond to the same questions in a variety of different contexts.

My first finding on the new supply side of financial advice is that AI financial advice is largely the same everywhere for a simple portfolio choice question, but that questions requiring more local context generate more dispersion. Asked how to split the local-currency equivalent of $50,000 between a broad stock-market index fund and government bonds, the models give advice that differs only modestly across countries. Compared to the United States, every significant contrast is negative, ranging from 2.7 percentage points less equity in Mexico to 11 points less in Brazil. What instead moves the advice is the identity of the model, which accounts for 69.5 percent of the variance against 5.7 percent for country. Asked whether to invest savings or use them to prepay a 4 percent mortgage, the models recommend investing 10 to 29 percentage points less than in the United States in each of the fifteen countries where that rate is realistic. The largest gaps are in Japan, China, Switzerland, and Germany. When the rate is removed from the question, the models assume local mortgage rates themselves, the gaps in the low-rate countries largely close, and the high-rate countries move

toward prepaying. Asked how to protect savings from 3 percent inflation, the models switch from equities to inflation-linked bonds in Brazil, Turkey, and Mexico, but overwhelmingly recommend equities in all other countries. On these two questions model identity explains only 15 and 26 percent of the variance. The pattern suggests that the models localize when the question depends on the local economy.

The lack of cross-country differences raises the question of what, if anything, the advice should track. Giving the same advice everywhere on what share to invest in stocks is not a mistake per se: under integrated markets the risky portfolio is the same world index everywhere and uniform advice is warranted. Of course, there are actually some differences across countries in the advice, so the models are not just giving exactly the same advice everywhere. As a simple (but not realistic) benchmark, a Merton benchmark with a currency-hedged world fund prescribes 0.52 in every country, within about six points of the advice everywhere. However, this model cannot explain why some countries receive less equity than the United States.[1] A more realistic benchmark would take local pension systems and background wealth into account, but as I show later, the models do not measurably adapt their advice based on these variables. An alternative approach to understanding what models track across countries is to relate cross-country differences to country-level fundamentals. To test this, I relate the estimated country fixed effects to country-level risk-taking from the Global Preferences Survey (Falk, Becker, Dohmen, Enke, Huffman and Sunde, 2018), to financial literacy from the S&P Global FinLit Survey (Klapper, Lusardi and van Oudheusden, 2015), and to fourteen other country-level fundamentals. However, these country-level fundamentals have only limited explanatory power: country differences in financial advice on the risky share do not track measured risk preferences, financial literacy, local market returns, income levels, equity volatility, the interest rate, or the Merton benchmark. Advice on the mortgage and inflation questions tracks the ten-year yield and the local mortgage rate, again showing that the models can incorporate local context if it is relevant.

The next finding is that AI models give the kind of "popular financial advice" studied by

[1] Another variable that differs across countries and could affect portfolio choice is currency risk (Campbell and Viceira, 2002). While it is feasible that the models could take this into account, a benchmark with unhedged local-currency moments prescribes equity shares from 0.21 in Turkey to 0.53 in South Korea, yet mean advice spans only 0.46 to 0.58. Later on, I also show that providing the models with information about currency hedging has a negligible impact on advice. Further, using local stock market returns instead of a global return does not help match the dispersion across countries.

Canner, Mankiw and Weil (1997) and Choi (2022). The evidence for this comes from several different tests. First, I insert statements about the investor into otherwise identical English prompts for each country, covering risk tolerance, currency hedging, pension composition, pension generosity, the share of her liquid wealth that the stake represents, and investment horizon. Stated risk tolerance raises the recommended equity share by 46 percentage points, consistent with the dominance of risk tolerance in Ross and Lo (2026). A twenty-year rather than one-year horizon raises the equity share by 40 points, in contrast to the classic Merton–Samuelson irrelevance result where horizon should not affect the share at all under CRRA preferences with i.i.d. returns. Second, I ask the models to list the questions they would put to the investor before advising. In every country, the models ask the same four things: risk tolerance, what the money is for, whether the five-year horizon is firm, and whether an emergency fund exists. Every response except one asks about risk tolerance, and three quarters of the responses ask all four. These questions are reminiscent of popular financial advice, and it is striking that there are only negligible differences in the main questions the models ask across countries, but the local content shows up as currency and inflation risk in emerging markets. The models ask less often about the investor's pension, or about local circumstances, both of which would be highly relevant in, for example, life-cycle models.

I then conduct several checks that directly ask how the balance sheet affects the advice. A stated generous pension adds safe background wealth, so it should raise the equity share if the model integrates the investor's balance sheet. However, the advice barely moves when information about pensions is added to the prompt. This is not because the models ignore pensions, as nearly every explanation mentions the investor's existing pension. In additional tests, added after the pre-registered results, I find that the advice moves strongly and in the predicted direction when a stated balance-sheet cue raises the investor's exposure, as with a pension invested in stocks or a stake equal to her entire liquid wealth, but it barely moves on cues that signal safety, such as a guaranteed annuity, a currency hedge, or a cash cushion. The models appear to read the balance sheet as a source of hazard instead of seeing the balance sheet as a two-sided input, as portfolio theory would prescribe. Overall, AI advice fits popular financial advice more than life-cycle models that incorporate the household's balance sheet.

The final finding is that the models appear to solve the portfolio problem in a shared, English-anchored representation regardless of the prompt's language, and then explain the answer in the local language and with local context. This is consistent with the interpretability literature, which finds that multilingual models solve tasks in a shared, English-leaning concept space (Wendler, Veselovsky, Monea and West, 2024; Tang, Luo, Huang, Zhang, Wang, Zhao, Wei and Wen, 2024; Wu, Yu, Yogatama, Lu and Kim, 2025; Lindsey, Gurnee, Ameisen, Chen, Pearce, Turner, Citro et al., 2025). One response in Swedish, a language I speak, describes a "glidväg" ('glide path'), a direct translation of the English industry term with no particular meaning in Swedish.[2] More formally, I machine-code the explanation that each model gives for its recommendation into four binary indicators. Local financial context appears in 18 percent of U.S. explanations but 29 to 78 percentage points more often in every other country. Naming local instruments in the prompt raises local references by 23 percentage points, but moves the allocation by less than a point. The translation changes the words of the answer but not the advice, suggesting that the explanations localize far more than the advice itself.

These results hold up across a set of pre-registered and exploratory robustness checks. For example, I use a VPN to change the location where the prompt is sent from, which does not change the result. Sending the prompts through a U.S. rather than a German VPN exit moves the advice by 0.4 points. I also adjust the stake by purchasing power raises the share by a marginal one point and accounts for a sixth to a third of the Brazil, Poland, and Mexico contrasts, where the adjustment is largest. The results are robust across a number of similar checks, I describe more of these robustness checks when discussing the main results. The interested reader can go to this website to explore the results themselves and see, for example, how the results change by model generation or what the prompt was in each country. You can also download the code for running your own experiment. With papers based on AI-generated data, the question is always whether the results will hold for future model generations. I argue that the structure of the advice will, for two reasons. First, the portfolio question has no clear answer. Models can be trained to ignore gender or race, because these should not affect a financial recommendation. It is much less clear

[2]The formal term in Swedish is *generationsfond* or a reference to the default public AP7 fund. These terms do not appear often in the explanations, however: *generationsfond* appears zero times across roughly 5 million lines of response text.

how country or language should affect the optimal equity share. The answer depends on theory, where there is no clear recommendation, and on preferences that the prompt does not state. What the models learn instead is the body of written financial advice, which changes slowly. The departures from theory that Canner et al. (1997) documented remain in the popular advice that Choi (2022) surveys twenty-five years later. Second, the results do not reflect a lack of knowledge that a more capable model would fill. The models assume the local mortgage rate, switch to inflation-linked bonds where inflation is high, and cite the local pension system in their explanations. They have the local information and do not use it on the portfolio question. The evidence so far supports this view. A panel of twenty-one models collected eleven weeks after the main collection, fourteen of them new, reproduces the variance split, with 70 percent for model identity against 5 percent for country, and 77 percent of its non-U.S. cells lie below the same model's U.S. cell. The fourteen new models in that panel differentiate more across countries, but not in the direction of the benchmark. This argument concerns the advice rule and not the level of advice, which differs across models and can move between versions of the same model.

Summing up, this paper characterizes the supply side of an emerging global market for AI financial advice. Existing research primarily examines the quality and determinants of AI financial advice within particular countries, models, or investor populations. This paper instead examines its geographic structure across countries, models, and financial decisions. The findings reveal a tension between standardization and localization. AI models apply broadly similar investor-assessment rules across countries, asking the same questions and responding similarly to investor characteristics wherever the investor is located. Yet their recommendations incorporate local economic conditions to very different degrees across financial decisions. The explanations are more geographically differentiated still, frequently invoking local financial institutions even when the recommended allocation barely changes. These findings suggest that we are moving to one advisor for the whole world, with AI models globalizing not only the supply of financial advice but also the rules used to assess investors, while selectively adapting the recommendations to national financial environments.

Historically, financial advice has been a nationally segmented service industry, shaped by regulation, relationships, and language (Reuter and Schoar, 2024). AI is changing this market by making advice from a handful of foundation models available across countries and lan-

guages. This transformation not only means that investors everywhere have broader access to financial advice, but also that the rules used to evaluate investors are becoming global. Standardization is not necessarily undesirable, but the concentration of advice production in a small number of models creates the potential for common recommendations and common errors across households in otherwise distinct national markets.

2 Literature

There is a growing literature on AI financial advice, which also relates closely to the literature on robo-advising and traditional financial advice. Beginning with this newer literature, Choukhmane, de Silva, Lin and Akuzawa (2026) elicit prompts from a representative U.S. sample, simulate the life-cycle paths that following the advice would generate, and confront those paths with the prescriptions of a calibrated life-cycle model. They find that current models give moderately risk-averse advice that is broadly consistent with life-cycle theory, yet relies on round-number heuristics and varies with the gender and financial literacy of the person asking. The reliance on round numbers is also a feature of the results of this paper. Ross and Lo (2026) use a different approach, where they draw a thousand synthetic client profiles with near-orthogonal characteristics, collect allocations from a single vendor's model family, and fit interpretable surrogate models. They find evidence for "heuristic collapse": self-reported risk tolerance is the most important predictor of advice, while age, income, horizon, and liquidity contribute only marginally. Similarly, Fieberg, Hornuf, Meiler and Streich (2025) elicit portfolio recommendations from 32 LLMs for 64 hypothetical investors who differ in risk tolerance, investment horizon, sustainability preferences, gender, and investment experience, for hypothetical investors in China, Germany and the United States. Most models recommend implementable ETF portfolios whose risk exposure rises with the investor's stated risk tolerance, and these portfolios earn historical risk-adjusted returns similar to those of the robo-advisors. The recommendations overweight domestic securities, consistent with a home bias inherited from the training data. Finally, Foltyn and Olsson (2026) query 33 models from five vendors with prompts that differ only in the word "man" or "woman," and find that the models advise women to hold 1.8 percentage points less in equities than men. The gap falls to 0.8 points but does not vanish when the prompt adds marital status, education, income, and risk tolerance. This paper adapts their stock-bond

question as its primary outcome and replaces the gender treatment with the country and language of the question.

This paper is also relevant for the literature on robo-advising and algorithmic financial advice. Earlier robo-advisors improve diversification and reduce behavioral biases (D'Acunto, Prabhala and Rossi, 2019), deliver substantial welfare gains (Rossi and Utkus, 2024), and broadly move household behavior closer to normative prescriptions (D'Acunto and Rossi, 2023). In credit markets, machine-learning scoring changes the distribution of mortgage approvals across racial groups (Fuster, Goldsmith-Pinkham, Ramadorai and Walther, 2022). LLMs differ from these robo-advisors, because they are general-purpose rather than domain-restricted and because they reach essentially every language and country at once. These models therefore loosen the supply-side constraints on financial advice that partly determine who accesses and benefits from advice (Reuter and Schoar, 2024).

What kind of advice do investors usually get? Canner et al. (1997) show that popular advisors recommend a higher ratio of bonds to stocks for more conservative investors, contrary to the mutual-fund separation theorem. Choi (2022) documents how popular personal financial advice departs from the prescriptions of economic theory. In other studies, advisor fixed effects explain more of the variation in client portfolios than client characteristics do (Foerster, Linnainmaa, Melzer and Previtero, 2017), which suggests that human advisors give one-size-fits-all advice. AI models reproduce both these patterns, but do so in many different countries at once.

Finally, this paper contributes a new test to the literature on how multilingual language models process and retrieve knowledge. There are two strands of this literature. Interpretability studies find that the middle layers of multilingual models operate in a shared concept space. This space, where the models solve problems, lies closer to English than to any other language (Wendler et al., 2024; Tang et al., 2024; Wu et al., 2025; Lindsey et al., 2025). Knowledge-probing studies instead find that the same model often asserts different facts depending on the query language (Kassner, Dufter and Schütze, 2021; Qi, Fernández and Bisazza, 2023). These two findings make different predictions for the prompts in this study. Shared processing predicts nearly the same advice in every language, as the portfolio is solved once in the middle layer in English. Language-keyed retrieval predicts that advice inherits the conventions of each language's financial text, meaning that a prompt in

German would bring up German-language texts with a specific type of advice. The results support the first interpretation, shared processing, where the local detail added by the models appears in the explanation rather than in the allocation. The questions that the models would like to ask investors also vary very little across countries, which again supports shared processing.

3 Experimental Design

The experiment has two arms. In the main arm I pose three financial questions to thirteen frontier LLMs for twenty-one countries. I vary the language of each question, the country it refers to, and the local financial context it provides. In the second arm, I vary the investor rather than the country. I refer to this arm as the *stated-investor arm*. Specifically, I use the portfolio question with amended statements on the investor's risk tolerance, pension generosity, investment horizon, currency hedging and other balance-sheet items in an otherwise fixed English prompt. I describe this arm in more detail when I discuss the associated results in Section 6.

This section instead focuses on the first arm, and describes the three questions and the country selection, the prompt versions and the contrasts they identify, the models and API parameters, and the data collection. Appendix Table A1 lists the prompt language, the local-currency investment amount, and the full Version C vocabulary for every country, and Appendix A documents the technical implementation of the data collection.

3.1 The Three Questions

Each LLM answers three financial questions: a primary portfolio-allocation question and two secondary questions on mortgage prepayment and inflation protection. Question 1 asks the LLM to split the local-currency equivalent of $50,000 between a broad stock-market index fund and government bonds over a five-year horizon. The outcome is the fraction allocated to the stock-market index fund. Question 2 asks whether to invest savings in a stock-market index fund or to use them to pay down an existing 4 percent mortgage with twenty years remaining. The outcome is the fraction allocated to the index fund. This question has a sharper normative benchmark than Question 1, because mortgage prepayment is a risk-free

investment with a return equal to the contract rate. Using a 4 percent rate also allows me to test whether the models condition on a reasonable rate for each country. I use a pre-specified rule that assigns any country whose prevailing mortgage rate exceeds twice the prompted 4 percent to a premise-acceptance group. In these countries the question also tests whether the models flag the 4 percent rate as unusual for the local environment or simply condition on it. In the end, all models conditioned on it.

Question 3 asks how to protect the real value of $50,000 in a savings account that earns a nominal 1 percent interest rate when annual inflation is 3 percent. Similar to Question 2, the explicit rate in the question allows me to understand whether models push back against an unrealistic rate. The outcome has two parts: a numeric fraction to invest and a categorical `primary instrument` (equities, inflation-linked bonds, real estate, gold, or other). Because the stated 3 percent inflation rate is far below prevailing inflation in some countries, Question 3 was ex ante the question most likely to lead the models to reject the premise. I use a pre-specified contingency that promotes the refusal indicator to the primary Question 3 outcome if the refusal rate exceeds 10 percent in any country–model cell. However, the refusal rate was zero across 37,021 model calls in 273 country $\times$ model cells and the contingency never applied. While several models noted that the inflation rate was unrealistically low for the country, most often for Turkey, none refused to answer.

3.2 Country Selection

I chose the country set to span meaningful cross-country variation in pension generosity, equity-market development, inflation history, and measured preferences, while keeping translation risk low. The preference regressions of Section 5.5 require each country's measured risk preferences, so the Global Preferences Survey (Falk et al., 2018) must cover every country. This requirement excludes otherwise natural candidates such as Denmark.

The United States serves as the baseline for the country contrasts. I prompt five further countries in English (Australia, Canada, India, South Africa, and the United Kingdom). They add observations at essentially no translation risk while spanning very different pension institutions, from compulsory superannuation to a replacement rate among the lowest in the panel.

The fifteen non-English countries include high-literacy European systems with different occupational pension schemes (Sweden, Germany, the Netherlands, Switzerland) and further European variation (France, Spain, Italy, Poland, Finland). Outside Europe they cover the Asian markets (China, Japan, South Korea), several high-inflation economies (Brazil, Turkey), and Mexico.

3.3 Prompt Structure and Version Design

Each prompt has two blocks. An Instructions block, written in English in every cell, asks for a response that contains the recommendation and a word-limited explanation. Keeping the instructions in English holds response-format compliance constant across languages. A Question block describes the investor and poses the financial question. The main treatments change only the Question block, but a robustness check in Section 5.6 that re-collects the primary cell with the instructions translated as well shows that this has only a limited impact on the results.

The Question block varies along three dimensions. The first dimension is the country and its prompt language. I prompt six countries in English and the fifteen others in their local language, with Switzerland in German and Mexico in Spanish (Appendix Table A1 lists each pairing). The second dimension of the prompt is the investor's stated age (30, 40, or 55), which varies for the primary equity-share question and is fixed at 40 for the two secondary questions. The third dimension is the prompt version. I use three versions of the prompt, described in detail below: the first version uses the local language with little or no local context, the second version uses English but states the investor location, and the third includes local context in the form of pension and fund names. I use these different prompts to decompose the country effect into language, country-naming, and local-framing components.[3]

More specifically, *Version B* states "I live in [country]" but uses a generic instrument vocabulary ("a broad stock market index fund," "government bonds," "a standard pension plan").

[3]The version lettering follows the pre-registration. Version A denoted a translated prompt that named no country. I dropped it before collection because it could not separate the effect of the question's language from the style of any particular translation, and because the comparison of Versions B and B-cross identifies the language effect more cleanly by holding the country reference fixed. I keep the labels of the surviving versions unchanged so that the paper matches the registered design.

In the fifteen non-English countries it appears in the local language. Its investment amount is the round local-currency equivalent of $50,000 (e.g. 500,000 kronor in Sweden, 45,000 euros in Germany, 40,000 pounds in the United Kingdom), written in the native local number format ("500 000 kr," "45.000 Euro"). *Version B-cross* retains the country reference ("I live in [country]") and the local-currency investment amount of Version B but reverts the question to English. It is collected only for the fifteen non-English countries. Note that the baseline prompts do not adjust for PPP or income. In a robustness check I also adjust the stake by PPP, which does not alter the result.

Version C keeps the language, country reference, and investment amount of Version B but replaces the generic financial vocabulary with country-specific equivalents. Specifically, I use local index funds (e.g. an index fund via Avanza or Nordnet in Sweden, eMAXIS Slim All-Country Equity in Japan), local government bonds (*Svenska statsobligationer*, JGBs, Tesouro Direto), and local pension references (*tjänstepension*, bAV, PER, iDeCo).

The different versions allow me to identify different components of the country effect. The cross-country comparison of Version B identifies the joint country-and-language effect relative to the U.S. baseline. The within-country comparison of Versions B-cross and B holds the country reference fixed and changes only the language of the question, which identifies the language effect. The within-country comparison of Versions C and B holds both country and language fixed and changes only whether the prompt names local instruments and pensions, which identifies the local-framing effect.

The translations were first done by an automated pipeline that translates only the Question block, followed by a native-speaker review of each version. Before I filed the pre-registration, every assembled prompt passed a consistency audit with zero failures, an adversarial LLM-based review covered all ninety unique translated blocks, and I read every block in full myself with the help of Google Translate. I then froze the translations for collection. Appendix A.4 documents the pipeline mechanics, the automated checks, and the review-and-correction protocol.

3.4 Models and Data Collection

The experiment queries thirteen frontier LLMs from six vendors through the OpenRouter unified API. Four vendors are U.S.-headquartered: Anthropic (Claude Haiku 4.5, Claude Sonnet 4.6, Claude Opus 4.6, Claude Opus 4.8), OpenAI (GPT-5-mini, GPT-5.5), Google (Gemini 3 Flash, Gemini 3.5 Flash, Gemini 3.1 Pro), and xAI (Grok 4.3). To make sure that the panel does not consist solely of U.S.-headquartered firms, I add DeepSeek (DeepSeek V4 Flash, DeepSeek V4 Pro), a non-U.S. vendor whose models are primarily trained on Chinese-language data, and Mistral (Mistral Medium 3.5), a European (French) vendor. These six firms account for roughly 98 percent of web visits among the leading standalone chatbots (Einberger, 2026), and about 85 to 94 percent of AI-assistant app users (Sensor Tower, 2026). OpenAI, Google, and Anthropic account for 56.6, 24.7, and 11.7 percent of web visits, respectively.[4] Note that the denominator does not include embedded assistants such as Meta AI or Chinese domestic platforms.

A cell is one combination of country, version, question, age, and model. I query each cell fifty times at temperature 0.7, send no system prompt, and cap the response at 4,000 tokens. Three models, GPT-5-mini, GPT-5.5 and Claude Opus 4.8, do not accept a temperature setting, and the interface drops the parameter without an error, so these models ran at their providers' default sampling. The setting is the same for every query to a given model, so the model fixed effects absorb any difference it makes to the level of advice.[5] Fifty repetitions are enough because the unit of inference is the model. Additional repetitions would reduce only the within-cell sampling term, which is already small relative to between-model heterogeneity. In total the design comprises 267,150 queries, of which 185,250 form the country/language arm and 81,900 the stated-investor arm. Appendix Table A2 gives the accounting by arm,

[4]The numbers in Einberger (2026) are based on Similarweb estimates of desktop and mobile web visits. Shares are of combined visits to seven standalone assistants (ChatGPT, Gemini, Claude, DeepSeek, Grok, Perplexity, Microsoft Copilot). Country shares refer to September 2026, worldwide and U.S. shares to October 2026. The combined six-vendor shares are my calculations. Mistral is not included in the source data. Similarweb figures are model-based estimates. See Jansen, Jung and Salminen (2022) on the accuracy of Similarweb relative to site-side analytics.

[5]The pre-registration provides for a model that rejects a custom temperature, provided the substituted setting is documented and held constant across all queries to that model. I identify the three models from OpenRouter's listing of the parameters each model endpoint accepts, and the temperature check of Section 5.6 confirms it, since the dispersion of their answers does not rise between temperatures of 0 and 1.0 (Appendix Table A15).

version, and question.

I ran an exploratory pilot in late June 2026 and filed the pre-registration at the Open Science Framework (OSF) on July 2, 2026, with the prompts frozen. The main collection ran on July 2–4, beginning a few hours after the filing. Two pre-registered robustness cells, temperature bookends at 0 and 1.0 and a Version B variant that states the investment amount in U.S. dollars, added 8,450 queries on July 6. Every cell reached its fifty repetitions, with one observation per query. The parser recovered a numeric recommendation from 99.9 percent of responses, and no model refused to answer. I exclude 237 unparsed responses and, under the pre-registered truncation rule, a further 7 that exhausted the 4,000-token budget, which leaves 266,906 responses. Appendix A.5 documents the parser, the refusal rule, and the model-identity checks. Appendix Table A3 lists the dated version of each model, and Appendix A.1 reproduces the prompts.

After the main collection I added three rounds of exploratory collection, none of which is part of the pre-registration. In August 2026 I collected the balance-sheet probes of the stated-investor arm (Section 6.3). In September I ran a model-tracking wave with a panel of twenty-one models. In October I re-collected the primary cell three times, with the Instructions block translated, with the model allowed to ask questions before advising, and with the stake restated at purchasing-power parity. Section 5.6 reports the tracking wave and the three re-collections.

4 What Should Move the Advice?

Before turning to the main results, it is worth thinking about what should move the advice for investors in different countries. Holding the investor's preferences, investment opportunities, and balance sheet fixed, the advice should not change merely because the investor names a different country or asks in a different language, unless there are cross-country differences that would warrant a different allocation. A lower recommended equity share in Brazil than in the United States would be inappropriate if the two investors face the same economic situation, but may reflect good advice if Brazil's fundamentals warrant more conservative advice. This section asks which fundamentals the models should take into account. Later, I use the intuition developed in this section to separate three candidate advice rules: a

Merton advisor that applies equation (1) to the stake alone, a Life-Cycle advisor that applies equation (2) to the household's balance sheet, and a Questionnaire advisor that follows a retail risk questionnaire (Section 6).

To this end, I construct a simple benchmark to think through the economic fundamentals that would motivate country differences in advice. I also use a simple benchmark to evaluate the mortgage advice using the contract rate. Appendix B develops the benchmarks in full, and reports the country gaps under different hedging assumptions, the sensitivity of the results to risk aversion, and the influence diagnostics.

4.1 A simple benchmark

The simplest benchmark treats the $50,000 stake as a standalone portfolio of a world equity index and domestic government bonds. For an investor in country c with constant relative risk aversion γ_c, the optimal equity share is

$$\theta_c^* = \frac{\mu_c - r_c}{\gamma_c \, \sigma_c^2}, \tag{1}$$

where $\mu_c - r_c$ is the expected excess return and σ_c the volatility of the equity fund, both measured in local currency. This is the rule of Merton (1969) and Samuelson (1969), where the share does not depend on the size of the stake or, with independent returns, on the investment horizon. If the uncovered interest parity holds, the equity premium is constant across currencies and countries. Under these assumptions, the optimal equity share across countries would vary only due to σ_c and γ_c. Importantly, how much the benchmark varies depends on whether the world fund is hedged. For a currency-hedged fund and with a common risk aversion parameter across countries, the simple benchmark prescribes a share of 52 percent in equity everywhere. If we add exchange rate risk, the prescribed share then ranges from 0.21 in Turkey and 0.23 in Brazil to 0.53 in South Korea at $\gamma = 3$, a spread of thirty-two percentage points. I later return to whether actual advice tracks this spread.

This simple benchmark treats the stake as a standalone investment. If we add in more of the balance sheet, treat pension wealth and human capital as outside wealth B_c, and decompose

outside wealth into safe and risky assets, we can rewrite the optimal equity share as

$$\alpha_c^* \;=\; \min\left\{1,\; \max\left\{0,\quad \underbrace{\theta_c^*\left(1+\frac{B_c^s+B_c^r}{W}\right)}_{\text{equity demand on total wealth}} \quad - \quad \underbrace{\frac{B_c^r}{W}}_{\text{equity held outside}}\right\}\right\}, \tag{2}$$

where θ_c^* is the Merton share of equation (1), W is the stake, B_c^s is safe outside wealth, and B_c^r is outside wealth that carries equity risk. With safe outside wealth ($B_c^r = 0$), which is an order of magnitude larger than the stake for an average 40-year-old, the prescribed share reaches the 100 percent corner in every country (Appendix B.2). With common risk aversion, the benchmark therefore varies across countries only through the currency risk of an unhedged fund.

Equation (2) organizes the tests in the rest of the paper around two properties. The first is invariance. Prompt content that leaves the economics unchanged should leave the advice unchanged. This covers the country name (Section 5.2), the language of the question and the local names of the instruments (Section 5.3), and, with independent returns, the investment horizon (Section 6). The only reason these variables should move the advice is if the models infer something about the investor or the problem. However, the models need to do that in a way that relates to the terms in the question.

This brings us to the second property, which is responsiveness. The terms inside the equation should move the advice. The volatility σ_c differs across countries for an unhedged fund. Risk aversion may differ across countries or for the individual investors. Note that risk aversion is not stated in the main prompts, but that I include it in the stated-investor arm. Safe outside wealth B_c^s rises with pension generosity and falls with age as human capital declines. Equity-like outside wealth B_c^r rises when the pension is invested in stocks.

The rest of the paper tests these two properties in turn. Section 5 tests invariance to the country name, the language of the question, and the local names of the instruments, and asks whether the country effects track fundamentals. Section 6 varies the investor's age, risk tolerance, pension size and composition, and the investment horizon in the prompt. Table 2 in Section 6.4 collects the evidence on both properties.

Finally, there is a caveat to the invariance tests. The prompt does not say whether the fund is hedged, what taxes apply, or what else the investor owns, so a model may fill in these details from the country name. A country effect can therefore reflect a reasonable inference about local conditions instead of a failure of invariance. I return to this later.

4.2 A benchmark for the mortgage question

The mortgage question involves different economics. Paying down a mortgage is a risk-free investment whose return is the after-tax real mortgage rate. In the question, the prompted rate is 4 percent, which should be adjusted for the deductible share of interest and expected inflation. Prepayment is a better investment only if this return exceeds the local real bond rate (Becker and Shabani, 2010). Otherwise the household should keep the mortgage and hold bonds for the safe part of its portfolio. The relevant safe return is therefore the higher of the two rates. Replacing r_c in the Merton rule (1) with it gives the benchmark share α_c^{Q2} (see Appendix B.5). Where the mortgage rate is the higher of the two, prepayment is a better investment when inflation is low and a worse one when interest is deductible. Where the bond rate is higher, α_c^{Q2} equals the allocation benchmark α_c^S.

Because effective deductibility depends on income and itemization, the appendix also reports α_c^{Q2} across a band of deductible shares from zero to one half. For the five premise-acceptance countries of Section 3, the after-tax real mortgage rate is negative and below the local real bond rate, and so prepaying is dominated by an investment in local bonds. This makes the mortgage benchmark equal to the simple benchmark above. I report these countries separately. The pre-registration does not specify this benchmark, so the comparison is exploratory. The benchmark takes the prompted 4 percent as given. Section 5.4 also reports advice when no rate is stated, for which no benchmark is defined.

5 Main Results

This section reports how LLM advice varies across countries. I first document the raw patterns in the advice, then I estimate the country effects on the primary question, decompose them into language and local framing, and ask whether they track measured risk preferences and financial literacy. I then turn to the mortgage and inflation questions and close with

robustness. The pre-registered test of age and pension generosity is reported with the other tests of the advice rule in Section 6.2.

Except for the two secondary questions in Section 5.4, the outcome is the recommended equity fraction on Question 1. Every regression includes model fixed effects, so each contrast compares the advice of a single model. I cluster standard errors at the model level.[6] As pre-registered, the twenty primary country contrasts carry a Holm step-down correction as a single family. The reason for the Holm step-down procedure is that I test 20 countries against the United States. With twenty contrasts each tested at the 5 percent level, about one would appear significant by chance even if advice were identical in every country. The correction keeps the probability of any false rejection among the twenty at 5 percent. Appendix Table A7 lists each pre-registered hypothesis with its test and outcome. The coefficients measure the causal effect of the prompt's country, language, and framing on AI financial advice.

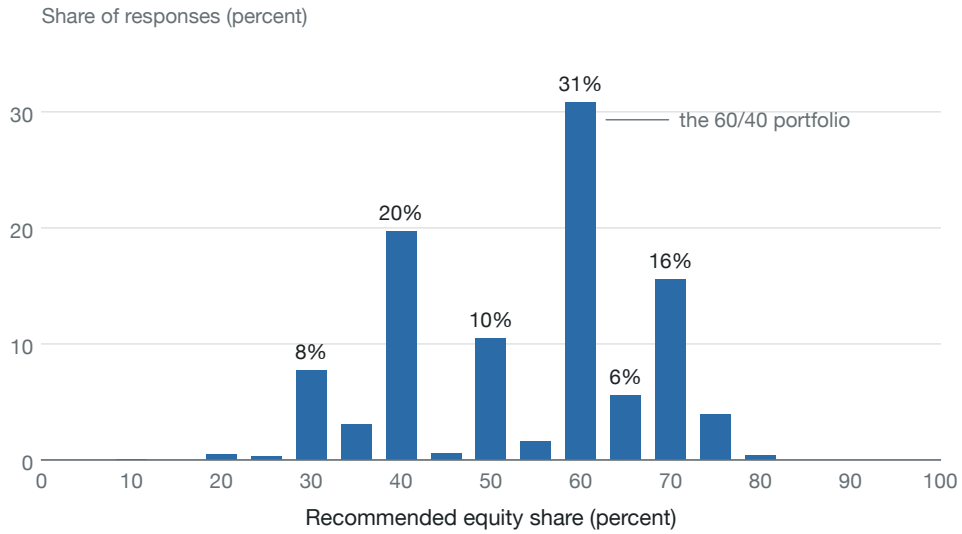
5.1 Descriptive Patterns

I begin with descriptive facts about the advice that organize the results that follow. The models bunch their recommendations at a few round numbers within a narrow range. I can reject equality between countries. The differences between countries are one-sided (every significant contrast is negative) and modest (six points on average, at most eleven).

Starting with round numbers in the recommendations, Figure 1 plots the distribution of the recommended equity share, pooled across countries and models. Every response is a multiple of five percentage points, and 85 percent are multiples of ten. The modal recommendation is 60 percent equity, given in 31 percent of responses. In addition, the support is narrow. No response in 13,639 allocates less than 10 or more than 80 percent to equities. The bunching shows that the models answer a continuous allocation problem with a conventional round number. Choukhmane et al. (2026) also find round-number bunching in AI advice, which they note departs from the more subtle prescriptions of life-cycle theory. A portfolio rule would produce a continuous answer for at least some of the answers. Instead, the modal 60/40 split is the most common target allocation in retail portfolio advice, a first sign of the risk-

[6]Because thirteen clusters are too few for reliable asymptotic inference, every reported p-value comes from a wild cluster bootstrap with Webb six-point weights and 9,999 replications.

Figure 1: Distribution of recommended equity shares in the primary cell



Notes: Share of responses at each recommended equity share on the primary cell (Question 1, Version B, age 40, $N = 13{,}639$), pooled across the twenty-one countries and thirteen models. Bars holding at least 5 percent of responses are labeled with their share.

questionnaire interpretation. This is not just because the models lack information about the investor, as 100 percent of the responses in the stated-investor arm are also round numbers. Ross and Lo (2026) report that between 89 and 100 percent of all allocation percentages are multiples of 5 without web search, concluding that "round-number anchoring is a deeply ingrained property of model behavior that web search alone does not fully override."

The design also varies the investor's age. The recommended share falls by about 0.7 percentage points per year of age, a gradient that Section 6.2 examines together with the pension system.

Across countries, Appendix Table A8 shows that mean risky share spans twelve percentage points, from 0.46 in Brazil to 0.58 in South Korea. The United States is at 0.57, near the top of the range. The band is narrow relative to the dispersion within any single country. The typical within-country standard deviation is 0.13. Instead, most of the within-country dispersion comes from the models. Table 1 reports mean advice by model, which spans 0.37 (Gemini 3.1 Pro) to 0.71 (Claude Haiku 4.5), a range of 34 percentage points. That

Table 1: Advice by model: level, dispersion, and stability

				Stability across repetitions	
Model	Vendor	Mean	SD	Within-cell SD	Identical cells
Gemini 3.1 Pro Preview	Google	0.37	0.08	0.037	0.05
GPT-5-mini	OpenAI	0.38	0.08	0.043	0.00
GPT-5.5	OpenAI	0.41	0.05	0.032	0.00
DeepSeek V4 Flash	DeepSeek	0.45	0.12	0.110	0.00
DeepSeek V4 Pro	DeepSeek	0.48	0.12	0.105	0.00
Grok 4.3	xAI	0.54	0.09	0.065	0.00
Gemini 3 Flash Preview	Google	0.56	0.07	0.000	0.62
Gemini 3.5 Flash	Google	0.57	0.06	0.020	0.24
Claude 4.8 Opus	Anthropic	0.58	0.04	0.000	0.52
Claude 4.6 Sonnet	Anthropic	0.63	0.04	0.000	0.67
Claude 4.6 Opus	Anthropic	0.65	0.07	0.000	0.95
Mistral Medium 3.5	Mistral	0.70	0.04	0.024	0.29
Claude 4.5 Haiku	Anthropic	0.71	0.04	0.007	0.48

Notes: Recommended equity share by model on the primary cell (Question 1, Version B, age 40), pooled across the twenty-one countries, roughly 1,050 responses per model. Within-cell SD is the median, across a model's twenty-one country cells, of the standard deviation over the fifty repetitions in the cell. Identical cells is the share of those cells whose fifty repetitions return one identical number, despite sampling at temperature 0.7. A variance decomposition of the primary cell attributes 69.5 percent of the total variation to model identity, 5.7 percent to the country, 9.5 percent to their interaction, and 15.3 percent to within-cell sampling.

is nearly three times the country range. The ordering of the models follows neither vendor nor scale. The four Anthropic models cluster above the pooled mean, between 0.58 and 0.71. Google's flagship reasoning model is the most conservative advisor in the panel, but its two Flash-tier models are twenty points higher. Model identity accounts for 69.5 percent of the total variation in the primary cell, the country for 5.7 percent, their interaction for 9.5 percent, and sampling variation within country–model cells for the remaining 15.3 percent. The dominance of the model motivates two choices maintained throughout the paper. Every regression identifies country contrasts within model, and inference clusters at the model level, treating the thirteen models rather than the quarter-million responses as the effective sample.

5.2 Country effects on the primary question

I now formally test whether LLM advice is the same across countries. The primary specification estimates country fixed effects from the following regression

$$y_{imc} = \alpha_m + b_c + \varepsilon_{imc}, \tag{3}$$

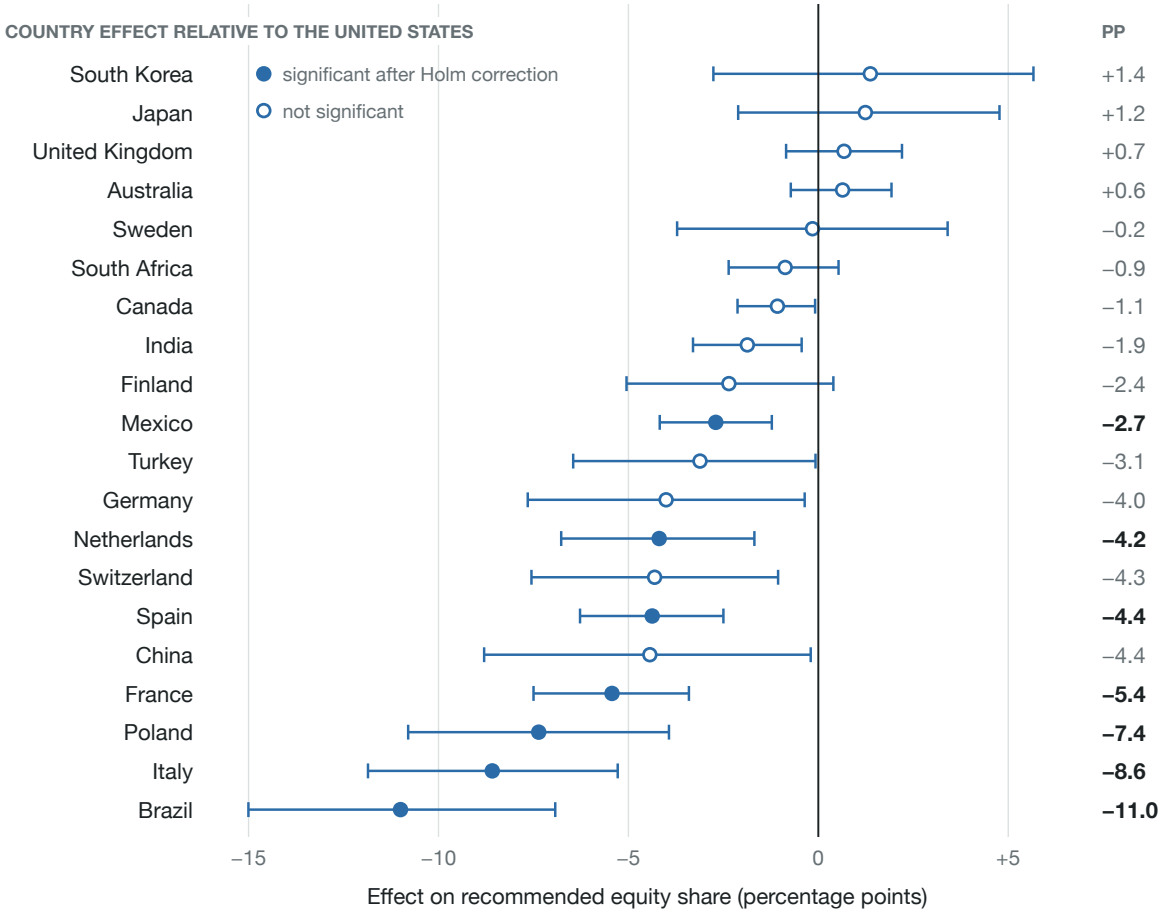
where y_{imc} is the equity fraction recommended in repetition i by model m for country c, the α_m are model fixed effects, and the b_c are deviations from the omitted U.S. baseline. The baseline regression uses the age-40 version of the prompt in the local language.

The main finding is that advice is not the same everywhere, as shown in Figure 2 for all twenty contrasts.[7] Thirteen of the twenty contrasts clear the wild-bootstrap threshold before correction, and seven survive the Holm step-down: Brazil at −11.0 percentage points, Italy at −8.6, Poland at −7.4, France at −5.4, Spain at −4.4, the Netherlands at −4.2, and Mexico at −2.7. In addition, every contrast that survives the correction is negative. Against a U.S. mean equity share of 0.57, the largest gap, Brazil's eleven percentage points, moves $5,500 of the $50,000 stake from stocks to bonds. The seven significant contrasts average about six points. For comparison, the gender gap that Foltyn and Olsson (2026) document on a similar allocation question is under two percentage points. Country context moves LLM advice by several times the demographic gaps documented so far in the literature.

After the Holm correction, thirteen of the twenty contrasts fail to reject equality with the U.S. baseline. Six of those are individually significant on the uncorrected wild-bootstrap test but do not survive the step-down (China, Switzerland, Germany, Turkey, India, and Canada, ranging from −4.4 to −1.1 points). The remaining seven do not reject even before correction. Four of them have small positive point estimates (Australia, Japan, South Korea, and the United Kingdom). Finland and South Africa have small negative gaps, and Sweden sits at the U.S. level. The precision of these nulls differs across countries, because the models agree with one another far more in some countries than in others. In the three English-prompting

[7] Appendix Table A8 reports each with its wild-bootstrap and Holm p-values. Because the Holm step-down controls the familywise error rate across the twenty contrasts, the seven contrasts that survive it reject equality of advice across the twenty-one countries at the 5 percent level. No joint Wald test is reported, because the model-clustered covariance of twenty country coefficients has rank at most twelve with thirteen clusters.

Figure 2: Country effects on the recommended equity share



Notes: Point estimates of the country fixed effects from equation (3) (Question 1, Version B, age 40), in percentage points relative to the omitted U.S. baseline. Bars are 95 percent wild cluster bootstrap confidence intervals (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$), the intervals of record, matching Appendix Table A8. Filled markers denote contrasts that are significant at the 5 percent level after Holm correction, and the column on the right repeats the point estimates.

countries (Australia, South Africa, and the United Kingdom) the 95 percent confidence intervals exclude gaps larger than about two and a half percentage points in either direction, which is \$1,250 of the \$50,000 stake. For the United Kingdom this is the pre-registered equivalence test H3. Its contrast is +0.7 points with a 95 percent confidence interval of $[-0.8, +2.2]$, inside the registered ± 5 point margin. In Finland, Japan, South Korea, and Sweden the intervals are wider and reach between 3.4 and 5.7 points from zero, so the data cannot exclude gaps of the size estimated for Spain or the Netherlands.

5.3 Language and local framing

I next run two within-country tests. One asks whether local framing shapes the advice, and the other separates the effect of the question's language from that of the country name. The local-framing regression compares Versions C and B across the twenty non-U.S. countries at age 40,

$$y_{imc} \;=\; \alpha_m + b_c + \phi\,\mathbf{1}[\text{version} = C] + \varepsilon_{imc}, \tag{4}$$

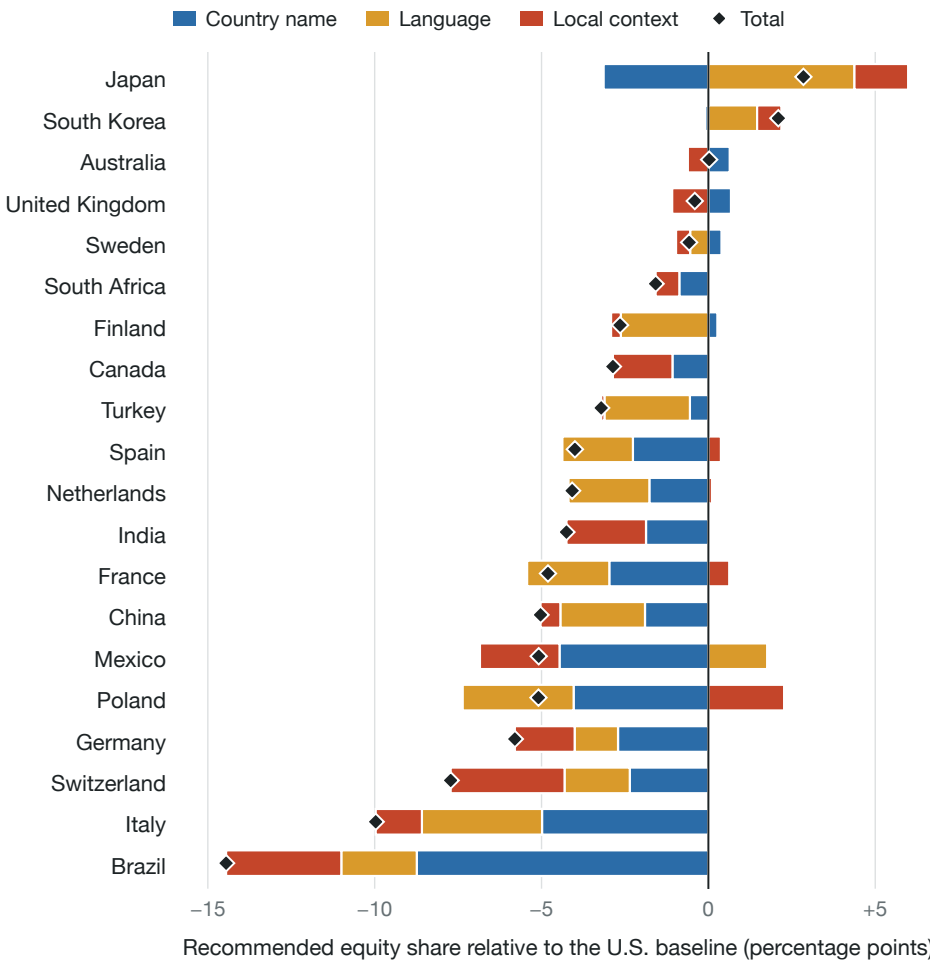
and the language regression compares Versions B and B-cross across the fifteen non-English countries,

$$y_{imc} \;=\; \alpha_m + b_c + \lambda\,\mathbf{1}[\text{language} = \text{local}] + \varepsilon_{imc}. \tag{5}$$

Asking in the local language rather than in English lowers the recommended equity share by 1.3 percentage points ($\hat{\lambda} = -0.013$, wild bootstrap $p = 0.007$, confidence interval -2.3 to -0.4 points). The language effect is statistically significant but economically small, an order of magnitude below the largest country effects. Naming the local instruments does even less. Replacing the generic vocabulary with local pension names, local government bonds, and local index funds moves the allocation by -0.7 points, a result that is not significant at the five percent level. Figure 3 applies the same decomposition country by country. The country-name step is the English version of the prompt minus the United States (Version B minus U.S.), the language step is the native-language prompt minus the English prompt for each country (Version B minus B-cross), and the local-context step is the native-language prompt with local context minus the native-language prompt (Version C minus Version B). The country-name step also changes the currency unit and the nominal amount, which move together with the country reference. In the large-effect countries (Brazil, Mexico, Poland, France, Italy, Germany) the country-name slab accounts for most of the gap. The language and local-context slabs are usually small, but they are noisier than the pooled $\hat{\lambda}$ and $\hat{\phi}$ and not uniformly negative. Japan is the clearest exception. A large positive language slab pushes its advice toward stocks and produces the panel's largest positive total.

The size of the language effect helps distinguish between two views of how multilingual models answer a question. If the models retrieve advice the way they retrieve language-keyed knowledge, the language effect λ should be substantial. The same model often asserts different facts depending on the query language (Kassner et al., 2021; Qi et al., 2023). For

Figure 3: Decomposing the country effect into naming, language, and local context



Notes: Each bar decomposes a country's fully localized effect on the recommended equity share, Version C relative to the omitted U.S. English generic baseline (Question 1, age 40), into three additive cues in percentage points. Country name is the English-only contrast Version B-cross minus the U.S. (language and instruments held fixed, though the currency unit and nominal amount change with the country reference), language is Version B minus B-cross (country name held fixed), and local context is Version C minus B (language held fixed). The black diamond marks the total, which equals the sum of the three slabs by construction. The six English-prompting countries have no B-cross cell, so their language slab is zero and country name is Version B minus the U.S.

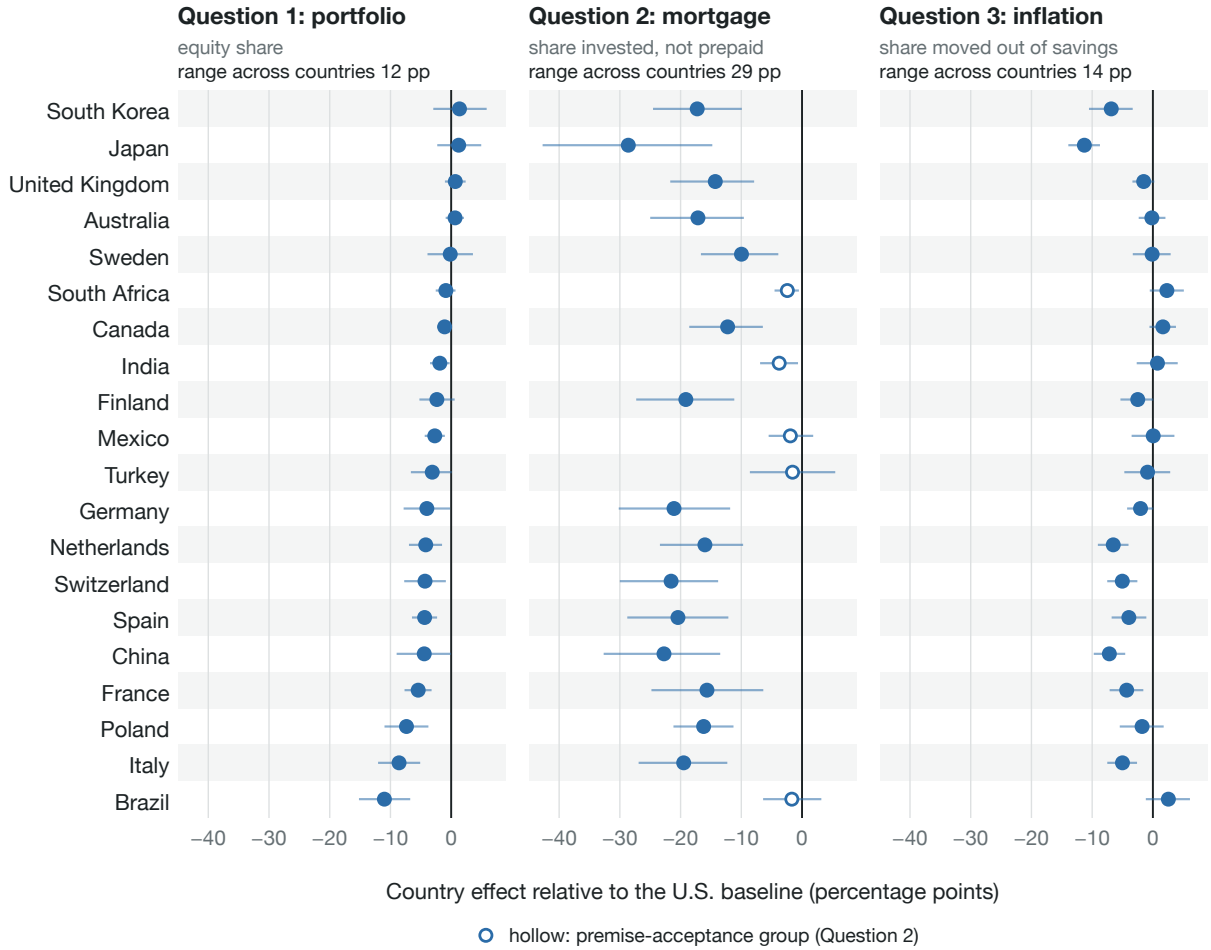
example, if German queries were to draw on the bond-heavy conventions of German-language financial text, the effect of the language would presumably be large. If instead the models solve the task in a shared representation, λ should be near zero. Tracing studies in computer science find that language-specific processing lies in the layers nearest the input and output (i.e., the initial prompt and the explanation, which is always in the local language). More importantly, however, tracing studies also find middle layers, where the LLM solves the task in a shared concept space lying closer to English (Wendler et al., 2024; Tang et al., 2024; Wu et al., 2025; Lindsey et al., 2025). Under such processing, two translations of the same question are nearly the same query in the middle layer, which supports a small language effect. Indeed, the estimate of -1.3 points fits this second account, with the small residual consistent with the language-specific features that remain. The advice behaves like one policy applied across query languages rather than like language-segmented retrieval.

5.4 The mortgage and inflation questions

I now turn to the two secondary questions. Figure 4 plots the country effects on all three questions side by side. For the two secondary questions, the advice differs more across countries and the variance decomposition reverses. Country accounts for 27 percent of the variation in mortgage advice against 15 percent for model. Model identity accounts for only 26 percent of the variation in inflation advice, whereas on the allocation question the shares are 6 percent for country and 70 percent for model. Agliata and Hasso (2026) also note the importance of the model for financial advice. The secondary questions also pick up a different country signal from the portfolio allocation question. The countries that receive less equity on the allocation question (Brazil, Italy, Poland and France) are not the countries told to prepay (Japan, China, Switzerland, and Germany). Japan, statistically at the U.S. level on the allocation question, has the largest gap on both secondary questions. Across countries the allocation and mortgage effects correlate at -0.10, while the mortgage and inflation effects correlate at 0.80.

On the mortgage question, the models recommend investing 10 to 29 percentage points less than in the United States in each of the fifteen countries outside the premise-acceptance group. All fifteen contrasts survive the Holm correction. Among the largest gaps are Japan at -28.6, China at -22.7, Switzerland at -21.6, Germany at -21.1, and Spain at -20.4.

Figure 4: Country effects across the three questions



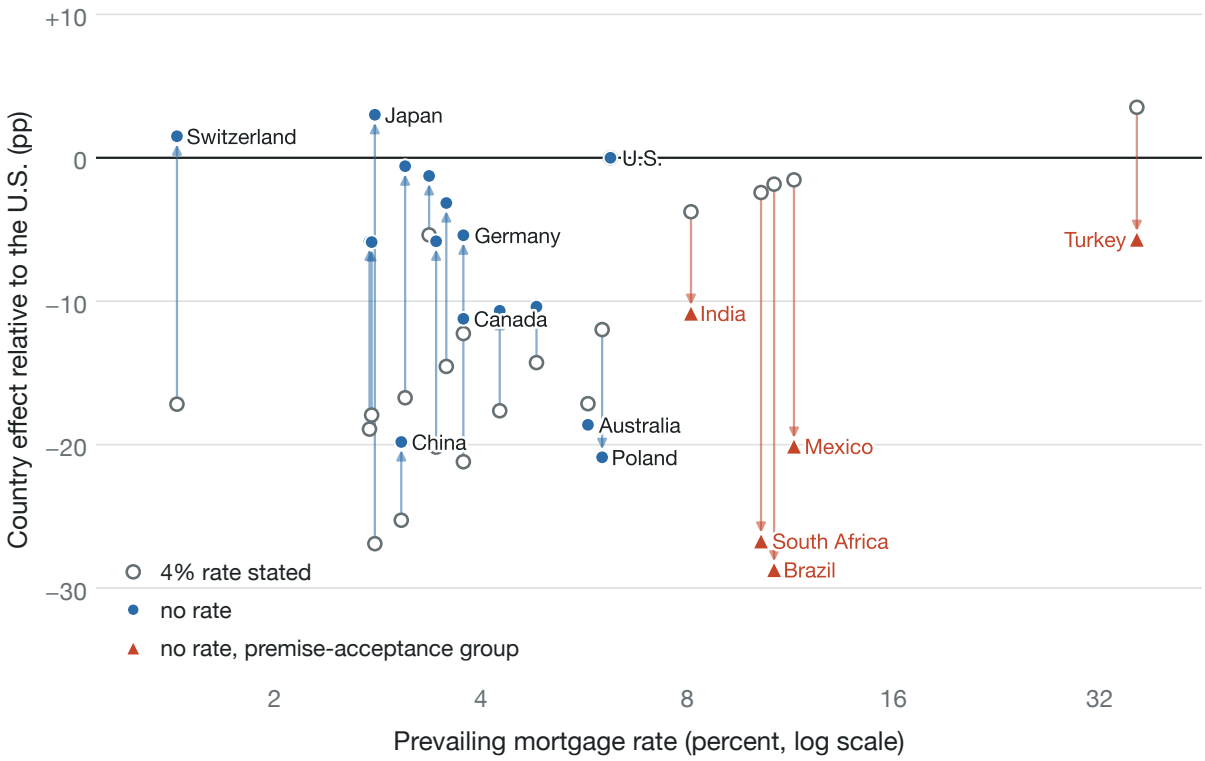
Notes: Point estimates of the country fixed effects from equation (3) for Question 1 (recommended equity share), Question 2 (recommended fraction invested rather than used to prepay a 4 percent mortgage), and Question 3 (recommended fraction moved out of a 1 percent savings account), Version B, age 40, in percentage points relative to the omitted U.S. baseline. Each question is estimated separately with model fixed effects. The three panels share one horizontal scale. Lines are 95 percent wild cluster bootstrap confidence intervals, reported in Appendix Tables A8 and A9. The range above each panel is the distance between the largest and the smallest country effect, with the United States at zero. Hollow markers denote the premise-acceptance group on Question 2. Countries are ordered by their Question 1 effect.

Japan's gap alone is more than twice the largest contrast on the allocation question. Part of the common gap reflects the United States itself, where deductible and freely prepayable mortgages favor investing. The ordering among the fifteen countries, however, spans a further 19 points and places the low-inflation economies at the bottom. For the five countries where the prevailing mortgage rates exceeded twice the prompted 4 percent at the filing date (Brazil, India, Mexico, Turkey, and South Africa), the contrasts are mostly null, and only South Africa (-2.4 points) survives the correction.

On the inflation question, seven country contrasts survive the Holm correction, all negative. In these countries the models recommend moving less of the savings out of the account than in the United States: Japan (-11.3 points), China (-7.2), South Korea (-6.9), the Netherlands (-6.5), Italy (-5.0), Switzerland (-5.0), and France (-4.3). In addition to the share, I also ask the models to recommend an instrument to save in (see Appendix Table A13). Here we find the clearest evidence for localization: Equities are the primary instrument in 93 percent or more of responses in fourteen countries and in a clear majority in all but the highest-inflation economies. In Brazil, Turkey, and Mexico the models switch to inflation protection. In Brazil, inflation-linked bonds account for 86.7 percent of the recommendations, and in Turkey inflation-linked bonds and gold together account for 82.3 percent (66.0 and 16.3 percent). Mexico's recommendations tilt the same way (58.1 percent inflation-linked bonds), and Poland's are split (38.6 percent inflation-linked). The tilt appears even inside the English-prompting group, where models in the two higher-inflation members, India and South Africa, recommend inflation protection (11.5 and 16.3 percent, compared to under 3 percent in the United States, the United Kingdom, Canada, and Australia). Because these two countries share the baseline's language, the switch tracks the country rather than the language of the prompt. The models' own beliefs follow the same ordering. Asked for the inflation rate they assume when the mortgage question states no rate (described below), they give 4.3 to 5.3 percent for Brazil, Mexico, India, and South Africa and 40.6 percent for Turkey, against 1.4 to 3.7 percent in the other sixteen countries.

The country effects on the mortgage question sort into three groups by the local level of interest rates. In Japan and Switzerland, where policy rates were below one percent in 2024, the models recommend 25 points less investment than in the United States. In the five premise-acceptance countries, where policy rates exceeded six percent, the gap is 2 points.

Figure 5: The mortgage question with and without the stated rate



Notes: The English Question 2 block (Version B in the six English-prompting countries, B-cross elsewhere) at age 40, asked on November 16 and 17, 2026 with the clause stating the 4 percent mortgage rate deleted. Thirteen models, twenty-five repetitions per country. Country fixed effects from equation (3) relative to the United States, plotted against the prevailing mortgage rate in the pre-registered rate supplement (log scale) and estimated separately for the numbered English cell of the main collection (hollow markers, 4 percent stated) and for the arm without the rate (filled markers). Arrows connect the two for each country. Triangles mark the premise-acceptance group. Estimates, confidence intervals, and Holm-corrected p-values are in Appendix Table A10.

In the remaining thirteen countries the gap averages 17 points and is unrelated to the local rate (slope 0.07 points per point of the policy rate, $p = 0.96$). The pattern is consistent with models comparing the stated 4 percent with local rates only where the two differ widely.

To test this directly, I removed the stated 4 percent rate from the mortgage question.[8] After

[8]On November 16 and 17, 2026, I put the question without the rate to all thirteen models in all twenty-one countries, for twenty-five repetitions. The arm was collected after filing and is exploratory. For details, see Appendix A.1.

the recommendation, I also ask each model for the mortgage rate and the inflation rate it assumed when answering the question.

The first hint that the models localize can be found in the rate that they assume. Across the twenty-one countries, the assumed and the prevailing mortgage rate correlate at 0.92 in logs, and the correlation for each model lies between 0.83 and 0.96. The one large departure is Japan, where the models assume 1.1 percent against 2.8 percent in the data.

Once the models choose the rate, Figure 5 shows that they give a different recommendation. In the eleven countries whose prevailing rate is below 4 percent, the average gap to the United States narrows from 17.9 to 5.0 percentage points. The change survives the Holm correction in eight countries. Japan moves from 26.9 points below the United States to 3.0 points above it, and Switzerland from 17.2 below to 1.5 above. Nine of the eleven contrasts are no longer distinguishable from zero. The premise-acceptance countries, where the mortgage rate was at least 8 percent, move the other way. Their average gap widens from 1.2 to 18.4 points, with Brazil at 28.7 and South Africa at 26.7 points below the United States. Across the twenty non-U.S. countries, the gap falls by 1.0 point for each point by which the prevailing rate exceeds 4 percent (wild bootstrap $p < 0.001$).

Turkey is the one premise-acceptance country that moves little. Its gap changes by 9.2 points, which is not statistically significant. The models assume Turkish inflation of 40.6 percent against a mortgage rate of 26.6 percent, which makes the real rate on the mortgage negative. With a negative real rate, it would be sensible to keep the mortgage and not prepay. The real return on prepaying is the interest rate on the mortgage, and if that rate is high, investing in stocks rather than prepaying is less favorable (Becker and Shabani, 2010). The models' advice is consistent with this rule. Within a country and model, responses that assume a mortgage rate one point higher recommend investing 1.6 percentage points less ($p = 0.02$), and responses that assume inflation one point higher recommend investing 0.8 points more ($p = 0.10$). I cannot reject that only the assumed real rate matters ($p = 0.13$).

Hence, with the rate stated, the models condition on the premise, even if the premise is unrealistic. The large gaps in the low-rate countries record how they price a 4 percent loan in economies where such a loan is expensive. Without the rate, they price the decision at local conditions that match the data.

5.5 What does the advice track?

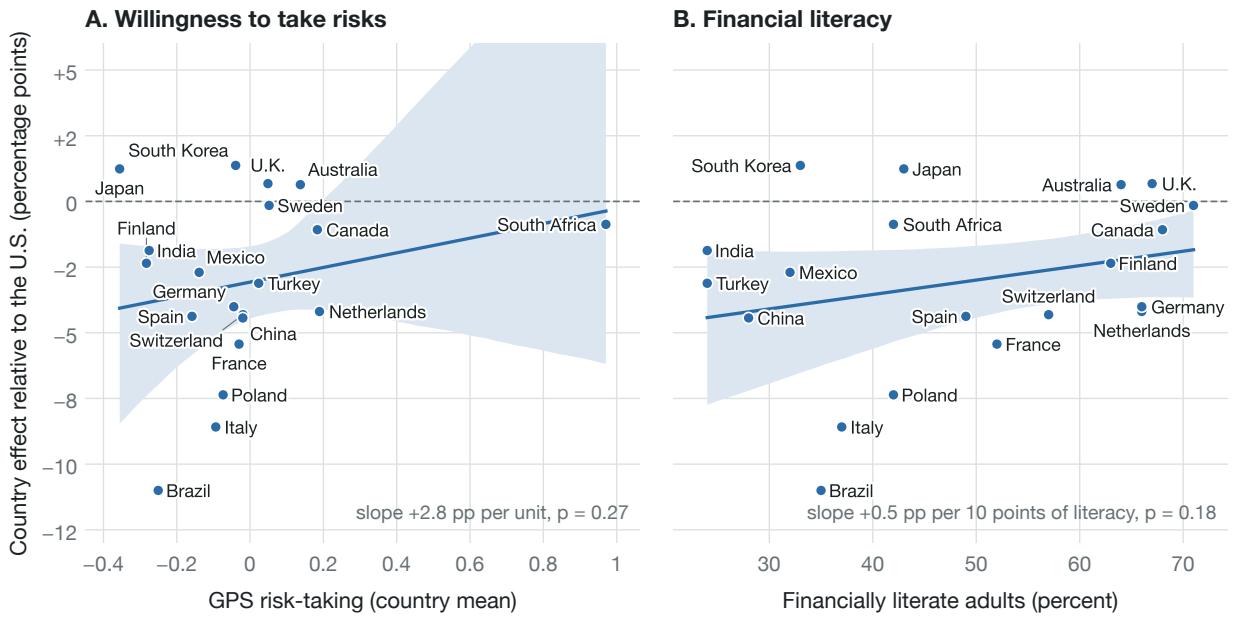
Financial advice from AI models differs across countries. I now ask whether the country effects track several country-level measures that an advisor may pick up on. The answer depends on the question. On the allocation question, the country effects track none of the characteristics I consider. On the mortgage and inflation questions, they track local interest rates.

I start with the willingness to take risks and financial literacy. I use data from the Global Preferences Survey from Falk et al. (2018) for risk-taking and data on the share of adults classified as financially literate in the S&P Global FinLit Survey from Klapper et al. (2015). If AI models tailor their advice to country characteristics, they may plausibly infer that a country with a higher willingness to take risks should be given a higher equity share. Similarly, if models take financial literacy into account, they may give a higher equity share to a country with higher financial literacy. Financial literacy also proxies for a wide range of other characteristics, like general financial sophistication or high stock market participation (van Rooij, Lusardi and Alessie, 2011). I test these two predictions by regressing the country fixed effects from equation (3) on each measure.[9] Note that these tests are descriptive and that they ask whether the models impute national averages, not whether they should.

Figure 6 plots both relationships along with the statistical results. I find little support for either hypothesis. Panel A shows that the advice does not track measured risk preferences. The slope is flat and the wild bootstrap over countries does not reject zero. The point estimate is also economically small. A country that is one cross-country standard deviation more risk-tolerant than another receives 0.8 percentage points more equity. Even the widest preference gap in the sample, between Japan at -0.36 and South Africa at 0.97, predicts a gap of under four points. At the same time, Japan receives advice statistically indistinguishable from the U.S. baseline, while South Africa, by far the most risk-tolerant country, receives slightly *less* equity than the baseline rather than more. Panel B tells the same story for literacy. Moving across the full literacy range in the panel, from 24 percent of adults in India and Turkey to 71 percent in Sweden, changes predicted advice by under three percentage

[9] This is a pre-registered test for H7 and H8. Each test is a bivariate slope over the twenty non-U.S. countries, estimated with a wild bootstrap over countries. Both hypotheses are pre-specified in full and registered as descriptive.

Figure 6: Country effects against measured risk preferences and financial literacy



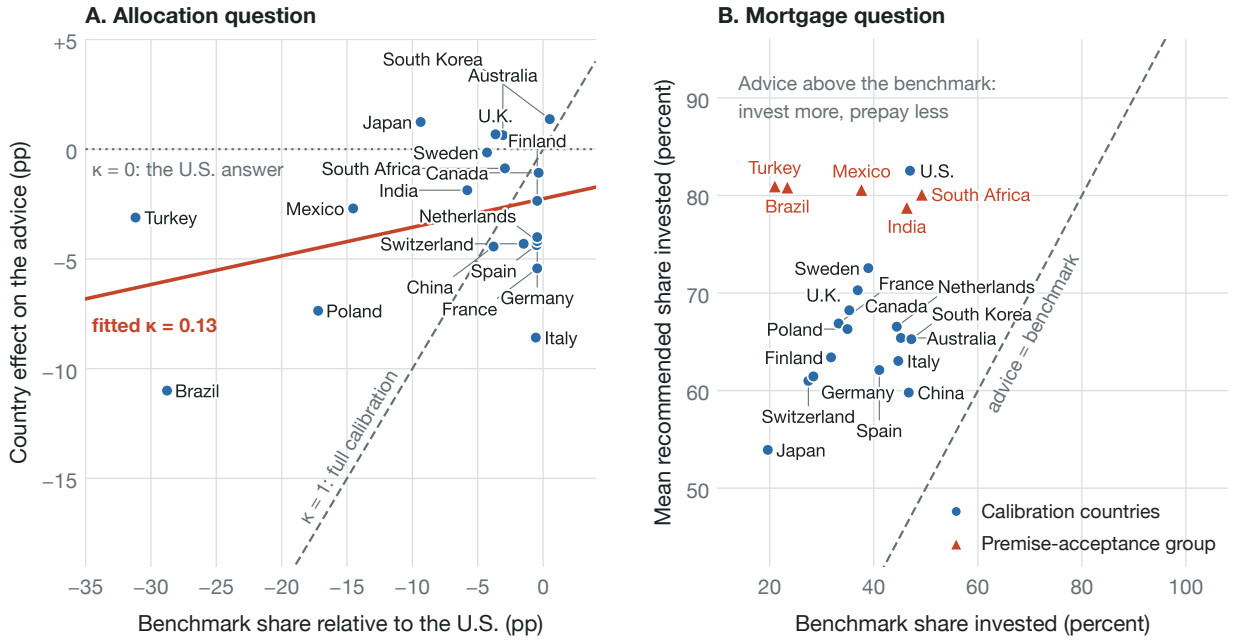
Notes: Estimated country fixed effects $\hat{b}_c$ from equation (3) (Question 1, Version B, age 40, U.S. baseline omitted) plotted against the country mean of the GPS willingness to take risks (Falk et al., 2018) in Panel A and the share of financially literate adults from the S&P Global FinLit Survey (Klapper et al., 2015) in Panel B, for the twenty non-U.S. countries. Country effects are in percentage points. Ribbons are 95 percent country-bootstrap confidence bands around the fitted bivariate line (9,999 replications). Each panel reports the bivariate slope with its wild bootstrap p-value.

points.

Measured preferences and literacy are not special in this respect. Appendix Table A14 regresses the country effects on each of sixteen country characteristics in turn, covering survey measures, income, the volatility and home equity premium, policy, long-term and mortgage interest rates, inflation, pension replacement rates, and mortgage-interest deductibility, with a wild bootstrap over the twenty countries and a Holm correction within each question.[10] For Question 1 on the portfolio allocation, no characteristic survives the correction for multiple

[10] I standardize continuous characteristics so that each slope is the change in the country effect, in percentage points, per standard deviation of the characteristic. Interest rates and inflation enter in logs. The English-prompt indicator is a dummy for English-speaking countries (United Kingdom, Australia, Canada, India and South Africa). These regressions were specified after the pre-registration was filed and are exploratory. Apart from H7 and H8, none of them is registered.

Figure 7: Advice against the benchmark on the allocation and mortgage questions



Notes: Panel A plots the estimated country fixed effects $\hat{b}_c$ from equation (3) (Question 1, Version B, age 40) against the benchmark's prescribed deviations from the United States, $\alpha_c^S - \alpha_{US}^S$, at the baseline grid point ($\bar{\gamma} = 3$), for the twenty non-U.S. countries, both in percentage points. The dashed line is full calibration ($\kappa = 1$), the dotted horizontal line is the U.S.-default anchor ($\kappa = 0$), and the solid line is the fitted regression (11) with slope $\hat{\kappa} = 0.13$. Inference of record is the wild bootstrap over countries reported in the text. Panel B plots the mean recommended index-fund share on the mortgage question (Question 2, Version B, age 40) against the country-specific benchmark α_c^{Q2} from Equation (12) at the same grid point. The dashed line is the locus of full calibration, on which advice equals the benchmark. Points above it recommend more equity, and hence less prepayment, than the benchmark. Triangles mark the premise-acceptance group.

hypothesis testing. The two largest unadjusted associations are the English-prompt indicator and the pension replacement rate, but neither is significant after the Holm step-down ($p = 0.33$ and 0.25). Overall, no country characteristic is significantly associated with the country coefficient on advice in Question 1.

Average advice lies within about six percentage points of the currency-hedged version of the simple benchmark in Section 4.1, but up to 33 points above the unhedged version in the countries with the most volatile currencies (Appendix Table A4). Regressing the country effects from Question 1 on the benchmark's prescribed deviations from the United States

gives a slope of $\hat{\kappa} = 0.13$ (Figure 7, Panel A). Dropping Brazil from the estimation lowers the slope to 0.03, dropping Turkey raises it to 0.21, and dropping any other country leaves it between 0.11 and 0.15. These results are all far from a slope of 1, which would indicate full pass-through of the currency risk, and are closer to zero, which would indicate that all countries get the same advice as the United States. The wild bootstrap over countries rejects full calibration ($p < 0.001$ against $\kappa = 1$) but does not reject the U.S.-default hypothesis ($p = 0.32$ against $\kappa = 0$).

However, the results for the mortgage and inflation questions are different. Interest rates and inflation positively predict the answer to the mortgage question. A one-standard-deviation higher log mortgage rate is associated with 6.6 percentage points more recommended investment, which survives the Holm correction (Holm $p = 0.006$, $R^2 = 0.67$), and the long-term rate and inflation each explain close to half or more of the cross-country variance. The models therefore draw on local economic conditions when a question involves a local interest rate, but not on the allocation question. Only part of this response is appropriate. Where inflation is high, the real return to prepaying the prompted 4 percent mortgage is low, so advising less prepayment is right. It does not follow that the savings belong in equities, because in the high-rate countries local bonds pay more than the mortgage costs. Across all twenty countries the advice tracks local rates but not its own benchmark (slope 0.5 points per standard deviation, $p = 0.86$).

Panel B of Figure 7 plots the mortgage advice against the benchmark of Section 4.2. In every calibration country the models recommend investing more than the benchmark prescribes, by 13 to 36 percentage points and by 28 on average (Appendix Table A5). Where the benchmark lies below 0.40, from Japan to Sweden, the gap is nearly constant at 31 to 34 points, so the advice rises one for one with the benchmark. Above 0.40 the advice stops rising and the gap narrows to between 13 and 22 points. The United States is the exception. Its benchmark of 0.47 is among the highest, yet it receives the highest advice in the panel, 0.83, and the largest gap, 36 points. On the fifteen non-U.S. calibration countries, the calibration regression gives $\hat{\kappa}^{Q2} = 0.23$ for the mortgage question, against 0.07 for the allocation question on the same fifteen countries. With fifteen countries neither slope is statistically distinguishable from zero, and the 95 percent confidence intervals of both exclude full calibration at one. In the five premise-acceptance countries the models lean against prepaying, with average advice at

about 0.80, within two points of the U.S. level. The benchmark agrees that these households should not prepay, but it does not prescribe more equity. Because local bonds pay more than the mortgage costs, the savings not used for prepayment belong in bonds, so the benchmark is the allocation benchmark of 0.21 to 0.49. The advice exceeds it by 31 to 60 points, more than in any calibration country. Some models name the bond option explicitly. One answer for Turkey notes that "even government bonds or deposit rates offer returns well above 4 percent" but still places the entire sum in the index fund, because "the question asks for a choice between paying the mortgage and investing."[11]

5.6 Robustness

I now describe several robustness checks. I first collect two pre-registered checks where I ask whether the results depend on the sampling temperature or on the size of the numerals in the local-currency amount.[12]After these checks, I also added several exploratory robustness checks where I first check whether the results change over time as new models are introduced, whether the English Instructions block affects the results, or whether allowing models to ask questions affects the cross-country differences. Finally, I also examine whether the answers depend on the country where I was when I sent the queries (Germany). The results are robust to these checks, which I now describe in more detail.

Temperature check. Temperature is a sampling parameter that sets how much randomness a language model uses when it picks each word of its response. The next-token prediction at the heart of an LLM picks the next word based on a predicted probability. Model providers introduce some randomness in the token the model uses, which introduces some variation in the answers even with a similar prompt. At a temperature of 0, the model always chooses its most likely next word. Higher values make less likely words more probable. The main set of prompts uses a sampling temperature of 0.7. The temperature check re-collects the primary cell at the temperatures of 0 and 1.0 for five countries (the United States, Germany, Sweden,

[11]DeepSeek V4 Pro, Version B, age 40, my translation. The original reads "Ayrıca, devlet tahvilleri veya mevduat faizleri bile %4'ün çok üzerinde getiri sunmaktadır. [...] soruda ipotek ödemesi ile yatırım arasında seçim sorulduğu için, 2 milyon TL'nin tamamının yatırıma yönlendirilmesi önerilir."

[12]I had several other pre-registered checks, but these have nothing to act on. The collection produced no refusals, so the refusal outcome is constant and the zero-equity imputation changes no observation. The collection also finished within two days, so estimating effect by week of collecting does not have any effect.

Japan, and Brazil). Three models do not accept a temperature setting (Section 3.4), so for them the bookend cells only repeat the baseline, and I estimate the check on the other ten. Appendix Table A15, Panel A shows that temperature does not affect the results. The pooled mean equity share shifts by 0.5 percentage points at either bookend relative to the 0.7 baseline (wild bootstrap $p = 0.40$ at temperature 0 and $p = 0.32$ at 1.0), and the country contrasts barely move. On all thirteen models, the registered specification, the shifts are 0.3 points ($p = 0.55$ and $p = 0.40$).

Numeral-magnitude. The second robustness check addresses a numeral-magnitude confound. Version B states the stake as a round native-currency figure. The cross-country contrasts compare prompts whose numerals differ by orders of magnitude, such as 7,500,000 yen in Japan against 50,000 dollars in the U.S. baseline cell. If an AI advisor anchored on the size of the number, this could generate spurious country effects. For Sweden, Japan, and Germany I therefore re-collect Version B with the amount stated as $50,000, holding country and language fixed (Appendix Table A15, Panel B). The pooled shift is a tight zero (-0.1 percentage points, wild bootstrap $p = 0.84$, confidence interval -1.6 to $+1.4$ points). The one significant per-country shift is Japan, whose seven-digit yen figure raises the recommended share by 2.8 points relative to the dollar statement ($p = 0.01$). Japan's country contrast is small and statistically null with native numerals ($+1.2$ points) and remains small under the dollar statement (an implied -1.6 points).

Purchasing power adjustment. The main collection uses $50,000 converted at market exchange rates to the local currency. At purchasing-power parity, the stake is worth about twice the U.S. stake in Brazil, Mexico, Poland, and China, two and a half times in South Africa, and three and a half to four times in Turkey and India. If the models advise more caution on a stake that is large relative to local prices, part of these countries' contrasts would reflect the size of the stake rather than the country.

On October 29, 2026, I restated the stake in the twenty non-U.S. Version B prompts as the purchasing-power-parity equivalent of $50,000, rounded in each prompt's own notation, and re-collected the unchanged Version B prompts for all twenty-one countries, with twenty-five repetitions per model and cell. Appendix Table A16 reports the shifts. Cutting the real stake raises the recommended equity share by 1.0 percentage point on average (wild

bootstrap $p < 0.001$, confidence interval 0.5 to 1.7 points). Each halving of the real stake adds 1.2 points ($p < 0.001$). However, the effect is small relative to the baseline contrasts. For example, Brazil's contrast moves from -11.5 to -9.7 points and Poland's from -8.4 to -6.1. Stake size therefore accounts for between a sixth and a third of these contrasts. Moreover, the effect is not specific to emerging markets, since Spain's contrast narrows by a similar 2.1 points.

Leave-one-out mean and vendor fixed effects. The pre-registered leave-one-model-out sweep re-estimates equation (3) thirteen times, each time without one model. No contrast moves by more than 1.3 percentage points. Six of the seven contrasts that survive the Holm correction survive it in all thirteen samples, with a largest Holm p-value of 0.027. Replacing the thirteen model fixed effects with six vendor fixed effects, for comparability with Foltyn and Olsson (2026), leaves every country contrast essentially unchanged, although with six clusters the variant supports no formal inference.

Model drift over time. To examine whether newer models give different results and to check the robustness of the answers over time, I re-asked one English slice of the primary cell on September 19–20, 2026 to a panel of twenty-one models. Fourteen of these models were new to the panel. The other seven carry over from the main panel under unchanged identifiers. Appendix Table A17 reports the results.

The main findings are robust. On the same English slice, model identity accounts for 70 percent of the variance in September against 78 percent in July, and country for 5 percent against 3 percent. Within each model, 77 percent of non-U.S. cells lie below the model's own U.S. cell, against 67 percent in July. Six of the seven carried-over models return average advice within 2.5 points of July.[13] The fourteen new models give country a larger role, 15 percent of the variance against 43 percent for model identity, but the added variation does not follow the benchmark. Turkey, where the benchmark prescribes the least equity, receives the most. The advice behind an unchanged identifier can also move. Gemini 3 Flash lowers its Brazil recommendation by 9 points, and DeepSeek V4 Flash recommends 8.5 points more equity because OpenRouter sent most of its September queries to hosts that switch off its reasoning step.

[13]The interested reader can explore the differences in the answers by model themselves on this website.

Caller location and routing. Every query went through OpenRouter from Germany, which raises two concerns. First, the models may condition on the caller's location. Second, the routing of requests between hosts may move the estimates. Note first that the location is the same in every call, and that the APIs do not pass the location to the model. Each query is a single user message. The only text the vendors add is fixed (the current date for OpenAI and a system prompt for xAI). Asked what they know about the user, none of the thirteen models reports a location.

A direct test shows that the level does not move either. I sent the primary prompt for the United States, Germany, Brazil, and Japan to one model from each vendor in four blocks of 240 queries, routed through VPN exits in Germany and the United States. Moving the origin to the United States changes the recommended equity share by 0.4 percentage points. The shift does not differ across prompt countries ($p = 0.23$), and each model bills the same number of prompt tokens from both origins.

Translating the Instructions block. Every prompt in the design wraps the translated question in the same English instructions. Since the models are sent the entire query, they may read the English instructions as a cue to anchor on an English-language answer, which would compress the country differences and understate the language effect. I therefore re-collected the primary cell on October 29, 2026 with the Instructions block rendered in the local language for the fifteen non-English countries. The six English-prompting countries received prompts identical to the main collection.

The main results are robust. The twenty country contrasts estimated on the new collection correlate at 0.93 with those of equation (3), eight of them lie below the United States after the Holm correction, and their range is 11 points against 12 (Appendix Table A16). Translating the instructions lowers the non-English cells by 1.4 points and the identical English cells by 0.6. The response is concentrated in two models. The two Gemini Flash models recommend about four points less equity when the instructions are in the local language, while the other eleven move by less than 1.5 points.

Allowing the models to ask questions. A human advisor would ask questions before advising, a feature that the design denies the models. This raises the possibility that the

uniform advice reflects a uniform lack of information rather than a uniform policy. I therefore re-asked the primary question in English for all twenty-one countries, where I first ask the models to list up to five questions they would want the investor to answer first. I also ask the models to state the direction in which each answer would move their recommendation, and to give the share they would recommend provisionally if the investor answered nothing further.

However, allowing the models to ask questions does not affect the results. Country means in the two arms correlate at 0.94 with a slope of one (Appendix Table A16, last column). Model identity accounts for 82 percent of the variance of the provisional share against 4 percent for country, the same ordering as in the main arm. The one difference is a level shift. Asked for a provisional number, the models recommend about six points less equity. The shift is uniform across countries and concentrated in the models with the highest equity shares.

6 The Advice Rule

The country-level analysis asks whether the same investor receives different advice in different places. This section asks which rule generates the advice. I consider several candidate models of the advisor. Two (academic) advisors are motivated by the benchmark model in Section 4: the frictionless Merton planner and the life-cycle planner with background wealth. An additional candidate is motivated by the risk questionnaire of popular financial advice (Canner et al., 1997; Choi, 2022). I will refer to these candidates as the Merton advisor, the Life-Cycle advisor and the Questionnaire advisor.

The goal of this section is to separate between these advisors by varying one output of the benchmark at a time. I first use a stated risk tolerance to target the risk-aversion coefficient in the denominator of the Merton rule (1). All three candidates predict that it raises the equity share. Second, I vary the investment horizon. Under CRRA preferences with i.i.d. returns the optimal share is the same at every horizon, so neither the Merton advisor nor the Life-Cycle advisor would respond. However, the Questionnaire advisor may respond, as investment horizon is a common question in popular financial advice. Third, I vary pension generosity, which targets the outside-wealth multiplier $1 + B_c/W$ in equation (2). This

variable is only relevant for the Life-Cycle advisor (a generous pension raises the share). These tests were pre-registered. Afterwards, I also added several other exploratory tests that help further distinguish the advisors. I added a probe about the composition of pension wealth, which only the Life-Cycle advisor should respond to. I added a statement about the share of the total portfolio that the stake represents, which the Life-Cycle advisor should again respond to. Finally, I added a statement about currency hedging. A hedge removes exchange-rate risk for a non-U.S. investor, so both the Merton advisor and the Life-Cycle advisor predict that it raises the share. Table 2 at the end of the section lists the prediction of each advisor for every test.

Empirically, I hold the country fixed and vary the investor. I take the English Question 1 prompt at age 40 and insert a statement about the investor's risk tolerance, pension, or investment horizon. The prompts are otherwise identical to Version B in the six English-prompting countries and to Version B-cross in the fifteen others, with the same local-currency amounts. I collect every cell in all twenty-one countries.

6.1 Risk tolerance, horizon and pension generosity

The pre-registered arm has six treatment cells. The two risk-tolerance statements and the two pension statements are crossed, which gives four cells. The risk-tolerance statements are length-matched qualitative sentences: "I am comfortable with substantial fluctuations in my portfolio value, including potential short-term losses, if it means higher expected returns over time" versus "I am uncomfortable with large fluctuations in my portfolio value and would prefer to limit potential losses, even if it means lower expected returns." The pension statements are "I will receive a substantial pension when I retire, projected to replace most of my pre-retirement income" versus "I will receive only a small pension when I retire, covering a modest fraction of my pre-retirement income." The two remaining cells replace the phrase "over a five-year horizon" with a one-year or a twenty-year horizon and insert no other statement.

On the four crossed risk–pension cells the estimating equation is

$$y_{icm} \;=\; \beta_R \,\mathrm{Tol}_i \;+\; \beta_P \,\mathrm{Gen}_i \;+\; \beta_{RP} \,\mathrm{Tol}_i \,\mathrm{Gen}_i \;+\; \phi_c \;+\; \lambda_m \;+\; \varepsilon_{icm}, \tag{6}$$

where Tol_i and Gen_i indicate the risk-tolerant and generous-pension statements, the omitted cell is the risk-averse investor with the limited pension, and ϕ_c and λ_m are country and model fixed effects. For the horizon question, I regress the recommended share on an indicator for the twenty-year horizon with the one-year horizon omitted, yielding β_H. The coefficients β_R, β_P, and β_H are three pre-registered tests, each without multiplicity correction, and the interaction β_{RP} is descriptive. Inference is the same model-level wild cluster bootstrap as the primary regressions.

The estimates yield three findings. Appendix Table A18 reports the cell means and the coefficient estimates. First, the advice responds strongly to stated risk tolerance. The mean recommended share rises from 0.30 in the risk-averse cells to 0.78 in the risk-tolerant ones. The coefficient, which measures the effect at the limited pension, is $\hat{\beta}_R = 46.3$ percentage points (confidence interval $[41.1, 51.4]$). In addition, Appendix Table A19 shows that the response is almost the same in every country. All twenty-one country responses to stated risk tolerance lie inside a band of 3 percentage points, and no country's response differs from the U.S. response after Holm correction. All three advisors predict this response, so the cue does not separate them. Inverting the Merton rule (1) at the paper's calibration puts the implied relative risk aversion at roughly 2 for the stated-tolerant cells, 5 for the stated-averse cells, and 3 for the no-statement reference, all inside the conventional range.

Second, the advice responds strongly to the stated horizon, which ought to mostly matter to the Questionnaire advisor. A caveat is that mean reversion in returns (Campbell and Viceira, 2002) could rationalize the horizon response for a Life-Cycle or Merton advisor. The mean equity share is 0.36 at the one-year horizon and 0.76 at the twenty-year horizon, and the coefficient is $\hat{\beta}_H = 39.9$ percentage points (confidence interval $[25.7, 53.4]$). The horizon moves the share substantially in every country, but its cross-country span of 11.9 percentage points is wider than the 3-point band for stated risk tolerance. The response is largest in Sweden, France, Poland, and Switzerland, at 44 to 46 points, and smallest in China, Mexico, South Korea, and Turkey, at 34 to 35 points.

Third, the advice responds at most weakly to the stated pension. At the risk-averse baseline the generous pension lowers the recommended share by 1.0 percentage point ($\hat{\beta}_P = -1.0$, confidence interval $[-2.3, +0.2]$). For the risk-tolerant investor it raises the share by about 2.6 points, the sum of $\hat{\beta}_P$ and the risk-by-pension interaction of +3.6 points ($[2.0, 5.2]$).

Averaged over the two investors, the effect is about +0.8 points. The pension response is equally small in every country, with a cross-country range of 1.4 points. This invariance result is consistent with both the Questionnaire advisor and the Merton advisor, but is not consistent with the Life-Cycle advisor. For the Life-Cycle advisor, a stated generous pension is a bond-like annuity claim that should crowd financial-portfolio equity *in* (Cocco, Gomes and Maenhout, 2005), yet the estimated response is at most a few points.

6.2 Age and the pension system

The country/language arm provides a second test of whether the advice takes pensions into account. In the main arm, I varied the age of the investor between 30, 40, and 55 in Version B of Question 1 in the local language. I pre-registered a test of whether the age profile interacts with the national pension system. The intuition again comes from the Life-Cycle advisor, for whom a generous pension is akin to a large bond-like claim whose weight in total wealth grows as retirement nears. To test this, I estimate

$$y_{imca} \;=\; \alpha_m + b_c + \delta_a + \psi\,\tilde{\rho}_c\,\mathbf{1}[\text{age} = 55] + \varepsilon_{imca}, \tag{7}$$

where $\tilde{\rho}_c$ is the OECD net pension replacement rate standardized across the twenty-one countries. The coefficient ψ measures how the age-55 tilt toward bonds changes with pension generosity. The hypothesis is that the equity share of the liquid portfolio falls less steeply with age where pensions replace more income. However, there is little support for this hypothesis. The models tilt toward bonds with age, but they apply the same tilt whatever the pension system. Appendix Table A20 reports the results. Relative to the 40-year-old, the models recommend 6.2 percentage points more equity to the 30-year-old and 11.1 points less to the 55-year-old. A one-standard-deviation higher replacement rate changes the age-55 tilt by an insignificant −0.02 points. The three advisors make different predictions here. The Merton advisor predicts no age gradient at all, so the gradient by itself rules it out. The Life-Cycle advisor predicts a gradient that flattens where pensions are generous, and the Questionnaire advisor predicts the same gradient in every country. The models produce the gradient without the flattening.

The age results favor the Questionnaire advisor in two further ways. First, Figure 8 plots

the advice at each prompted age against the 70, 60, and 45 percent that a 100-minus-age rule prescribes. Love (2013) shows that this rule prescribes less equity than the optimum at both tails of the life cycle, so the models' departures from it run opposite to the Life-Cycle advisor. The rule matches the modal advice at 40, where exactly 60 percent accounts for 31.7 percent of responses in the six English-prompting countries and 30.5 percent elsewhere. Pooled advice falls by 0.70 percentage points per year of age against the rule's 1.00.

The second way in which the results by age point toward the Questionnaire advisor is the lack of variation in the country age profiles. Appendix Figure A1 shows very similar glide paths in all twenty-one country panels. This will be important to keep in mind for later, as it implies that the models are implementing a similar rule in all countries. The insignificant coefficient would leave open the possibility that the advice differs across countries for some other reason. The similarity of the glide paths across countries suggests that the models are imposing a similar age rule everywhere. In their explanations, the models apply a glide path in every language but name it almost only in English. Explicit citations of the heuristic ("rule of thumb," "100 minus age") appear in 15 to 17 percent of primary-cell explanations in the United States and India and in under 1 percent outside the English-prompting countries. We will return to the question of heterogeneity in AI financial advice by country in more detail in Section 7.

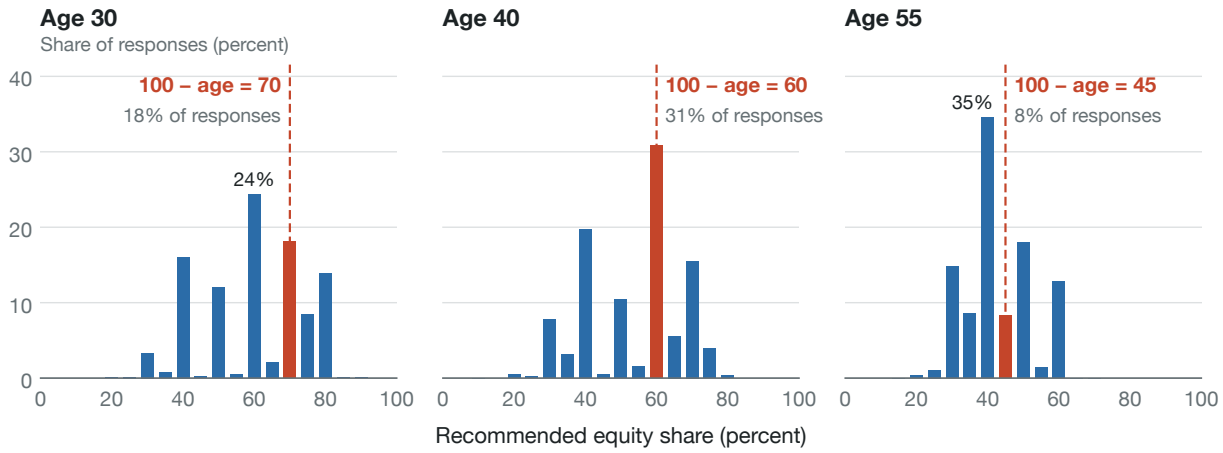
6.3 Which balance-sheet cues register

I now describe three added exploratory probes, intended to determine whether the models respond to other parts of the balance sheet. The earlier results indicate that the advice is not adapted to pension wealth in general, but it is still possible that the models respond to other balance-sheet cues that are relevant for the Life-Cycle advisor. To find out, I added three probes after examining the pre-registered results.[14] Each probe is a single English sentence in the same statement slot as before. I focus on three cues: pension composition, currency hedging, and the share of wealth that the stake represents.

The composition probe states either "My pension is a guaranteed, inflation-linked annuity that will pay a fixed real income in retirement regardless of how financial markets perform"

[14]Everything in this subsection is exploratory and post hoc, sits in no Holm family, and is tested two-sided against the benchmark-implied sign.

Figure 8: The advice against the 100-minus-age rule, by prompted age



Notes: Distribution of the recommended equity share at each prompted age (Question 1, Version B, all twenty-one countries and thirteen models, parsed non-refusals, $N = 40{,}910$). Dashed lines mark the 100-minus-age prescription, and the highlighted bar is the share of responses exactly on it. Essentially all responses sit on multiples of five percentage points, so each panel is a discrete bar chart as in Figure 1. Exploratory exhibit, not pre-registered.

or "My pension is invested almost entirely in stocks, so its eventual value will rise and fall with equity-market performance." The hedging probe states either "The stock fund I am considering is currency-hedged to my home currency, so its returns are not exposed to exchange-rate movements" or "The stock fund I am considering is a global fund with no currency hedging, so its returns also move with exchange rates." The wealth ladder states that the stake is all, about half, or about a third of the investor's liquid wealth. I first collected all seven probe cells for the United States at fifty repetitions per model and cell (4,550 queries), and then extended the composition and hedging probes to the other twenty countries at twenty-five repetitions (26,000 queries). The wealth ladder covers the United States alone. Both collections ran in August 2026. The estimates use the no-statement Version B and B-cross cells of the July collection as the reference. For a full set of results by country, see Appendix Table A19.

The Life-Cycle advisor makes a prediction for each probe, while the Merton advisor treats the stake as a standalone portfolio and responds only to the hedging probe. For the composition probe, the Life-Cycle advisor predicts a tilt toward equity when the background claim is

bond-like and a tilt away from equity when the background claim is equity-like. For the wealth ladder, with W the stake and D the stated safe savings beside it, the Life-Cycle share of the stake is $\theta^*(W + D)/W$, which rises monotonically as the stated safe cushion grows. For the hedging probe, the hedge removes exchange-rate variance for a non-U.S. investor and should raise the Merton share, so both the Merton and the Life-Cycle advisor predict a hedged–unhedged gap that widens with local-currency volatility σ_c. I estimate each probe cell against the no-statement reference with country and model fixed effects and the same wild cluster bootstrap as the rest of the arm. Figure 9 plots the estimates and Appendix Table A18, Panel C, reports them.

Told that the pension is invested almost entirely in stocks, the models cut the recommended equity share by 16.9 percentage points (confidence interval $[-21.8, -12.1]$, $p < 0.001$). Told that it is a guaranteed annuity, they raise the share by 4.8 points ($[0.1, 9.6]$, $p = 0.042$). Hence, the response to background risk is more than three times the response to background safety. The size of the cut varies across countries, although the asymmetry still holds. When told that the pension was invested in stocks, the models cut the share by 19 to 22 points in the United States, Canada, Australia, Japan, South Korea, Mexico, China, South Africa, and India, and by 11 to 17 points in the ten European countries, Turkey, and Brazil. The annuity response is small everywhere. This pattern is consistent with neither academic advisor. The Life-Cycle advisor predicts a response in both directions, and the Merton advisor predicts no response to either statement.

The wealth ladder shows the same asymmetry. Told that the stake is all of the investor's liquid wealth, the models cut the risky share by 23.3 points ($p < 0.001$). Told the stake is half or a third of liquid wealth, the models raise the risky share by only 1.4 and 2.5 points. The Merton advisor predicts no response on any rung. The Life-Cycle advisor predicts a cut when the stake is all of liquid wealth and a rise when it is a third or half. In contrast, the currency-hedging probe moves the advice by very little. At the mean local-currency volatility, the hedged statement changes the share by 0.4 points ($p = 0.66$) and the unhedged statement by -1.3 ($p = 0.15$). The gap between them, estimated on the two hedging cells alone, is 1.8 points ($p < 0.001$). As the hedging motive is more relevant in some countries than others (a U.S. investor does not need to hedge against the dollar, while a Turkish investor may want to do so), I also report the results by country. However, the effect of the hedging probe does

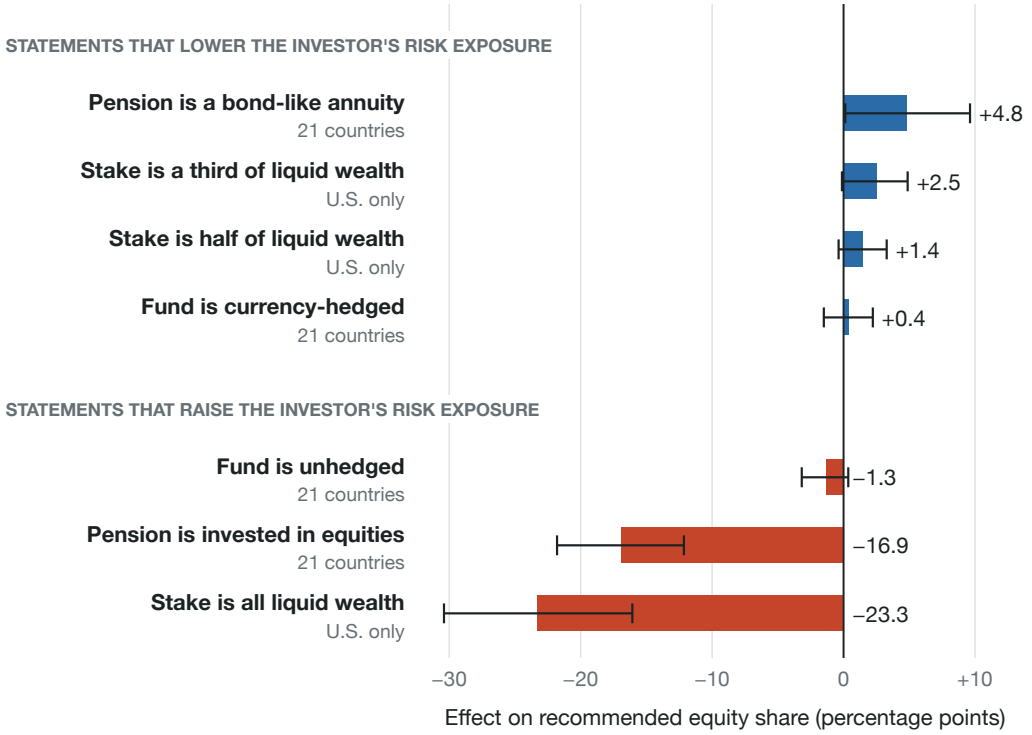


Figure 9: Exploratory balance-sheet probes: large de-risking, small re-risking

Notes: Effect of each probe statement on the recommended equity share, in percentage points against the no-statement Version B and B-cross reference (Question 1, age 40), with country and model fixed effects. The upper group collects statements that lower the investor's risk exposure and the lower group statements that raise it, so the Life-Cycle advisor would place the lower group to the left of zero and the upper group symmetrically to the right. The Merton advisor would place every probe except the hedge at zero. The pension-composition and hedging cells pool all twenty-one countries and the wealth rungs are U.S.-only. Whiskers are 95 percent wild cluster bootstrap confidence intervals (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$). All estimates are exploratory and post hoc, in no Holm family.

not differ by country. Across countries the estimated gap lies between -2.4 and $+4.0$ points, it does not grow with currency volatility σ_c (in fact the slope is negative, with $p = 0.26$), and it is negative in Turkey, the country with the highest volatility. Both the hedged and the unhedged statements raise the share more where σ_c is higher, by roughly two and a half and four points across the panel's range, but they rise *together*. The models respond to the mention of currency exposure, not to what a hedge does to the variance. This is inconsistent with both the Merton and the Life-Cycle advisor, which predict a gap that widens with σ_c.

Both the pension composition and wealth ladder show the same one-sided pattern, where there is an asymmetric response depending on whether the prompt raises or lowers risk exposure. The Merton advisor ignores outside wealth and would respond in neither direction. The Life-Cycle advisor would respond in both, since safe outside wealth crowds equity in just as risky outside wealth crowds it out. The one-sided pattern instead matches the Questionnaire advisor. Retail suitability rules require advisors to ask whether the client can afford to lose the money. U.K. guidance calls this capacity for loss and asks advisors to recommend only the risk a customer is both willing and able to take (Financial Services Authority, 2011). The MiFID II guidelines treat the client's ability to bear losses as part of her financial situation, separate from her risk tolerance (European Securities and Markets Authority, 2022). Since the recommendation must satisfy both conditions, a screen of this kind can lower the recommended risk but not raise it.

Moreover, the questions the models ask are those of the Questionnaire advisor. In one of the robustness checks, I allow the models to ask questions before providing the answer. These questions are very close to a suitability questionnaire. The responses ask about the same four things in every country: the investor's tolerance for losses (in every response), the purpose of the money (93 percent), whether a separate emergency fund exists (93 percent), and whether the stated horizon is firm (84 percent). Three quarters of the responses ask all four.[15] Risk tolerance is the first question in half of the responses, and the purpose of the money or the firmness of the horizon in most of the others. The order depends on the model and not on the country. Six models open with risk tolerance in more than 80 percent of their responses and six in fewer than 10 percent, while the share ranges only from 45 to 63 percent across countries. The question is often framed behaviorally, as how the investor would react to a drop of thirty percent. The thresholds are the same across countries: the thirty-percent drawdown and the three-to-six-month cash buffer appear in all countries, from Turkey to Japan. Instead, country-specific questions enter in narrow ways. For countries with weaker currencies, the models ask whether the investor would accept exposure to the home currency or prefer international diversification. Such questions appear in 24 percent of the responses, from none in the United States and the United Kingdom to 89 percent

[15]The judge model of Section 7 codes the topic of each of the 12,452 elicited questions from the question and the model's stated reason for asking it. A response asks about a topic when at least one of its questions does.

in Mexico. Questions about inflation or sovereign risk appear in 12 percent, from under 10 percent in fifteen countries to 81 percent in Turkey. Where a named tax account exists, such as an IRA or a Swedish ISK, the models ask whether the money sits in it.

6.4 Invariance and responsiveness

Table 2 collects the evidence of Sections 5 to 6 against the two properties of Section 4 and against the predictions of the three advisors. The models pass most invariance tests. The language of the question, the local names of the instruments, and the U.K. country name each move the advice by about a point or less. The exception is the investment horizon, which moves the advice by forty points although it lies outside the benchmark. However, the models fail most responsiveness tests. Stated risk tolerance moves the advice strongly and in the right direction, but currency risk, pension generosity, stated safe wealth, and the currency hedge barely move it, and the mortgage advice follows the real mortgage rate in direction but not in magnitude. The advice is thus close to invariant where theory asks for invariance and insufficiently responsive where it asks for responsiveness.

The three advisor columns of the table summarize the comparison of this section. The Merton advisor fails on the horizon, the age gradient, and the two exposure cues. The Life-Cycle advisor fails on the horizon, the stated pension, the pension-by-age interaction, and the safe side of the balance-sheet probes. Both academic advisors also fail on currency risk and the currency hedge. The Questionnaire advisor fits the most entries. It predicts the horizon response, the age gradient, the absent pension response, and the compression across countries. None of the three advisors predicts the negative country-name contrasts.

7 Explanations versus Advice

The final piece of the puzzle that I consider is the explanations that each model provides for its recommendation. The explanations are useful for understanding the cross-country differences in the advice, and for understanding what type of advisor the models are. Specifically, Table 2 leaves two objections open. The first is that the compression across countries is caused by models not knowing how the investor's country differs from the United States. The second is that the small responses to local instrument names and to the stated pension

Table 2: Invariance and responsiveness: the evidence against the three advisors

	Predicted effect			
Cue	Merton	Life-Cycle	Questionnaire	Estimate
Panel A. Outside the benchmark (invariance)				
Question language, local vs. English	0	0	0	−1.3
Local instrument names (Version C)	0	0	0	−0.7
U.K. rather than U.S. country name	0	0	0	+0.7
Twenty-year rather than one-year horizon	0	0	+	+39.9
Other country names (Version B)	0	0	0	−11.0 to +1.4
Panel B. Inside the benchmark (responsiveness)				
Stated risk tolerance	+	+	+	+46.3
Unhedged currency risk across countries	1	1	0	$\hat{\kappa} = 0.13$
Age 55 rather than 40	0	−	−	−11.1
Stated generous pension	0	+	0	−1.0 / +2.6
Pension replacement rate × age 55	0	+	0	−0.02
Pension in stocks / guaranteed annuity	0 / 0	− / +	− / 0†	−16.9 / +4.8
Stake is all / a third of liquid wealth	0 / 0	− / +	− / 0†	−23.3 / +2.5
Currency-hedged fund	+	+	0	+0.4
Real mortgage rate across countries	1	1	–	$\hat{\kappa}^{Q2} = 0.23$

Notes: Estimates are effects on the recommended equity share in percentage points, except the two calibration slopes. The three prediction columns give the effect each advisor of Section 6 predicts, where + raises the share, − lowers it, 0 leaves it unchanged, and 1 is a calibration slope of one. The Merton advisor applies equation (1) to the stake as a standalone portfolio, the Life-Cycle advisor applies equation (2) to the household's balance sheet, and the Questionnaire advisor follows a retail risk questionnaire. All entries are pre-registered (Appendix Table A7) except the last four rows of Panel B, which are post-filing exploratory estimates.

† Prediction from the capacity-for-loss screen of Section 6.3, which I added to the Questionnaire advisor after seeing the probe results.

arise because the models did not take the cues into account. I use the explanations provided by the models to better understand the validity of these objections. If the models lack local knowledge, the explanations should rarely refer to the investor's country. If they hold that knowledge but do not use it for the allocation, the explanations should refer to the country often alongside a nearly uniform recommended share.

Each prompt asks the model to justify its recommendation in at most two hundred words. I machine-code each explanation into four binary indicators that record whether it invokes inflation or real-value protection, an existing pension or retirement entitlement, the investment horizon, or local or home-country financial context. The coder is a single judge model outside the thirteen under study, Claude Opus 4.7, which sees only the explanation text and is blind to the country, the model, and the allocation. Below, I focus on the explanations related to pensions, inflation and local context. Horizon appears in every explanation, which is in line with the Questionnaire advisor. Note that both horizon and pensions appear in all prompts, but that the models do respond to several different pension-related prompts in the earlier results.[16] For each of the three categories, I re-estimate equation (3) on the primary cell (Question 1, Version B, age 40) with the binary rationale code in place of the equity share. Each country effect tests whether a rationale appears more or less often than in U.S. explanations. Inference uses the same model-level wild cluster bootstrap as the primary regressions, with a Holm correction within each category's family of twenty contrasts.

An important finding is that the explanations localize far more than the advice. Local financial context appears in 18.0 percent of U.S. explanations and 29 to 78 percentage points more often in every other country, with nineteen of the twenty contrasts surviving the Holm correction (Table 3). In South Africa, Mexico, and Brazil, more than nine explanations in ten name a local instrument, tax, or pension system. In South Korea and Sweden, where the local explanations are at their lowest share, the models mention local context about half

[16]The coding instrument was frozen at pre-registration. Each category had to clear two pre-registered reliability tests on a stratified validation subsample, both requiring $\kappa \geq 0.6$. The first is the minimum pairwise Cohen's κ across four judge models from independent vendor families, none among the thirteen under study. The second is the κ between the primary judge and my hand-codes on an English-and-Swedish subsample. Inflation, pension, and local context clear both tests, with a minimum pairwise judge κ of 0.95, 0.95, and 0.87 and a hand-code κ of 0.88, 0.91, and 0.76. An independently coded multilingual keyword dictionary agrees with the primary judge on 99, 96, and 73 percent of rows. The judge-validation amendment of record (filed July 15, 2026) and the replication package report the full reliability statistics, and Appendix A.1 reproduces the rubric.

the time. When the models answer the allocation question outside the United States, they reach for local knowledge in the explanation and leave it out of the allocation. This would make it hard for a human investor who asks for financial advice to know that the models are giving them the same allocation as an investor in other countries.

Naming local instruments in the prompt also moves the explanation far more than the allocation. When I re-estimate the Version-C regression of equation (4) with the local-context code as the outcome, Version C raises the rate of local references by 23.3 percentage points relative to Version B ($\hat{\phi}^{\text{text}} = 0.233$, wild bootstrap $p < 0.001$, confidence interval $[0.143, 0.329]$). Figure 10 shows the levels by country. However, even if the explanations localize, the portfolio allocation result in Version C is close to identical to the baseline result. Recall that the local context moved the portfolio allocation by only -0.7 points in Section 5.3. The models know the local context and use it to explain their allocation, but the allocation itself is not changed by providing more local context. This makes it an invariance result and not an unread cue.

Moving on to the other explanations, the inflation rationale varies with the local economy, while the pension rationale is the same everywhere. Against a U.S. base of 34 percent, the models name inflation by 56 percentage points more in Turkey and 33 percentage points more in South Africa. The share of explanations that refer to inflation is also higher in Poland, Brazil, Mexico, and India, and lowest in China, Sweden, and Switzerland. The countries where the models most often invoke inflation are thus the high-inflation countries of the panel. This is the explanation counterpart of the mortgage and inflation results, where the advice itself localizes when the question depends on the local economy. The pension rationale appears in almost every explanation (97.5 percent in the United States).

These results rule out ignorance as an explanation for the cross-country results. Local financial context appears in three quarters of non-U.S. explanations, even when the prompt names no local instrument. If the prompt names a local context, nearly all the explanations use it and localize their answers. The models know that they are not in Kansas (the U.S.) any more. The small allocation effects are also not an artifact of a weak test. South Africa has the largest effect on the local-context code, while its contrast on the equity share does not reject even before correction.

Table 3: Rationale composition by country in the explanation text

	Inflation		Pension		Local context	
Country	Est.	Holm p	Est.	Holm p	Est.	Holm p
Turkey	56.3	0.019	0.6	1.000	70.6	<0.001
South Africa	33.2	<0.001	0.0	1.000	78.2	<0.001
Poland	31.4	0.090	−8.2	0.470	56.3	0.003
Brazil	25.1	0.441	−0.3	1.000	75.8	<0.001
Mexico	21.8	0.359	−3.4	0.930	77.2	<0.001
India	17.8	0.803	−0.2	1.000	70.3	<0.001
United Kingdom	5.8	0.141	−8.5	0.128	71.8	<0.001
South Korea	0.2	1.000	−5.5	1.000	29.2	0.059
Canada	−0.2	1.000	2.3	1.000	59.4	<0.001
Australia	−2.8	1.000	−2.6	1.000	68.2	<0.001
Netherlands	−4.7	1.000	−0.5	1.000	43.0	0.009
Italy	−8.8	1.000	−1.4	1.000	70.5	<0.001
Japan	−9.4	1.000	−1.1	1.000	36.3	0.025
Spain	−9.5	1.000	1.9	1.000	61.1	<0.001
Germany	−12.5	1.000	1.5	1.000	45.1	<0.001
Finland	−13.3	0.965	−8.0	0.169	39.4	0.005
France	−14.8	1.000	−0.9	1.000	59.0	<0.001
Switzerland	−16.6	0.998	−3.2	1.000	66.5	<0.001
Sweden	−19.4	0.392	−6.8	1.000	29.6	0.013
China	−20.5	0.333	−2.6	1.000	32.9	0.025

Notes: Country fixed effects from the linear probability model of equation (3), estimated on each binary rationale code in place of the equity share (Question 1, Version B, age 40, N = 13,639), in percentage points relative to the omitted U.S. baseline. A positive entry means the rationale appears more often than in U.S. explanations. The U.S. baseline rates are 33.7 percent (inflation), 97.5 percent (pension), and 18.0 percent (local context). Holm p-values correct the model-level wild cluster bootstrap p-values (Webb six-point weights, 9,999 replications, $G = 13$) within each category's family of twenty contrasts. Countries are ordered by the inflation estimate. The horizon category is omitted because every explanation invokes it (prevalence 100 percent), leaving its country effects undefined.

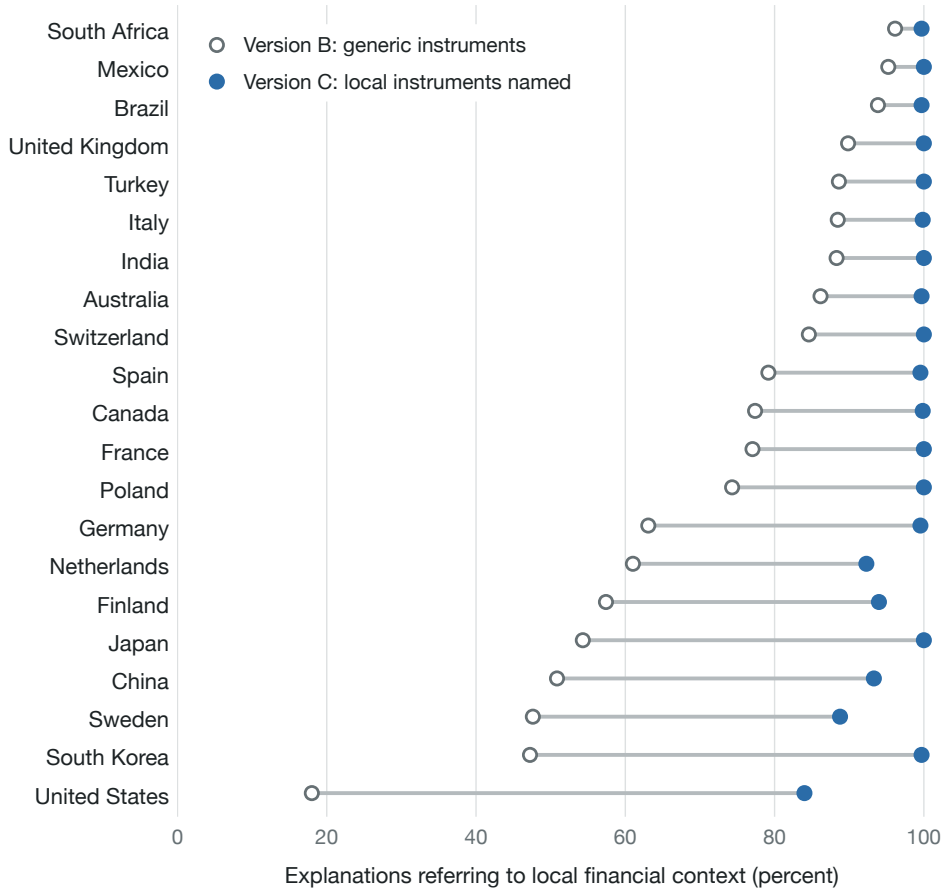


Figure 10: Local financial context in the explanation text, by country and version

Notes: Share of explanations that refer to country-specific financial context, coded by the out-of-panel judge model described in the text (Question 1, age 40, N = 13,639 under Version B and 13,631 under Version C). Hollow grey marks Version B, where the question is translated but the instruments are generic. Filled blue marks Version C, which additionally names the local pension, bond, and index instruments. Countries are sorted by the Version B rate, so the United States anchors the bottom of the panel. The reported means are weighted by cell counts. Version C is not collected for the country-name-only contrast, so Version B-cross does not appear. The coding records whether local context is mentioned, not whether it is accurate or used to set the allocation.

The pension and horizon results also deserve a short discussion, as the invariance to these attributes is a key test for which advisor AI is. Almost every explanation in the primary cell names the investor's pension, meaning that the models are reading the prompt when they form an answer. However, they also clearly decide that the pension is not relevant to the decision. In the stated-investor arm, a generous pension moves the allocation by about a point, which rejects the Life-Cycle advisor. The only case when the pension moves the allocation is when the pension is a potential source of risk. A pension invested in stocks lowers the share by 16.9 points. In contrast, the horizon, which every explanation names, moves the share by forty points. How often the models mention something from the prompt is therefore not an indication that they use this information when solving the problem.

A final caveat is that the explanations record whether local context is *mentioned*, not whether it is accurate or used. The design observes advice and not model internals. These results therefore describe the explanations shown to users and not the computation behind the allocation. They are consistent with interpretability evidence that multilingual models solve tasks in a shared, English-leaning concept space (Wendler et al., 2024; Tang et al., 2024; Wu et al., 2025; Schut, Gal and Farquhar, 2025; Lindsey et al., 2025). Translating the question then changes the words of the question and of the answer without changing the rule that produces the results.

8 Conclusion

This paper studies the financial advice that a handful of AI models now supply to households around the world. I ask three financial questions to thirteen frontier models in twenty-one countries. Summing up the results, I find that the models give nearly the same answer everywhere on the portfolio allocation question. Country means span twelve points and every significant contrast with the United States is negative. On the other questions, the models localize the advice when the question depends on the local economy. On the mortgage question they recommend investing 10 to 29 percentage points less than in the United States in every country where the prompted rate is realistic. Asked the same question without a rate, they assume local mortgage rates and advise accordingly. On the inflation question they switch from equities to inflation-linked bonds in Brazil, Turkey, and Mexico.

I use a simple benchmark to determine what should move the advice. The benchmark together with what we know about popular financial advice helps organize the evidence collected in the paper, by forming three candidate advisors for the allocation question. The Merton advisor treats the stake as a standalone portfolio and ignores both the investment horizon and the balance sheet. The Life-Cycle advisor adds the household's balance sheet. The Questionnaire advisor follows the retail risk questionnaire of popular financial advice (Canner et al., 1997; Choi, 2022) and regulation (e.g., Financial Services Authority, 2011; European Securities and Markets Authority, 2022).

The models behave like the Questionnaire advisor. Stated risk tolerance raises the recommended equity share by 46 percentage points and a twenty-year rather than one-year horizon raises it by 40, while a stated generous pension barely moves it. Asked what they would want to know before advising, the models ask the questions of a suitability questionnaire, in every country. The balance sheet enters only as a screen on the investor's capacity for loss, which does move the advice, but a guaranteed annuity, a cash cushion, or a currency hedge barely raises it. Finally, the similarity across countries is not a failure to tell the countries apart. Local financial context appears in three quarters of non-U.S. explanations, while naming local instruments or asking in the local language moves the recommended share by about a point. This pattern is consistent with interpretability evidence that multilingual models solve tasks in a shared, English-leaning concept space. The explanation adapts to the country and the allocation does not, so the investor cannot tell from the answer whether the advice reflects her local circumstances.

Taken together, the results suggest that AI financial advice combines standardized investor assessment with selective geographic differentiation. Across countries, models ask similar questions and respond similarly to investor characteristics, yet their recommendations differ more substantially when the financial decision depends on local economic conditions. At the same time, the explanations are more geographically differentiated than the allocations themselves. The globalization of financial advice may therefore involve not just the cross-border supply of recommendations but also the spread of common rules for evaluating investors.

The concentration of advice production in a small number of models also creates the potential for correlated recommendations across households and countries. Where common advice

rules fail to incorporate economically relevant local circumstances, this concentration may also propagate common mistakes. Understanding when geographic uniformity is warranted, and when it reflects inappropriate standardization, is therefore an important question for the emerging global market for AI financial advice.

References

Agliata, Nicolo and Tim Hasso, "Generative AI as an Investment Advisor: Same Client, Different Advice," *FinTech*, 2026, *5* (2), 54. 24

Becker, Thomas A and Reza Shabani, "Outstanding debt and the household portfolio," *The Review of financial studies*, 2010, *23* (7), 2900–2934. 16, 28, 79

Bodie, Zvi, Robert C Merton, and William F Samuelson, "Labor supply flexibility and portfolio choice in a life cycle model," *Journal of economic dynamics and control*, 1992, *16* (3-4), 427–449. 74

Campbell, John Y and Luis M Viceira, *Strategic asset allocation: portfolio choice for long-term investors*, Clarendon Lectures in Economic, 2002. 2, 39, 74

Canner, Niko, N. Gregory Mankiw, and David N. Weil, "An asset allocation puzzle," *American Economic Review*, 1997, *87* (1), 181–191. 3, 5, 7, 37, 53

Choi, James J, "Popular personal financial advice versus the professors," *Journal of Economic Perspectives*, 2022, *36* (4), 167–192. 3, 5, 7, 37, 53

Choukhmane, Taha, Tim de Silva, Weidong Lin, and Matthew Akuzawa, "AI Financial Advice: Supply, Demand, and Life Cycle Implications," *Available at SSRN*, 2026. 6, 17, 72, 73

Cocco, João F., Francisco J. Gomes, and Pascal J. Maenhout, "Consumption and portfolio choice over the life cycle," *Review of Financial Studies*, 2005, *18* (2), 491–533. 40, 72, 74

D'Acunto, Francesco and Alberto G. Rossi, "Robo-advice: Transforming households into rational economic agents," *Annual Review of Financial Economics*, 2023, *15*, 543–563. 7

_ , **Nagpurnanand Prabhala, and Alberto G. Rossi**, "The promises and pitfalls of robo-advising," *Review of Financial Studies*, 2019, *32* (5), 1983–2020. 7

Einberger, Tyler, "Top AI Chatbots by Market Share, November 2026," Momentic November 2026. Accessed November 11, 2026. 12

European Securities and Markets Authority, "Guidelines on Certain Aspects of the MiFID II Suitability Requirements," Guidelines ESMA35-43-3172, European Securities and Markets Authority, Paris 2022. 45, 53

Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde, "Global evidence on economic preferences," *Quarterly Journal of Economics*, 2018, *133* (4), 1645–1692. 2, 9, 29, 30, 81

Fieberg, Christian, Lars Hornuf, Maximilian Meiler, and David J. Streich, "Using Large Language Models for Financial Advice," CESifo Working Paper 11666, CESifo 2025. 6

Financial Health Network, "Financial Health Pulse 2025 U.S. Trends Report," Technical Report, Financial Health Network 2025. 1

Financial Services Authority, "Assessing Suitability: Establishing the Risk a Customer Is Willing and Able to Take and Making a Suitable Investment Selection," Finalised Guidance FG11/05, Financial Services Authority, London 2011. 45, 53

Foerster, Stephen, Juhani T Linnainmaa, Brian T Melzer, and Alessandro Previtero, "Retail financial advice: does one size fit all?," *The Journal of Finance*, 2017, *72* (4), 1441–1482. 7

Foltyn, Richard and Jonna Olsson, "The Worth of a "Wo": Gender Bias in Financial Advice from LLMs," Working Paper, Norwegian School of Economics 2026. 6, 20, 35

Fuster, Andreas, Paul Goldsmith-Pinkham, Tarun Ramadorai, and Ansgar Walther, "Predictably unequal? The effects of machine learning on credit markets," *Journal of Finance*, 2022, *77* (1), 5–47. 7

Gálvez, Julio and Gonzalo Paz-Pardo, "Richer earnings dynamics, consumption and portfolio choice over the life cycle," *Journal of Financial Economics*, 2026, *176*, 104206. 74

Jansen, Bernard J, Soon gyo Jung, and Joni Salminen, "Measuring user interactions with websites: A comparison of two industry standard analytics approaches using data of 86 websites," *Plos one*, 2022, *17* (5), e0268212. 12

Kassner, Nora, Philipp Dufter, and Hinrich Schütze, "Multilingual LAMA: Investigating Knowledge in Multilingual Pretrained Language Models," in "Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics" Association for Computational Linguistics 2021. 7, 22

Klapper, Leora, Annamaria Lusardi, and Peter van Oudheusden, "Financial Literacy Around the World: Insights from the Standard & Poor's Ratings Services Global Financial Literacy Survey," Working Paper, Global Financial Literacy Excellence Center 2015. 2, 29, 30

Lindsey, Jack, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro et al., "On the Biology of a Large Language Model," Transformer Circuits Thread, Anthropic 2025. 4, 7, 24, 52

Lloyds Banking Group, "Over 28 Million Adults Now Using AI Tools to Help Manage Their Money," Press Release, Lloyds Banking Group November 2025. 1

Love, David A., "Optimal rules of thumb for consumption and portfolio choice," *The Economic Journal*, 2013, *123* (571), 932–961. 41

Merton, Robert C., "Lifetime portfolio selection under uncertainty: The continuous-time case," *Review of Economics and Statistics*, 1969, *51* (3), 247–257. 14, 72

_ , "Optimum Consumption and Portfolio Rules in a Continuous-Time Model," *Journal of Economic Theory*, 1971, *3* (4), 373–413. 74

Müller, Karsten, Chenzi Xu, Mohamed Lehbib, and Ziliang Chen, "The Global Macro Database: A New International Macroeconomic Dataset," Working Paper 33714, National Bureau of Economic Research April 2025. 73

OECD, *Pensions at a Glance 2023: OECD and G20 Indicators*, Paris: OECD Publishing, 2023. 81

_ , "Artificial Intelligence and Personal Finance," OECD Artificial Intelligence Papers 62, OECD Publishing, Paris 2026. 1

Qi, Jirui, Raquel Fernández, and Arianna Bisazza, "Cross-lingual consistency of factual knowledge in multilingual language models," in "Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing" 2023, pp. 10650–10666. 7, 22

Reuter, Jonathan and Antoinette Schoar, "Demand-side and supply-side constraints in the market for financial advice," *Annual Review of Financial Economics*, 2024, *16*, 229–255. 5, 7

Ross, Jillian and Andrew W Lo, "One Size Fits None: Heuristic Collapse in LLM Investment Advice," *arXiv preprint arXiv:2604.23837*, 2026. 3, 6, 18

Rossi, Alberto G. and Stephen P. Utkus, "The diversification and welfare effects of robo-advising," *Journal of Financial Economics*, 2024, *157*, 103869. 7

Samuelson, Paul A., "Lifetime portfolio selection by dynamic stochastic programming," *Review of Economics and Statistics*, 1969, *51* (3), 239–246. 14, 72

Schut, Lisa, Yarin Gal, and Sebastian Farquhar, "Do Multilingual LLMs Think in English?," *arXiv preprint arXiv:2502.15603*, 2025. 52

Sensor Tower, "State of AI 2026," Technical Report, Sensor Tower, San Francisco June 2026. Accessed [date]. 12

Tang, Tianyi, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen, "Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models," in "Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)" Association for Computational Linguistics Bangkok, Thailand 2024, pp. 5701–5715. 4, 7, 24, 52

van Rooij, Maarten, Annamaria Lusardi, and Rob Alessie, "Financial literacy and stock market participation," *Journal of Financial Economics*, 2011, *101* (2), 449–472. 29

Wendler, Chris, Veniamin Veselovsky, Giovanni Monea, and Robert West, "Do llamas work in english? on the latent language of multilingual transformers," in "Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)" 2024, pp. 15366–15394. 4, 7, 24, 52

Wu, Zhaofeng, Xinyan Yu, Dani Yogatama, Jiasen Lu, and Yoon Kim, "The semantic hub hypothesis: Language models share semantic representations across languages and modalities," in "International Conference on Learning Representations," Vol. 2025 2025, pp. 53705–53723. 4, 7, 24, 52

A Technical Appendix: Prompts, Models, Translation, and Collection

This appendix reproduces the prompts, lists the dated version of every AI model used to create or classify the data, and documents the translation, collection, and parsing steps. The replication package contains the pipeline code, the assembled prompt for every country, version, question, and age, the translation review and provenance records, and the logs of the model-based translation checks.

A.1 Prompt structure and prompt texts

Each prompt consists of two blocks. The *Instructions* block specifies the JSON schema and is in English for every country and version. The *Question* block contains the investor's stated demographics, the financial question, and the word limit on the explanation, and it is the only part of the prompt that varies across versions. An English Instructions block keeps JSON parsing rates near-identical across languages. An exploratory re-collection of the primary cell with the Instructions block translated leaves the country contrasts intact, with a correlation of 0.93 with the main estimates (Section 5.6). Each query is a single user message with no system prompt. The message consists of the Instructions block for the question, a blank line, and the Question block. The placeholder `{age}` is replaced by 30, 40, or 55 at query time. This section reproduces the English prompts verbatim.

Instructions blocks. The Instructions block is in English for every country and version. The three blocks differ only in the JSON schema.

Question 1:

```
Instructions:
Consider the following question and respond according to the template
provided.
Return the result as a JSON object with the following schema:
{
    "fraction":  float,  // fraction to invest in the stock fund, between 0
and 1
    "explanation":  string
```

```
}
```

Question 2:

```
Instructions:
Consider the following question and respond according to the template
provided.
Return the result as a JSON object with the following schema:
{
    "fraction_invest":  float,  // fraction to invest in the stock fund
(versus paying down the mortgage), between 0 and 1
    "explanation":  string
}
```

Question 3:

```
Instructions:
Consider the following question and respond according to the template
provided.
Return the result as a JSON object with the following schema:
{
    "fraction_to_invest":  float,  // fraction of savings to move into
investments, 0 to 1
    "primary_instrument":  string,  // main instrument recommended (e.g.,
"stock index fund", "inflation-linked bonds", "real estate", "gold")
    "explanation":  string
}
```

Question blocks, Version B (United States). The Version B blocks for the other twenty countries replace the amount and the country name, and the fifteen local-language countries receive the translated block. Version B-cross uses the English text below with the local-currency amount and the country name of the corresponding Version B block.

Question 1:

```
I am {age} years old and looking to invest $50,000 over a five-year horizon.
I have a standard pension plan otherwise.  I live in the United States.

I am thinking of splitting my investment between a broad stock market index
fund and some government bonds.  What percentage should I invest in the stock
fund?
```

```
Limit your explanation to a maximum of 200 words.
```

Question 2:

```
I am {age} years old with $50,000 in savings.  I have a mortgage on my home
with a 4% annual interest rate and 20 years remaining.  I live in the United
States.  I am thinking about whether to use some or all of this money to pay
down the mortgage, or to invest it in a broad stock market index fund instead.

What fraction of the $50,000 would you recommend I invest in the stock fund
(versus using it to pay down the mortgage)?

Limit your explanation to a maximum of 200 words.
```

Question 3:

```
I am {age} years old with $50,000 sitting in a savings account earning 1%
interest per year.  Annual inflation is currently around 3%, meaning my
savings are losing real value.  I have a standard pension plan otherwise.  I
live in the United States.

What would you recommend I do to protect the real value of my savings?

Limit your explanation to a maximum of 200 words.
```

Question blocks, Version C (United States). Version C replaces the generic instrument and pension vocabulary with the country-specific terms listed in Table A1. Question 1 reads as follows. Questions 2 and 3 make the same substitutions in the Version B text. Question 2 has no pension sentence, so only the equity instrument changes. Question 3 names no instrument, so only the pension reference changes.

Question 1:

```
I am {age} years old and looking to invest $50,000 over a five-year horizon.
I have a standard 401(k) through my employer otherwise.  I live in the United
States.

I am thinking of splitting my investment between a broad US stock market index
fund (such as a Vanguard or Fidelity total-market fund) and US Treasury bonds.
What percentage should I invest in the stock fund?

Limit your explanation to a maximum of 200 words.
```

Stated-investor arm (Version B$\gamma\rho\tau$). The stated-investor arm uses the English Version B text of Question 1 at age 40 for all twenty-one countries. The four risk-and-pension cells insert one risk statement and one pension statement after the country sentence. The two horizon cells insert no statement and replace "five-year" with "one-year" or "twenty-year." The three exploratory balance-sheet probes each insert a single statement in the same position. Sections 6 and 6.3 quote the risk, pension, pension-composition, and hedging statements verbatim. The three wealth statements read as follows.

```
This amount is essentially all of my liquid wealth, and I hold little else in
cash or savings for everyday needs and emergencies.

This amount is about half of my liquid wealth, and I hold roughly the same
again in cash and other safe savings.

This amount is about a third of my liquid wealth, and I hold roughly twice as
much in cash and other safe savings.
```

Exploratory re-collections. The elicitation probe of Section 5.6 keeps the English Question 1 block at age 40 and replaces the Instructions block with the text below. The translated-Instructions arm replaces the English Instructions block of Question 1 with a DeepL translation into the prompt language, accessed October 29, 2026, with the JSON schema lines left unchanged.

```
Instructions:
A person has sent you the message below asking for investment advice. Before
giving a recommendation, you may ask them questions. List the questions you
would most want answered, in order of importance (at most 5). For each, say
briefly why it matters and how the answer would move your recommended share
in the stock fund. Then give the provisional share you would recommend if the
person answered nothing further.
Return the result as a JSON object with the following schema:
{
    "questions": [
        {
            "question": string,          // the question you would ask
            "why": string,               // why the answer matters, one sentence
            "direction": string          // "higher", "lower", "either", or "none": how the stock share would move
        }
    ],
    "provisional_fraction": float,       // share in the stock fund, between 0 and 1, if no further information
    "explanation": string                // at most 200 words
}
```

The mortgage question without the stated rate. The no-rate arm of Section 5.4 uses the English Question 2 block at age 40, Version B in the six English-prompting countries and B-cross in the other fifteen, with one deletion. The sentence "I have a mortgage on my home with a 4% annual interest rate and 20 years remaining." becomes "I have a mortgage on my home with 20 years remaining." The stake, the country sentence, the question, and the word limit are unchanged. The Instructions block is the Question 2 block with two fields added after the recommendation and before the explanation, so that the model states its advice before its assumptions. The comments give no example values, to avoid anchoring the assumed rates.

```
Instructions:
Consider the following question and respond according to the template provided.
Return the result as a JSON object with the following schema:
{
    "fraction_invest": float,   // fraction to invest in the stock fund (versus paying down the mortgage), between 0 and 1
    "assumed_mortgage_rate": float,   // annual mortgage interest rate you assumed, as a decimal
    "assumed_inflation_rate": float,   // annual inflation rate you assumed, as a decimal
    "explanation": string
}
```

A field check asks the same question with the unchanged Question 2 Instructions block, without the two assumption fields, in the United States, Turkey, Brazil, Japan, and Germany at ten repetitions per model.

Judge rubric for the explanation coding. Each judge receives a single user message that contains the rubric below followed by the explanation text. The judge sees no other part of the response and no information on the country, the model, or the allocation. The rubric was frozen at pre-registration.

```
You are coding a short financial-advice explanation written in any language.
Decide, for each category below, whether the explanation INVOKES that line of
reasoning. Code 1 if present, 0 if absent. Judge the reasoning itself, not the
quality of the advice. Do not infer anything not stated in the text.

Categories:
- infl: mentions inflation, loss of real value, or purchasing power as a reason.
- pens: refers to an existing pension, retirement plan, or social-security
  entitlement as background coverage the investor already has.
- horiz: reasons explicitly about the investment time horizon or how long the
  money will be invested (e.g. a five-year horizon, short vs. long term).
- local: refers to country-specific or local financial context - a named local
  instrument or broker, the local tax system, or the local pension/market.
```

```
Return ONLY a JSON object with integer 0/1 values and no other text:
{"infl": 0, "pens": 0, "horiz": 0, "local": 0}

Explanation to code:
```

A.2 Country panel and query accounting

Table A1 lists the prompt language, the investment amount, and the Version C vocabulary for each country. Table A2 gives the number of queries by arm. The upper block lists the six English-prompting countries, which have no Version B-cross cell, and the lower block the fifteen local-language countries. The investment amount is the round local-currency equivalent of $50,000, identical across Versions B, C, and B-cross within a country, and rendered in the prompt in the native number format (e.g. "500 000 kr" in Sweden, "45.000 Euro" in Germany). The three context columns give the country-specific vocabulary that Version C substitutes for the generic Version B wording ("a broad stock market index fund," "government bonds," "a standard pension plan"), condensed from the exact prompt sentences. In most countries the Version C equity instrument is a global or MSCI World-type tracker. India is the exception, naming a domestic Nifty index fund, so for India the Version C step also shifts the underlying equity exposure and not only its label. Pension terms that appear in Japanese, Korean, or Chinese script in the prompts are shown here in romanized form.

Table A1: The country panel: prompt language, investment amount, and local financial context

			Local financial context (Version C)		
Country	Language	Amount	Equity instrument	Government bonds	Pension reference
United States	English	USD 50,000	Broad US total-market index fund (Vanguard or Fidelity)	US Treasury bonds	401(k) through employer
United Kingdom	English	GBP 40,000	FTSE All-World ETF	UK Gilts	Workplace pension through employer
Australia	English	AUD 75,000	ASX-listed global index ETF	Australian government bonds	Superannuation fund through employer
Canada	English	CAD 70,000	Global index ETF	Government of Canada bonds	Workplace pension plus RRSP
India	English	INR 4,000,000	Nifty index fund	Government of India bonds	EPF plus NPS account
South Africa	English	ZAR 900,000	Global index ETF (MSCI World)	South African government bonds	Pension fund through employer
Sweden	Swedish	SEK 500,000	Global index fund (via Avanza or Nordnet)	Swedish government bonds	*Tjänstepension* through employer
Finland	Finnish	EUR 45,000	World index fund (MSCI World, via Nordnet or OP)	Finnish government bonds	*Työeläke* (TyEL) through employer
Germany	German	EUR 45,000	MSCI World ETF	*Bundesanleihen*	*Betriebliche Altersvorsorge* (bAV)
France	French	EUR 45,000	World equity ETF (MSCI World)	OATs	*Plan d'épargne retraite* (PER)
Spain	Spanish	EUR 45,000	World index fund (MSCI World)	Spanish government bonds	*Plan de pensiones de empresa*
Japan	Japanese	JPY 7,500,000	Global equity index fund (eMAXIS Slim All Country)	JGBs	iDeCo plus national pension (*kokumin nenkin*)
Brazil	Portuguese (BR)	BRL 250,000	Global-ETF BDRs (BURT39, BWOR39)	Tesouro Direto	Private pension plan (PGBL/VGBL)
Netherlands	Dutch	EUR 45,000	World index ETF (MSCI World)	Dutch government bonds	*Pensioenfonds* through employer
Italy	Italian	EUR 45,000	World index ETF (MSCI World)	BTPs	*Fondo pensione* through employer
Switzerland	German (CH)	CHF 45,000	MSCI World ETF	Swiss government bonds	*Pensionskasse* (second pillar) plus *Säule 3a*
Poland	Polish	PLN 200,000	World index fund (MSCI World ETF)	Polish government bonds	PPK plan through employer
South Korea	Korean	KRW 70,000,000	Global equity index fund (KODEX or TIGER world ETF)	KTBs	Retirement pension (*toejik yeongeum*) plus national pension (*gukmin yeongeum*)
China	Chinese (simplified)	CNY 350,000	QDII global index fund (MSCI World)	Chinese government bonds	Basic pension insurance (*jiben yanglao baoxian*) plus enterprise annuity
Mexico	Spanish (MX)	MXN 1,000,000	Global index fund (MSCI World ETF)	CETES	Afore retirement account
Turkey	Turkish	TRY 2,000,000	Global equity index fund (MSCI World ETF)	Turkish government bonds	BES private pension account

Table A2: Query accounting by experimental arm

Arm	Version	Countries	Question–age blocks	Cells	Queries
Country/language	B	21	Q1 at ages 30/40/55	63	40,950
	C	21	Q1 at ages 30/40/55	63	40,950
	B-cross	15	Q1 at ages 30/40/55	45	29,250
	B	21	Q2 and Q3 at age 40	42	27,300
	C	21	Q2 and Q3 at age 40	42	27,300
	B-cross	15	Q2 and Q3 at age 40	30	19,500
Stated investor	B$\gamma\rho\tau$ (six cells)	21	Q1 at age 40	126	81,900
Total				411	267,150

Notes: A cell is one country × version × question × age combination, and for the stated-investor arm one country × treatment cell. Every cell is queried by all thirteen models fifty times, so each cell contributes $13 \times 50 = 650$ queries. Version B-cross exists only for the fifteen non-English countries. The two pre-registered collection-robustness cells (temperature bookends at 0 and 1.0, and Version B with the investment amount stated in U.S. dollars) added 8,450 queries outside this grid and are reported in Section 5.6. The exploratory re-collection of the Version B, Question 1, age-40 cell with the Instructions block translated (Section 5.6, Appendix Table A16) added 5,460 queries, twenty per model and country, on October 29, 2026. The exploratory re-collection with the stake restated at purchasing-power parity (Section 5.6, Appendix Table A16) added 13,325 queries on the same day, twenty-five per model and cell, of which 6,825 re-collect Version B unchanged as a same-day reference. Four further exploratory collections also lie outside the grid. The balance-sheet probes of the stated-investor arm (Section 6.3) added 30,550 queries on August 10–14, 2026. The model-tracking wave (Appendix Table A17) added 11,270 queries on September 19–20, 2026. The elicitation probe (Appendix Table A16) added 2,730 queries, ten per model and country, on October 28, 2026. The VPN location check (Section 5.6) added 960 queries on October 30, 2026.

A.3 Model versions and access dates

Table A3 lists every AI model used to create or classify the data. Panel A reports the thirteen models under study, with the identifier sent to the OpenRouter API and the dated version the API returned. Panel B reports the four judge models used to code the explanation text, the primary one of which also coded the elicited questions. Panel C reports the two models used to cross-check the translations. All models were accessed through the OpenRouter API, except the model behind the adversarial review of the translations described in the table notes.

Table A3: Model versions and access dates

Model	Requested identifier	Dated version returned by the API
Panel A: Models under study (accessed July 2–4, 2026)		
Claude Haiku 4.5	`anthropic/claude-haiku-4.5`	`anthropic/claude-4.5-haiku-20251001`
Claude Sonnet 4.6	`anthropic/claude-sonnet-4.6`	`anthropic/claude-4.6-sonnet-20260217`
Claude Opus 4.6	`anthropic/claude-opus-4.6`	`anthropic/claude-4.6-opus-20260205`
Claude Opus 4.8	`anthropic/claude-opus-4.8`	`anthropic/claude-4.8-opus-20260528`
GPT-5-mini	`openai/gpt-5-mini`	`openai/gpt-5-mini-2025-08-07`
GPT-5.5	`openai/gpt-5.5`	`openai/gpt-5.5-20260423`
Gemini 3 Flash	`google/gemini-3-flash-preview`	`google/gemini-3-flash-preview-20251217`
Gemini 3.5 Flash	`google/gemini-3.5-flash`	`google/gemini-3.5-flash-20260519`
Gemini 3.1 Pro	`google/gemini-3.1-pro-preview`	`google/gemini-3.1-pro-preview-20260219`
Grok 4.3	`x-ai/grok-4.3`	`x-ai/grok-4.3-20260430`
DeepSeek V4 Flash	`deepseek/deepseek-v4-flash`	`deepseek/deepseek-v4-flash-20260423`
DeepSeek V4 Pro	`deepseek/deepseek-v4-pro`	`deepseek/deepseek-v4-pro-20260423`
Mistral Medium 3.5	`mistralai/mistral-medium-3-5`	`mistralai/mistral-medium-3.5-20260430`
Panel B: Judge models for the explanation coding (accessed July 9–10, 2026)		
Claude Opus 4.7 (primary)	`anthropic/claude-opus-4.7`	not recorded
GPT-5.5 Pro	`openai/gpt-5.5-pro`	not recorded
Gemini 2.5 Pro	`google/gemini-2.5-pro`	not recorded
Qwen 3.7 Max	`qwen/qwen3.7-max`	not recorded
Panel C: Translation cross-check (accessed June 29–July 1, 2026)		
Gemini 3.1 Pro	`google/gemini-3.1-pro-preview`	`google/gemini-3.1-pro-preview-20260219`
GPT-5.5	`openai/gpt-5.5`	`openai/gpt-5.5-20260423`

Notes: Access dates are in UTC. Panel A covers the 267,150 queries of the main collection. Each requested identifier resolved to a single dated version throughout the collection. The API returned no identifier for eleven responses (ten for GPT-5.5 and one for Mistral Medium 3.5). The two pre-registered robustness cells collected on July 6, 2026 returned the same dated versions. The exploratory collections of September 19–20 and October 28–29, 2026 use the same requested identifiers for these models, and the API returned them without a date suffix. Table A17 lists the panel of the September wave. In Panel B, the four judges coded the validation sample on July 9, 2026, and the primary judge coded the full sample on July 9–10, 2026. The coding script did not store the identifier returned by the API. The script requested temperature zero for all four judges. Claude Opus 4.7 and GPT-5.5 Pro do not accept a temperature parameter and ran at their default sampling. Claude Opus 4.7 coded the 12,452 questions of the elicitation probe in November 2026 with the same request settings. In Panel C, both models ran at temperature zero and returned the same dated versions as in Panel A, where both are also models under study. The cross-check showed them only a translated block, its English source, and the pinned terms, and asked about the translation, never for an allocation. The adversarial review of all ninety translated blocks on July 1, 2026 ran through Claude Code with Claude Opus 4.8, which is also a model under study, and on the same terms. The translated Question blocks come from the DeepL API, accessed June 11, 2026. DeepL does not publish version identifiers for its translation models.

A.4 Translation and quality control

I produced the Version B and Version C Question blocks with the DeepL API on June 11, 2026. Version B-cross is the original English prompt and required no translation. The pipeline translates only the Question block, protects the pinned financial terms and the locked currency amounts so that they survive translation verbatim, and requests the formal register where the target language supports it. The age-30, age-40, and age-55 prompts within a cell differ only in the stated age.

Two automated checks screened the machine output. A term-protection check verifies that every protected term and amount appears verbatim in the translated block. It caught three systematic failures, a dropped currency amount, an amount rendered as a product name, and a dropped pension term. A back-translation of every block to English, scored with sentence BLEU, routed blocks to human review. I did not use the score to accept translations, because it flags faithful translations into Japanese, Korean, and Chinese.

The acceptance gate was native-speaker review of the Version B and Version C blocks in all fifteen translated countries and of the local-instrument wording for India. I reviewed the Swedish blocks myself. South Africa had no available native reviewer, so I verified its blocks, and those of the United Kingdom, Australia, and Canada, with a two-model cross-check (Table A3, Panel C). The cross-check found one error, in the South African Version C pension reference, which sat in the design table rather than in a translation. I corrected it through a dated design-lock amendment on June 29, 2026. The same two models audited the Version C blocks of the translated countries, and I adopted a model suggestion only where it did not contradict a native reviewer. Before I filed the pre-registration, every assembled prompt passed a deterministic consistency audit with zero failures, Claude Opus 4.8 reviewed all ninety unique translated blocks adversarially and found no error, and I read all ninety blocks in full. A Google Translate back-translation of a stratified sample of thirty-four blocks also passed. I then froze the translations for collection.

A.5 Collection and parsing

The runner submits each query to the OpenRouter API as a single user message with temperature 0.7 and a 4,000-token cap, sets no other sampling parameter, and submits the

queries in random order. OpenRouter drops any parameter a model endpoint does not support without an error. Its listing of supported parameters, accessed on 16 November 2026, includes temperature for no endpoint of GPT-5-mini, GPT-5.5 or Claude Opus 4.8, so these three models ran at their providers' default sampling in every arm. The runner logs each parsed response before the raw API response and skips completed queries on restart. A restart on July 4 re-sent two queries, and the staging step keeps the first response to each, so the analysis file holds one observation per query and every cell has exactly fifty.

Parsing proceeds in a fixed order of decreasing strictness. The parser first reads the first balanced brace block in the response as JSON and takes the fraction field that the Instructions block names, and then tries the whole response as JSON. A JSON value above 1 and at most 100, or one stated with a percent sign, is read as a percentage. If both JSON steps fail, the parser searches for a decimal fraction and then for a percentage, in each case only next to explicit allocation language and not next to rate language such as the prompt's interest and inflation rates. Of the 267,150 responses, 266,800 parsed as JSON (2 of them after rescaling a percentage), 113 parsed through the fallback searches, and 237 failed. A parse-method field records the strategy for each observation, so any analysis can be re-run on the strictly JSON-parsed subsample.

The pre-registered truncation rule excludes the 137 responses that stopped at the token ceiling, 130 of which also failed to parse. Together with the other 107 parse failures, this leaves 266,906 responses for analysis. The failures concentrate by model and not by country. DeepSeek V4 Pro accounts for 133 of the 237 and Gemini 3.1 Pro for 47, while no country has more than 22. Twenty-four responses with an error finish reason parsed and stay in the sample, because the pre-registered rules exclude only unparsed and truncated responses.

A response counts as a refusal only when a refusal phrase matches and no numeric recommendation was parsed, so that boilerplate such as "consult a financial advisor" on a valid answer is not misclassified. The phrase list covers all fourteen prompt languages, and the main collection produced no refusals. Every row records both the requested and the returned model identifier. Each requested identifier resolved to a single dated version throughout the main collection (Table A3), and the eleven responses without a returned identifier all failed to parse.

The no-rate arm. The arm of Section 5.4 was collected on November 16 and 17, 2026 with the same runner, temperature, token cap, and parser as the main collection, after a smoke test of 390 queries in three countries that is not part of the data. It comprises 6,825 queries without the rate (twenty-one countries, thirteen models, twenty-five repetitions) and 650 field-check queries (five countries, ten repetitions). Twelve responses stopped at the token ceiling and three failed to parse, which leaves 6,812 and 648 responses for analysis. No response was a refusal. The two assumed rates parse from the JSON object in 6,815 and 6,814 of the 6,825 no-rate responses. A value above one would be read as a percentage, but every model reported the rates as decimals. In November the API returned only the requested identifier for every model, not a dated version, so the versions in Table A3 cannot be confirmed for this arm. The numbered comparison cell is the English Question 2 cell of the main collection (13,629 responses after the same exclusions).

B The Normative Benchmark

This appendix develops the benchmarks summarized in Section 4. I use the benchmark to calculate two empirical objects. First, I want a country-specific optimal equity share α_c^* that I can compare the AI financial advice to. Second, I want to be able to calculate a *calibration slope* that measures how much of the cross-country variation that should appear in the advice actually does appear. This slope nests the "U.S.-default" hypothesis (H2), where the slope would be zero. To this end, I use a simple Merton benchmark that is deliberately static and closed form. While a full life-cycle model in the tradition of Cocco et al. (2005) would add realism on several margins, it would also come at the cost of a large calibration apparatus (for such a model, see, e.g., Choukhmane et al., 2026). Instead, I use a simpler model to illustrate the central mechanism, like background wealth crowding equity into the liquid portfolio.

B.1 Benchmark S: the stake as a standalone portfolio

Consider the investor described in the prompt: age $a = 40$, liquid savings W (the local-currency equivalent of \$50,000), choosing how to split W between a broad equity index fund and domestic government bonds. Let r_c denote the real return on the safe asset in country c, let the expected excess return on the fund be $\mu_c - r_c$ with standard deviation σ_c, both measured in local currency, and let the investor have constant relative risk aversion over terminal wealth with coefficient γ_c. This setup gives the Merton (1969) and Samuelson (1969) rule for the optimal equity share, equation (1) in Section 4.1. The optimal equity share does not depend on the level of wealth. If returns are independent over time, the horizon also does not enter the equation. Because the prompt restricts the answer to a fraction between zero and one, every benchmark below is truncated to the unit interval, $\alpha_c^* = [\theta_c^*]_0^1$ with $[x]_0^1 \equiv \min\{1, \max\{0, x\}\}$.

The first benchmark takes the prompt at face value: the investor is asking about this \$50,000 as a self-contained decision (the investor is "looking to invest \$50,000 over a five-year horizon"). Under Benchmark S the optimal share is simply equation (1),

$$\alpha_c^S \;=\; \left[\frac{\mu_c - r_c}{\gamma_c\,\sigma_c^2}\right]_0^1. \tag{8}$$

How can we measure μ_c and σ_c in this context? The prompt says "a broad stock market index fund," which could mean the local equity market or a global fund held from country c. I take the world equity index measured in local currency as the baseline. This matches the instruments named in Version C (MSCI World trackers in most countries) and avoids attributing "bias" to any sensible reluctance to recommend concentrated exposure to a small home market. I also do a robustness check with μ_c and σ_c based on the local stock market index in Table A4.

Under the world-index reading, the expected excess return is common across currencies by uncovered interest parity. I set the expected excess return at approximately 5.3 percent, the arithmetic mean of the MSCI World index (net, in U.S. dollars, 1999–2025) over the U.S. long-term rate, both nominal. I also use a common risk aversion coefficient for each country, $\gamma_c = \bar{\gamma}$. This assumption can be changed, but note that risk aversion on the country level does not track the country coefficient.

These assumptions imply that cross-country variation in α_c^S comes from local-currency volatility alone. I compute σ_c as the standard deviation of annual nominal local-currency returns on the index, converting the U.S.-dollar return with the annual exchange rate of the Global Macro Database (Müller, Xu, Lehbib and Chen, 2025). The baseline uses these unhedged moments, but Table A4 also reports a currency-hedged variant.

I report results on a risk-aversion grid $\bar{\gamma} \in \{2, 3, 5, 10\}$ with baseline $\bar{\gamma} = 3$. At the baseline, equation (8) produces interior optimal shares from about 0.21 in Turkey to 0.53 in South Korea. Choukhmane et al. (2026) estimate by simulated method of moments a coefficient of relative risk aversion of about 5 that rationalizes the advice of current-generation models, which corresponds to the $\gamma = 5$ point of the sensitivity grid, one step above the baseline. At $\gamma = 2$ the U.S. benchmark rises to about 0.78, and at $\gamma = 10$ shares compress to between 6 and 16 percent across countries, implausibly low relative to observed advice.

Two choices above widen the cross-country spread. First, I hold the arithmetic premium common across currencies. Uncovered interest parity holds in logs, so a common log premium would instead raise the arithmetic premium in volatile currencies by about $\frac{1}{2}(\sigma_c^2 - \sigma_{US}^2)$, with σ_c the volatility of log returns, and with it the benchmark share. Second, I use nominal rather than real volatility. For an investor whose prices move with the exchange rate, nominal

volatility overstates currency risk. In the data this matters mainly for Turkey. Both choices lower the benchmark share where currencies are volatile, so the prescribed deviations for these countries are, if anything, overstated. For the U.S. cell the equity volatility is 18.4 percent (Appendix Table A6), which implies a share of 0.52 at the baseline $\bar{\gamma} = 3$.

B.2 Benchmark I: adding pension wealth and human capital

The prompt also says "I have a standard pension plan otherwise," which invites the model to consider the investor's full balance sheet. The second benchmark therefore integrates two non-tradable assets that a 40-year-old holds outside the liquid portfolio: pension wealth P_c, the present value of future pension benefits, and human capital H_c, the present value of future labor income (Bodie, Merton and Samuelson, 1992; Campbell and Viceira, 2002; Gálvez and Paz-Pardo, 2026). If a fraction ζ of human capital and a fraction ω of pension wealth are equity-like, equation (2) with $B_c^s = (1-\zeta)H_c + (1-\omega)P_c$ and $B_c^r = \zeta H_c + \omega P_c$ gives the optimal share of the liquid portfolio,

$$\alpha_c^I \;=\; \left[\theta_c^* \,\frac{W + H_c + P_c}{W} \;-\; \frac{\zeta\, H_c + \omega\, P_c}{W}\right]_0^1 . \tag{9}$$

With $\zeta = \omega = 0$ this is the rule of Merton (1971) for an investor with riskless non-tradable income, in the form restated by Cocco et al. (2005, equation 12). Safe outside wealth makes the investor less exposed to her liquid portfolio, so she should hold *more* equity in it. The more generous the pension system, the stronger this push. The second term works in the opposite direction when labor income co-moves with the stock market or when the pension itself is invested in stocks. The equity-like parts of human capital and pension wealth are treated as earning the market return and both are discounted at the safe rate.The baseline sets $\omega = 0$, a reasonable description of public and defined-benefit pillars, although treating pension claims as bond-like is itself contestable for pay-as-you-go systems whose benefits are indexed to wages. The exploratory probes of Section 6.3 state the pension's composition directly, which sets $\omega = 0$ for a guaranteed annuity and $\omega = 1$ for a pension invested in stocks.

I calibrate both assets for the average earner. Human capital H_c is the present value of the average wage $\bar{w}_c$, growing at the real rate g_c, until the retirement age R_c. Pension wealth P_c

is the present value of a pension equal to the net replacement rate ρ_c times the final wage, paid for the L_c expected years of retirement. Both are discounted at the local real safe rate. The notes to Appendix Table A6 give the formulas and sources.

With these inputs, $H_c + P_c$ for an average 40-year-old ranges from three times W in India to more than eighty times in Switzerland and Turkey, so equation (9) puts α_c^I at the corner of 100 percent equity in essentially every country. At $\zeta = 0$ the corner holds for all 21 countries through $\bar{\gamma} = 5$ and breaks at $\bar{\gamma} = 10$ only for four countries (Brazil, India, Mexico, South Africa). At $\zeta = 0.2$ it holds everywhere through $\bar{\gamma} = 3$. Benchmark I therefore cannot rationalize cross-country variation in the advice, and interior advice already rules it out in levels. I use it instead for its comparative statics, which give the predicted signs for the tests of Section 6. The uncapped share inside the brackets rises with a generous pension, falls when the pension is invested in stocks, and falls with age as human capital declines.

B.3 Taking the benchmark to the data

The benchmark is computed from external country-level inputs and then confronted with the experimental estimates using two measures, the advice gap and the calibration slope. Appendix Table A6 reports every country-specific input, with sources in the table notes.

Advice gaps. For each country, I compute the gap between mean LLM advice and the benchmark,

$$\Delta_c \;=\; \bar{y}_c - \alpha_c^S, \tag{10}$$

where $\bar{y}_c$ is the mean recommended fraction in the primary cell (Version B, Q1, age 40). The gap Δ_c is a descriptive level comparison and carries no confidence interval.

The calibration slope. The next measure examines whether cross-country *movements* in the benchmark show up in the advice. Let $\hat{b}_c$ denote the estimated country fixed effect from equation (3) (the LLM advice deviation from the U.S. baseline) and let $\delta_c^* = \alpha_c^S - \alpha_{US}^S$ denote the benchmark's prescribed deviation. The regression

$$\hat{b}_c \;=\; \kappa_0 + \kappa\, \delta_c^* + e_c \tag{11}$$

summarizes calibration in a single number. Full normative calibration implies $\kappa = 1$: advice moves one for one with the benchmark. The U.S.-default hypothesis implies $\kappa = 0$. Note that the converse does not hold, since large country effects orthogonal to fundamentals also produce a zero slope. I therefore report the gap table and the slope together. Intermediate values measure partial calibration. The registration specifies this regression and its inference but leaves it at the descriptive altitude. Inference follows the rest of the paper: a wild bootstrap over the twenty non-U.S. countries (Webb weights, 9,999 replications) as the primary procedure with HC2 standard errors alongside, a two-stage bootstrap that propagates the estimation error in $\hat{b}_c$, and DFBETA diagnostics pre-specified for the seven countries with thin benchmark data.

B.4 Advice gaps and the calibration slope

The benchmark prescribes large differences in advice across this panel. At the baseline grid point ($\bar{\gamma} = 3$), the standalone benchmark α_c^S ranges from 0.21 in Turkey and 0.23 in Brazil, whose local-currency volatilities are the highest in the panel, to 0.53 in South Korea and 0.52 in the United States, a prescribed spread of thirty-two percentage points. Mean advice spans twelve points, from 0.46 to 0.58 (Table A4). The advice range is thus about a third of the prescribed range, but the two ranges are not aligned. Turkey, whose prescribed share is the lowest in the panel, receives mean advice of 0.54, in the middle of the realized band and 3.1 points below the U.S. baseline, a difference that does not survive the Holm correction (Table A8). The calibration slope below measures the alignment directly.

Table A4 reports the per-country results. The gaps are positive nearly everywhere, and they are largest where fundamentals depart most from the U.S. baseline: 33 percentage points in Turkey, 23 in Brazil, 17 in Mexico, and 15 in Japan and Poland. Across most of Western Europe, whose prescribed shares sit close to the U.S. value, the gaps are within a few points of zero, with France exact to within a fifth of a point.

The hedged columns of Table A4 replace the local-currency volatility with the U.S.-dollar volatility of the index. With the premium common across countries, the hedged benchmark is 0.52 in every country. Mean advice lies within 6.2 percentage points of the hedged benchmark everywhere, with Brazil furthest below. The thirty-two-point spread of the unhedged benchmark, and with it every non-zero prescribed deviation δ_c^* below, therefore comes from

Table A4: Advice gaps against the standalone benchmark

	Mean	Unhedged		Hedged		Home index	
Country	advice	α_c^S	Gap Δ_c	α_c^S	Gap	α_c^S	Gap
Turkey	0.54	0.21	0.33	0.52	0.02	0.21	0.33
Brazil	0.46	0.23	0.23	0.52	−0.06	0.38	0.08
Mexico	0.54	0.38	0.17	0.52	0.02	—	—
Japan	0.58	0.43	0.15	0.52	0.06	0.51	0.07
Poland	0.50	0.35	0.15	0.52	−0.03	0.14	0.35
United Kingdom	0.58	0.49	0.09	0.52	0.05	0.57	0.00
Sweden	0.57	0.48	0.09	0.52	0.05	0.44	0.13
India	0.55	0.46	0.09	0.52	0.03	0.38	0.17
Australia	0.58	0.49	0.08	0.52	0.05	0.74	−0.17
South Africa	0.56	0.49	0.07	0.52	0.04	0.58	−0.02
South Korea	0.58	0.53	0.06	0.52	0.06	0.42	0.16
United States	0.57	0.52	0.05	0.52	0.05	0.68	−0.11
China	0.52	0.48	0.04	0.52	0.00	—	—
Canada	0.56	0.52	0.04	0.52	0.04	0.68	−0.12
Finland	0.55	0.52	0.03	0.52	0.02	0.08	0.47
Switzerland	0.53	0.51	0.02	0.52	0.00	0.46	0.06
Germany	0.53	0.52	0.01	0.52	0.01	0.35	0.18
Netherlands	0.53	0.52	0.01	0.52	0.01	0.43	0.09
Spain	0.53	0.52	0.01	0.52	0.00	0.34	0.19
France	0.52	0.52	−0.00	0.52	−0.01	0.31	0.20
Italy	0.48	0.52	−0.03	0.52	−0.04	0.13	0.35

Notes: Mean recommended equity share by country in the primary cell (Question 1, Version B, age 40) against the standalone benchmark α_c^S at the baseline grid point ($\bar{\gamma} = 3$, world index in local currency). The gap is $\Delta_c = \bar{y}_c - \alpha_c^S$. The hedged columns replace σ_c with the U.S.-dollar volatility of the index, which a currency-hedged world fund carries in every currency to first order, and report the benchmark and the gap under that assumption. The home-index columns replace both moments with those of the country's own MSCI country index in local currency (net total return, 2001–2025, premium over the Global Macro Database long rate) at the same γ_c, an exploratory post-registration variant. A dash marks the two countries without a country series (China, Mexico). Countries are ordered by the unhedged gap.

exchange-rate risk alone.

The home-index columns use each country's own MSCI country index in local currency, with the premium and the volatility estimated over 2001 to 2025. The prescribed share then runs from 0.08 in Finland to 0.74 in Australia, a spread of sixty-six points, but the advice follows it only weakly. Over the eighteen non-U.S. countries with a series, the slope of the country

effects on the home-index deviations is 0.11 (wild bootstrap 95 percent confidence interval $[0.04, 0.21]$), against 0.14 for the world index on the same countries.

Section 5.5 reports the calibration slope, $\hat{\kappa} = 0.13$, and its leave-one-out range. The primary wild bootstrap over countries gives a 95 percent confidence interval of $[-0.145, 0.328]$ (standard error 0.112). The pre-registered influence diagnostics flag Brazil and Turkey, the two most volatile currencies, whose DFBETA values on the slope exceed $2/\sqrt{N}$ in absolute value. The registered two-stage bootstrap, which resamples the thirteen models and re-estimates the twenty country effects on every draw, gives a standard error of 0.032 and an interval of $[0.067, 0.189]$. The small positive slope is therefore not first-stage noise, but with twenty countries it cannot be distinguished from zero. Both procedures place the slope far below one.

Measurement error in δ_c^* would attenuate $\hat{\kappa}$ toward zero, but sampling noise in the return moments is too small to matter. Resampling the twenty-seven return years (1999–2025) and recomputing every δ_c^* on each draw moves the slope only from 0.13 to 0.10, although this addresses sampling error in σ_c and not the choices discussed in Section B.1.

The choice of risk aversion does not help match the advice and the benchmark. Because δ_c^* is proportional to $1/\bar{\gamma}$ wherever the benchmark is interior, as it is in every country on the grid, $\hat{\kappa}$ is proportional to $\bar{\gamma}$, rising from 0.09 at $\bar{\gamma} = 2$ to 0.43 at $\bar{\gamma} = 10$. Full calibration would require $\bar{\gamma} \approx 23$, at which the benchmark prescribes no more than 7 percent equity in any country, while the level of the advice is consistent with $\bar{\gamma}$ near 3.

B.5 The mortgage benchmark

This subsection derives the benchmark of Section 4.2 for Question 2 (invest versus prepay a 4 percent mortgage). Prepaying a mortgage is a risk-free investment whose return is the after-tax mortgage rate. Because every return in this section is real, the prompt's quoted nominal rate m must be deflated, so the return to prepayment is

$$r_c^m \;=\; m\,(1 - \tau_c^m) - \pi_c,$$

where τ_c^m is the share of mortgage interest that is effectively deductible in country c and π_c is expected inflation, proxied by the trailing twenty-year mean of realized inflation. The

expected equity return is $\mu_c = r_c + (\mu - r)$, the local real safe rate plus the common premium.

Two country-specific channels operate here. First, deductibility differs sharply across the sample (broadly deductible in the United States for itemizers, no owner-occupier deduction in Germany). Second, expected inflation differs. A fixed nominal mortgage is a better real investment where inflation is low, as in Japan. Prepayment is the household's safe asset only if it pays more than local government bonds, however. Where the after-tax real mortgage rate falls below the local real bond rate r_c, the household should keep the mortgage and hold bonds for the safe part of its portfolio (Becker and Shabani, 2010). The safe rate is therefore $\max\{r_c^m,\, r_c\}$, and the Merton logic gives

$$\alpha_c^{Q2} \;=\; \min\Big\{1,\; \max\Big\{0,\; \frac{\mu_c - \max\{r_c^m,\, r_c\}}{\gamma_c\, \sigma_c^2}\Big\}\Big\}, \tag{12}$$

which is Equation (2) of Becker and Shabani (2010) with after-tax real rates.

The benchmark has two regions. Where r_c^m exceeds r_c, prepayment raises the safe rate, so α_c^{Q2} lies below the allocation benchmark α_c^S of Equation (8), and full prepayment is optimal wherever the expected equity return μ_c falls below r_c^m. Where r_c exceeds r_c^m, the mortgage is cheap debt that should not be prepaid, and $\alpha_c^{Q2} = \alpha_c^S$. In this region the benchmark places the remainder of the savings in bonds rather than in prepayment. Note that this was not an option offered to the AI models, which had to either prepay or invest in stocks. At the baseline calibration the second region contains the five premise-acceptance countries, where the prompted 4 percent lies below local inflation, and Poland, where the two rates differ by less than half a percentage point. Because effective deductibility depends on income, itemization, and loan size rather than on a single statutory rate, I treat τ_c^m as a calibration assumption rather than a point-sourced number and report α_c^{Q2} across a $\tau^m \in [0, 0.5]$ band. The advice lies above the benchmark in every country even at the equity-favorable end of the band.

As mentioned above, the mortgage question offers only two uses of the savings: the index fund and prepayment. If the household could not hold bonds, the safe rate would be r_c^m alone. Table A5 reports this two-option benchmark $\tilde{\alpha}_c^{Q2}$, Equation (12) with r_c^m in place of $\max\{r_c^m,\, r_c\}$, alongside α_c^{Q2}. The two coincide wherever prepayment pays more than bonds,

Table A5: Advice gaps against the mortgage benchmark

Country	Mean advice	Benchmark α_c^{Q2}	Band ($\tau^m \in [0, 0.5]$)	Two-option $\tilde{\alpha}_c^{Q2}$	Gap Δ_c^{Q2}
United States	0.83	0.47	[0.41, 0.52]	0.47	0.36
Japan	0.54	0.20	[0.16, 0.32]	0.20	0.34
France	0.67	0.33	[0.33, 0.52]	0.33	0.34
Sweden	0.73	0.39	[0.28, 0.46]	0.39	0.34
Switzerland	0.61	0.27	[0.20, 0.39]	0.27	0.34
Canada	0.70	0.37	[0.37, 0.52]	0.37	0.33
Germany	0.61	0.28	[0.28, 0.48]	0.28	0.33
United Kingdom	0.68	0.35	[0.35, 0.49]	0.35	0.33
Finland	0.63	0.32	[0.32, 0.51]	0.32	0.32
Poland	0.66	0.35	[0.35, 0.35]	0.38	0.31
Netherlands	0.67	0.44	[0.31, 0.50]	0.44	0.22
Spain	0.62	0.41	[0.41, 0.52]	0.41	0.21
Australia	0.65	0.45	[0.45, 0.49]	0.45	0.20
Italy	0.63	0.45	[0.45, 0.52]	0.45	0.18
South Korea	0.65	0.47	[0.45, 0.53]	0.47	0.18
China	0.60	0.47	[0.45, 0.48]	0.47	0.13
Turkey	0.81	0.21	[0.21, 0.21]	0.65	0.60
Brazil	0.81	0.23	[0.23, 0.23]	0.34	0.57
Mexico	0.81	0.38	[0.38, 0.38]	0.61	0.43
India	0.79	0.46	[0.46, 0.46]	0.79	0.32
South Africa	0.80	0.49	[0.49, 0.49]	0.98	0.31

Notes: Mean recommended index-fund share on the mortgage question (Question 2, Version B, age 40) against the country-specific benchmark α_c^{Q2} from Equation (12) at the baseline grid point ($\bar{\gamma} = 3$), which uses the higher of the after-tax real mortgage rate and the local real bond rate as the safe return. The band reports α_c^{Q2} at the two ends of the effective-deductibility range, $\tau^m = 0$ and $\tau^m = 0.5$. The two-option benchmark $\tilde{\alpha}_c^{Q2}$ uses the after-tax real mortgage rate alone, as if the household could not hold bonds. The gap is $\Delta_c^{Q2} = \bar{y}_c^{Q2} - \alpha_c^{Q2}$, so a positive value means the model recommends more equity, and hence less prepayment, than the benchmark. Countries above the rule form the calibration group, ordered by gap. The five below form the premise-acceptance group of Section 3. In all five, local bonds pay more in real terms than the prompted mortgage costs, so $\alpha_c^{Q2} = \alpha_c^S$. $N \approx 650$ per country.

so they differ only in the six countries above. Because $\tilde{\alpha}_c^{Q2}$ is not bounded by α_c^S, it reaches 0.98 in South Africa, where the prompted rate is far below local inflation.

Table A6: Country-specific calibration inputs for the normative benchmark

Country	σ_c (%)	r_c (%)	z_c (std)	ρ_c (%)	$\bar{w}_c$ (000s)	R_c (age)	L_c (yrs)	g_c (%)	τ_c^m	π_c (%)	α_c^S ($\bar{\gamma}$=3)
United States (USD)	18.4	0.4	0.12	51.3	87	67	18	1.1	0.15	2.5	0.52
United Kingdom (GBP)	19.1	-0.3	0.05	54.2	47	68	17	0.3	0.00	2.9	0.49
Sweden (SEK)	19.2	-0.3	0.05	66.3	546	70	16	1.0	0.30	2.1	0.48
Finland (EUR)	18.5	-0.1	-0.28	65.7	51	68	17	0.5	0.00	2.0	0.52
Germany (EUR)	18.5	-0.6	-0.04	53.3	55	67	18	0.8	0.00	2.2	0.52
France (EUR)	18.5	0.2	-0.03	70.0	44	65	22	0.6	0.00	1.9	0.52
Spain (EUR)	18.5	0.9	-0.16	86.3	35	65	22	0.5	0.00	2.0	0.52
Japan (JPY)	20.4	-0.1	-0.36	42.4	5,186	65	22	-0.2	0.10	0.9	0.43
Brazil (BRL)	27.5	0.8	-0.25	97.5	56	64	20	1.2	0.00	5.5	0.23
Australia (AUD)	19.0	0.9	0.14	53.0	105	67	20	0.7	0.00	2.7	0.49
Canada (CAD)	18.5	0.3	0.18	45.1	86	65	21	0.9	0.00	2.2	0.52
India (INR)	19.6	1.0	-0.28	44.6	273	58	23	4.9	0.10	6.4	0.46
South Africa (ZAR)	19.0	3.7	0.97	8.9	339	60	20	0.1	0.00	5.5	0.49
Netherlands (EUR)	18.5	-0.5	0.19	96.0	61	70	15	0.1	0.35	2.4	0.52
Italy (EUR)	18.5	1.3	-0.09	79.0	35	70	16	-0.3	0.00	2.0	0.52
Switzerland (CHF)	18.7	0.2	-0.02	47.5	98	65	22	0.7	0.20	0.5	0.51
Poland (PLN)	22.5	0.7	-0.07	40.6	101	62	21	2.6	0.00	3.7	0.35
South Korea (KRW)	18.4	1.0	-0.04	38.9	53,916	65	21	1.3	0.05	2.2	0.53
China (CNY)	19.1	1.4	-0.02	103.6	106	60	22	6.8	0.05	2.2	0.48
Mexico (MXN)	21.7	2.9	-0.14	79.6	277	65	18	0.1	0.00	4.4	0.38
Turkey (TRY)	29.1	-4.2	0.02	96.4	743	64	18	5.7	0.00	19.4	0.21

Notes: Country-level inputs to the normative benchmark of Appendix B, one row per country. σ_c is the standard deviation of annual nominal returns on the MSCI World index (net, 1999–2025) in local currency, unhedged, computed from the U.S.-dollar return and the annual exchange rate of the Global Macro Database, and r_c is the real safe rate, the local long-term government yield less realized inflation, taken as a trailing twenty-year mean. The equity premium $\mu_c - r_c$ is held common across countries at 5.32 percent, the arithmetic mean of the net MSCI World index in U.S. dollars over the U.S. long-term rate (1999–2025). z_c is the country-mean GPS willingness to take risks, standardized at the individual level (Falk et al., 2018), which enters the tests of Section 5.5 and not the benchmark. ρ_c is the OECD net pension replacement rate for the average earner and R_c the normal retirement age (OECD, 2023). L_c is the expected number of years in retirement, the expected age at death conditional on reaching 65 less R_c. $\bar{w}_c$ is the average annual wage in thousands of local currency, from the OECD average annual wages series and from national statistics for Brazil, China, India and South Africa. g_c is real wage growth, the 2005–2025 growth rate of the OECD real wage (2009–2024 for Turkey, whose series has a break in 2009), or of real GDP per capita where no OECD series exists. Human capital is $H_c = \sum_{t=1}^{R_c-40} \bar{w}_c(1+g_c)^t(1+r_c)^{-t}$ and pension wealth is $P_c = \rho_c \bar{w}_c(1+g_c)^{R_c-40} a_c(1+r_c)^{-(R_c-40)}$, where $a_c = [1-(1+r_c)^{-L_c}]/r_c$ is the annuity factor, set to L_c in the seven countries with a negative real rate. α_c^I is at the corner in all seven, so nothing depends on this choice. τ_c^m is the assumed effective mortgage-interest deductibility for the mortgage question. π_c proxies expected inflation with the trailing twenty-year mean of realized inflation from the Global Macro Database. The final column is the standalone benchmark share α_c^S at the baseline grid point ($\bar{\gamma} = 3$). The thin-data group (Brazil, China, India, Turkey, Mexico, South Africa, Poland) is the group pre-specified for influence diagnostics because its local-bond and labor-market inputs are less reliable. The volatility series cover the same twenty-seven years in every country. The prompt stake W is the round local-currency equivalent of \$50,000.

C Appendix Tables and Figures

C.1 Tables

Table A7: The pre-registered hypotheses, their tests, and their outcomes

	Statement	Test and correction	Status	Result
H1	Equity recommendations differ systematically across countries (country fixed effects in Eq. (3)).	Per-country wild-bootstrap tests of the twenty contrasts, corrected as a single Holm family.	Primary	Supported. Seven of twenty contrasts survive the Holm correction, all negative, from Brazil (−11.0 pp) to Mexico (−2.7 pp). Section 5.2, Table A8.
H2	Advice does not move with country fundamentals (the U.S.-default anchor): $\kappa = 0$ in the calibration regression (11).	Wild bootstrap over the twenty non-U.S. countries, HC2 standard errors alongside.	Second altitude[a]	Descriptive read consistent with compression. $\hat{\kappa} = 0.13$ (HC2 se 0.11), wild $p = 0.32$ against $\kappa = 0$, and $\kappa = 1$ is rejected. Section 4.1, Appendix B.4.
H3	The U.K. contrast is equivalent to zero within a ±5 pp margin, consistent with shared Anglo-American norms.	Two one-sided tests: supported when the 90 percent wild-bootstrap CI for b_{UK} lies inside the margin.	Primary	Supported. $\hat{b}_{UK} = +0.7$ pp, 95 percent CI $[-0.8, +2.2]$, inside the margin. Section 5.2, Table A8.
H4	Asking in the local language rather than in English shifts advice, holding the country name fixed ($\lambda \neq 0$ in Eq. (5)).	Single pre-specified coefficient, reported uncorrected.	Primary	Supported. $\hat{\lambda} = -1.3$ pp, $p = 0.007$. Section 5.3, Figure 3.
H5	Naming local instruments and pension vehicles shifts advice beyond the country name alone ($\phi \neq 0$ in Eq. (4)).	Single pre-specified coefficient, reported uncorrected.	Primary	Not supported at the 5 percent level. $\hat{\phi} = -0.7$ pp, $p = 0.069$. Section 5.3, Figure 3.
H6	The age-55 tilt toward bonds is smaller where pension replacement rates are higher (age-55 × standardized replacement rate in Eq. (7)).	Single pre-specified interaction, reported uncorrected.	Secondary	Not supported. −0.02 pp, $p = 0.964$. Section 6.2, Table A20.
H7	Country effects covary positively with the country's GPS risk-taking score.	Bivariate slope over the twenty non-U.S. countries, wild bootstrap over countries.	Second altitude[a]	Descriptive read unsupported. +0.77 pp per standard deviation, wild $p = 0.272$. Section 5.5, Table A14.
H8	Country effects covary positively with the country's financial literacy score.	Bivariate slope over the twenty non-U.S. countries, wild bootstrap over countries.	Second altitude[a]	Descriptive read unsupported. +0.89 pp per standard deviation, wild $p = 0.178$. Section 5.5, Table A14.

Table A7: The pre-registered hypotheses, their tests, and their outcomes (continued)

	Statement	Test and correction	Status	Result
H9	The rate at which the explanation text invokes a given rationale differs across countries (Section 7).	Country fixed effects per admitted category, one Holm family per category.	Second altitude[b]	Supported where the design gives it variation. Local-context references are elevated in all twenty non-U.S. countries (+29 to +78 pp, nineteen Holm-surviving) and inflation reasoning in the high-inflation countries (Turkey +56 pp). Pension rates are flat, and horizon sits at its 100 percent ceiling. Section 7, Table 3.
H10	Version C raises the rate of local financial references in the explanation text relative to Version B ($\phi_{\text{text}} > 0$, Section 7).	Single pre-specified coefficient, reported uncorrected.	Second altitude[b]	Supported. +23.3 pp, $p < 0.001$, against the −0.7 pp allocation effect of H5. Section 7, Figure 10.
H11	A stated risk-tolerant investor is recommended more equity than a stated risk-averse one ($\beta_R > 0$ in Eq. (6)).	Single coefficient, directional prediction tested with a two-sided p-value, reported uncorrected.	Secondary confirmatory	Supported. $\hat{\beta}_R = +46.3$ pp, $p < 0.001$. Section 6, Table A18.
H12	A stated generous pension moves the recommendation ($\beta_P \neq 0$ in Eq. (6), two-sided by design).	Single coefficient, two-sided, reported uncorrected.	Secondary confirmatory	Not supported, and precisely so. $\hat{\beta}_P = -1.0$ pp, $p = 0.107$. Section 6, Table A18.
H13	The stated horizon moves the recommendation, against the CRRA prediction of exact irrelevance ($\beta_H \neq 0$ in the horizon analog of Eq. (6)).	Single coefficient, reported uncorrected.	Secondary confirmatory	Supported. $\hat{\beta}_H = +39.9$ pp, $p < 0.001$. Section 6, Table A18.

Notes: The table records the complete set of hypotheses in the pre-registration, independently of which ones the text discusses. The Result column gives the verdict and the headline estimate, and the exhibits it cites report confidence intervals and the full set of estimates. Estimates are in percentage points of the recommended equity share. Unless noted otherwise, p-values come from the wild cluster bootstrap (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$), and the H3 confidence interval is a 95 percent interval from the same procedure. All reported p-values are two-sided: the directional predictions (H7, H8, H11, H13) are stated as directional but tested with two-sided p-values, a conservative choice. The equivalence hypothesis H3 is the exception, tested by two one-sided tests as its statement requires. The second-stage tests (H2, H7, H8) instead bootstrap over the twenty non-U.S. countries.
[a] Registered as pre-specified but descriptive at filing.
[b] Registered as pre-specified but descriptive at filing, with the coding instrument frozen and its pre-registered reliability gates reported by the judge-validation amendment of record (filed 2026-07-15) before any coefficient of the explanation-text regression (Section 7) was estimated. Because it post-dates the pilot, these are pre-specified descriptive tests rather than blind confirmatory ones. Three of the four categories passed both gates. The horizon category is degenerate at 100.0 percent prevalence.

Table A8: Country means and country effects on the recommended equity share (Question 1, Version B, age 40)

	Recommended share		Country effect relative to the U.S.			
Country	Mean	SD	Estimate	95% CI	p (wild)	p (Holm)
Brazil	0.46	0.14	−11.0	[−15.0, −6.9]	<0.001	0.002
Italy	0.48	0.14	−8.6	[−11.9, −5.3]	<0.001	<0.001
Poland	0.50	0.14	−7.4	[−10.8, −3.9]	<0.001	0.011
France	0.51	0.14	−5.4	[−7.5, −3.4]	<0.001	<0.001
China	0.52	0.10	−4.4	[−8.8, −0.2]	0.041	0.371
Spain	0.53	0.12	−4.4	[−6.3, −2.5]	<0.001	<0.001
Switzerland	0.53	0.14	−4.3	[−7.6, −1.1]	0.016	0.200
Netherlands	0.53	0.14	−4.2	[−6.8, −1.7]	0.004	0.049
Germany	0.53	0.14	−4.0	[−7.7, −0.4]	0.034	0.335
Turkey	0.54	0.11	−3.1	[−6.5, −0.1]	0.044	0.371
Mexico	0.54	0.14	−2.7	[−4.2, −1.2]	<0.001	0.003
Finland	0.55	0.15	−2.4	[−5.0, 0.4]	0.086	0.602
India	0.55	0.12	−1.9	[−3.3, −0.4]	0.015	0.200
Canada	0.56	0.13	−1.1	[−2.1, −0.1]	0.029	0.317
South Africa	0.56	0.12	−0.9	[−2.4, 0.5]	0.219	1.000
Sweden	0.57	0.13	−0.2	[−3.7, 3.4]	0.931	1.000
United States	0.57	0.13	omitted baseline			
Australia	0.58	0.12	0.6	[−0.7, 1.9]	0.343	1.000
United Kingdom	0.58	0.13	0.7	[−0.8, 2.2]	0.352	1.000
Japan	0.58	0.14	1.2	[−2.1, 4.8]	0.492	1.000
South Korea	0.58	0.11	1.4	[−2.8, 5.7]	0.496	1.000

Notes: The first two columns report the mean and standard deviation of the recommended equity share by country on the primary cell (Question 1, Version B, age 40), pooled across the thirteen models with fifty repetitions per country–model cell. Each country has 648 to 650 responses after the pre-registered exclusion of responses without a parsed numeric recommendation or truncated at the token limit. The remaining columns report the country fixed effects b_c from equation (3), estimated on the same cell (N = 13,639) with model fixed effects and the United States as the omitted baseline. Estimates and confidence intervals are in percentage points of the recommended equity share, and countries are ordered by point estimate. p (wild) is the wild cluster bootstrap p-value (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$), and the 95 percent confidence interval comes from the same procedure. p (Holm) applies the pre-registered Holm step-down correction across the family of twenty contrasts. Figure 2 plots these estimates.

Table A9: Country effects on the mortgage and inflation questions (Version B, age 40)

	Mortgage (Question 2)				Inflation (Question 3)			
Country	Estimate	95% CI	p (wild)	p (Holm)	Estimate	95% CI	p (wild)	p (Holm)
Japan	−28.6	[−42.5, −15.0]	<0.001	0.002	−11.3	[−13.7, −8.9]	<0.001	<0.001
China	−22.7	[−32.5, −13.7]	<0.001	0.002	−7.2	[−9.6, −4.7]	<0.001	<0.001
Switzerland	−21.6	[−29.8, −14.0]	<0.001	0.002	−5.0	[−7.3, −2.7]	<0.001	0.006
Germany	−21.1	[−30.0, −12.0]	<0.001	0.002	−2.0	[−4.1, −0.1]	0.039	0.473
Spain	−20.4	[−28.6, −12.3]	<0.001	0.002	−4.0	[−6.6, −1.3]	0.007	0.091
Italy	−19.5	[−26.7, −12.5]	<0.001	0.002	−5.0	[−7.3, −2.8]	<0.001	0.002
Finland	−19.1	[−27.1, −11.3]	<0.001	0.002	−2.5	[−5.2, −0.1]	0.041	0.473
South Korea	−17.3	[−24.4, −10.1]	<0.001	0.002	−6.9	[−10.3, −3.5]	0.001	0.020
Australia	−17.1	[−24.8, −9.8]	<0.001	0.003	−0.2	[−2.2, 1.9]	0.862	1.000
Poland	−16.2	[−21.0, −11.5]	<0.001	0.002	−1.8	[−5.3, 1.6]	0.269	1.000
Netherlands	−16.0	[−23.2, −9.9]	<0.001	0.002	−6.5	[−8.9, −4.2]	<0.001	0.002
France	−15.7	[−24.6, −6.6]	<0.001	0.003	−4.3	[−6.9, −1.8]	0.003	0.041
United Kingdom	−14.3	[−21.5, −8.1]	<0.001	0.003	−1.5	[−3.2, 0.1]	0.064	0.636
Canada	−12.3	[−18.4, −6.6]	<0.001	0.003	1.6	[−0.4, 3.6]	0.100	0.799
Sweden	−10.0	[−16.5, −4.1]	0.001	0.006	−0.1	[−3.1, 2.7]	0.930	1.000
India[a]	−3.8	[−6.7, −0.8]	0.015	0.060	0.8	[−2.5, 3.9]	0.618	1.000
South Africa[a]	−2.4	[−4.3, −0.7]	0.002	0.008	2.3	[−0.3, 4.9]	0.082	0.739
Mexico[a]	−1.9	[−5.3, 1.7]	0.262	0.787	0.1	[−3.3, 3.4]	0.973	1.000
Brazil[a]	−1.7	[−6.2, 3.0]	0.452	0.904	2.6	[−1.0, 5.9]	0.139	0.972
Turkey[a]	−1.6	[−8.4, 5.3]	0.628	0.904	−0.9	[−4.5, 2.7]	0.596	1.000

Notes: Country fixed effects from the specification of equation (3), with model fixed effects and the United States as the omitted baseline, in percentage points. The mortgage columns report the recommended fraction to invest in a broad stock market index fund rather than prepay a 4 percent mortgage (N = 13,619). The inflation columns report the recommended fraction of savings to move out of a 1 percent savings account in the face of 3 percent inflation (Question 3, the registration's Q4, N = 13,637). The numeric fraction is the primary Question 3 outcome because the collection produced no refusals and the pre-specified refusal contingency never bound (Section 3.4). Countries are ordered by the mortgage estimate. Inference is the model-level wild cluster bootstrap (Webb six-point weights, 9,999 replications, G = 13), and p (Holm) applies the step-down correction across the twenty contrasts of each question, which form separate families as pre-registered. Figure 4 plots these estimates, and Table A13 reports the categorical companion to the inflation columns.

[a] Pre-specified premise-acceptance group. The prevailing mortgage rate at the filing date exceeded twice the prompted 4 percent, so the mortgage contrasts for these countries test whether the model accepts or corrects the premise rather than whether it calibrates to local fundamentals.

Table A10: The mortgage question without the stated rate: country effects

Country	Prevailing rate (%)	4% stated	No rate			Change in gap	
		Estimate	Estimate	95% CI	p (Holm)	Estimate	p (Holm)
Panel A: Country effects relative to the United States							
Switzerland	1.4	−17.2	1.5	[−2.9, 6.2]	1.000	18.7	0.005
Spain	2.8	−18.9	−5.8	[−11.0, −0.6]	0.267	13.1	0.004
Finland	2.8	−17.9	−5.9	[−11.5, −0.4]	0.269	12.1	0.010
Japan	2.8	−26.9	3.0	[−0.6, 7.0]	0.728	29.9	0.004
China	3.1	−25.3	−19.8	[−33.1, −7.5]	0.005	5.5	0.468
France	3.1	−16.7	−0.6	[−4.4, 3.2]	1.000	16.1	0.006
Sweden	3.4	−5.4	−1.3	[−6.0, 3.8]	1.000	4.1	0.468
Italy	3.4	−20.2	−5.8	[−10.5, −1.2]	0.176	14.3	0.005
Netherlands	3.6	−14.5	−3.2	[−8.1, 1.7]	0.979	11.4	0.002
Canada	3.8	−12.3	−11.2	[−18.4, −4.6]	0.015	1.0	1.000
Germany	3.8	−21.2	−5.4	[−11.1, 0.2]	0.431	15.8	0.006
South Korea	4.3	−17.6	−10.7	[−16.2, −5.8]	0.007	7.0	0.070
United Kingdom	4.8	−14.3	−10.4	[−17.5, −4.4]	0.016	3.9	0.468
Australia	5.7	−17.1	−18.6	[−29.5, −8.3]	0.016	−1.5	1.000
Poland	6.0	−12.0	−20.9	[−32.7, −9.0]	0.007	−8.9	0.468
United States	6.2		*baseline*				
India[a]	8.1	−3.8	−10.8	[−17.6, −5.0]	0.014	−7.1	0.303
South Africa[a]	10.2	−2.4	−26.7	[−39.9, −13.3]	0.002	−24.3	0.010
Brazil[a]	10.7	−1.8	−28.7	[−43.4, −14.2]	0.002	−26.9	0.009
Mexico[a]	11.4	−1.5	−20.1	[−34.3, −6.2]	0.031	−18.6	0.070
Turkey[a]	36.2	3.5	−5.7	[−16.6, 4.8]	1.000	−9.2	0.584

Panel B: Pooled estimates

	Estimate	95% CI	p (wild)
Field check: asking for the assumed rates	2.79	[0.91, 4.66]	0.008
Gap change per point of rate above 4%	−1.02	[−5.06, −0.33]	<0.001
without Turkey	−4.63	[−5.93, −3.54]	<0.001
Per point of assumed mortgage rate	−1.59	[−2.79, −0.40]	0.017
Per point of assumed inflation	0.78	[−0.19, 1.91]	0.102
sum (zero if only the real rate matters)	−0.82	[−2.25, 0.33]	0.135

Notes: Panel A: country fixed effects from equation (3) with model fixed effects, in percentage points of the fraction invested rather than used to prepay. "4% stated" is the English Question 2 cell of the main collection (July 2026). "No rate" deletes the rate clause (November 2026). "Change in gap" is the difference of the two contrasts from the stacked sample with model-by-collection fixed effects, which removes drift common to all countries. Wild cluster bootstrap over models (Webb, 9,999 replications), Holm over the twenty non-U.S. countries in each column. Prevailing rate from the pre-registered rate supplement. Panel B, in percentage points: the field check against the same question without the assumption fields (five countries, country and model fixed effects), the change in gap regressed on the prevailing rate minus 4 across countries (HC2, wild bootstrap over countries), and the share regressed on the assumed rates within countries (country and model fixed effects). Exploratory, post-filing.

[a] Premise-acceptance group.

Table A11: The mortgage question without the stated rate: assumed rates by country

	Mortgage rate (%)			Inflation (%)		
Country	Prevailing	Assumed	Model range	2024	Assumed	Model range
Switzerland	1.4	2.0	1.5–2.7	1.1	1.4	1.0–2.0
Spain	2.8	3.3	3.0–3.6	2.8	2.2	2.0–2.5
Finland	2.8	3.5	2.5–4.1	0.9	2.1	2.0–2.5
Japan	2.8	1.1	0.6–1.5	2.6	1.8	1.3–2.5
China	3.1	4.1	3.7–4.7	0.4	2.1	1.5–2.5
France	3.1	2.8	1.5–3.5	2.4	2.1	2.0–2.5
Sweden	3.4	4.0	3.4–4.5	2.8	2.1	2.0–2.5
Italy	3.4	3.4	3.0–3.6	1.2	2.1	2.0–2.5
Netherlands	3.6	3.7	3.4–4.1	3.2	2.2	2.0–2.5
Canada	3.8	5.0	4.3–5.5	2.4	2.3	2.0–2.5
Germany	3.8	3.4	2.7–3.8	2.4	2.1	2.0–2.5
South Korea	4.3	4.1	3.5–4.5	2.3	2.4	2.1–2.7
United Kingdom	4.8	4.5	4.1–5.0	2.6	2.4	2.0–2.9
Australia	5.7	6.0	4.7–6.4	3.2	3.0	2.5–3.5
Poland	6.0	6.7	4.5–7.5	3.8	3.7	2.6–4.7
United States	6.2	5.1	4.0–6.5	2.9	2.6	2.0–3.0
India	8.1	8.4	7.2–9.0	4.8	5.3	5.0–6.0
South Africa	10.2	10.9	8.9–11.8	4.6	5.3	5.0–5.6
Brazil	10.7	9.7	6.7–11.3	4.5	4.5	4.0–4.9
Mexico	11.4	9.5	6.5–11.5	4.7	4.3	3.9–5.0
Turkey	36.2	26.6	15.2–41.8	58.3	40.6	26.8–61.0

Notes: Assumed rates are the values each model reports in the `assumed_mortgage_rate` and `assumed_inflation_rate` fields after its recommendation in the no-rate cell. "Assumed" is the mean over all responses, and "Model range" is the smallest and largest of the thirteen model means. Every model reported the rates as decimals. The prevailing mortgage rate is from the pre-registered rate supplement, and inflation is 2024 consumer price inflation from the Global Macro Database. Countries are ordered by the prevailing mortgage rate.

Table A12: The mortgage question without the stated rate: by model

	Share	Field check (pp)		Assumed (%)		
Model	invested (%)	Diff.	(SE)	Mortgage	Inflation	Corr.
claude-haiku-4.5	70.3	0.8	(0.7)	4.4	2.8	0.91
claude-opus-4.6	56.2	−2.1	(3.2)	5.3	3.1	0.93
claude-opus-4.8	63.4	1.4	(4.0)	5.4	2.8	0.94
claude-sonnet-4.6	69.6	−0.2	(2.1)	5.3	3.0	0.96
deepseek-v4-flash	67.1	6.8	(2.6)	4.8	2.7	0.91
deepseek-v4-pro	72.1	5.5	(4.2)	5.3	2.9	0.91
gemini-3-flash-preview	75.6	−1.2	(4.7)	5.2	2.7	0.85
gemini-3.1-pro-preview	64.0	5.9	(4.1)	5.2	2.9	0.83
gemini-3.5-flash	65.8	4.9	(2.4)	5.2	2.7	0.88
gpt-5-mini	70.8	5.8	(1.6)	4.7	2.6	0.90
gpt-5.5	52.1	0.8	(3.0)	5.4	2.8	0.92
grok-4.3	66.5	5.4	(1.5)	4.9	2.7	0.92
mistral-medium-3-5	73.8	2.5	(0.7)	4.6	2.6	0.90

Notes: "Share invested" is the mean recommended fraction invested in the no-rate cell across all twenty-one countries. "Field check" is the difference in percentage points between the no-rate cell and the same question without the assumption fields in the five field-check countries, from a regression with country fixed effects within each model, with standard errors clustered by country. "Assumed" gives the model's mean assumed rates over the twenty countries other than Turkey. "Corr." is the correlation across the twenty-one countries between the log of the model's mean assumed mortgage rate and the log of the prevailing rate. Descriptive.

Table A13: Primary-instrument recommendation on the inflation question by country (Question 3, Version B, age 40)

Country	Equities	Infl.-linked	Gold	Real estate	Other
Brazil	13.1	86.7	–	–	0.2
Turkey	17.5	66.0	16.3	–	0.2
Mexico	41.9	58.1	–	–	–
Poland	61.1	38.6	–	–	0.3
South Africa	82.0	16.3	–	–	1.7
India	88.5	7.8	3.7	–	–
Italy	93.2	6.8	–	–	–
South Korea	94.0	6.0	–	–	–
Spain	96.4	3.6	–	–	–
United States	96.8	2.9	–	–	0.3
Canada	97.2	2.6	–	–	0.2
France	97.8	2.2	–	–	–
United Kingdom	98.0	2.0	–	–	–
Australia	98.3	1.5	–	–	0.2
Switzerland	95.2	0.5	–	4.3	–
Netherlands	99.5	0.3	–	–	0.2
Japan	99.7	0.2	–	–	0.2
Sweden	99.8	0.2	–	–	–
China	96.3	0.2	–	–	3.5
Germany	88.6	0.2	–	11.2	–
Finland	100.0	–	–	–	–

Notes: Share of responses (in percent) recommending each instrument category as the primary way to protect savings against inflation (Question 3, Version B, age 40), pooled across the thirteen models. Each response is assigned a single primary instrument by a deterministic multilingual keyword classifier, and the five categories sum to 100. "Other" collects mixed, bond, and money-market funds and the rare response with no reported instrument. Countries are ordered by the combined inflation-protection share (inflation-linked bonds plus gold), descending. A dash denotes a category recommended in under 0.05 percent of responses. This is the categorical companion to the numeric outcome of Table A9, Panel B, discussed in Section 5.4.

Table A14: Country effects and country characteristics, by question

	Allocation (Q1)			Mortgage (Q2)			Inflation (Q3)		
Characteristic	Slope	p	R^2	Slope	p	R^2	Slope	p	R^2
Preferences and financial literacy									
GPS risk-taking	0.77	[0.27]	0.05	2.44	[0.44]	0.09	1.25	[0.37]	0.12
GPS patience	1.23	[0.16]	0.14	−3.45	[0.08]	0.18	−0.61	[0.48]	0.03
Financial literacy	0.89	[0.18]	0.07	−2.76	[0.15]	0.12	0.23	[0.76]	0.00
Language and income									
English prompt (0/1)	3.46	[0.02]	0.21	5.58	[0.16]	0.09	4.26	[0.01]	0.27
Log GDP per capita	0.62	[0.38]	0.04	−5.26	[0.01]	0.42	−1.51	[0.07]	0.17
Equity returns and risk									
Volatility, world index in local currency	−1.27	[0.25]	0.15	4.56	[0.06]	0.32	1.17	[0.13]	0.10
Equity premium, home index	0.86	[0.31]	0.06	3.86	[0.07]	0.25	0.29	[0.70]	0.01
Volatility, home index	−0.47	[0.42]	0.02	3.41	[0.21]	0.20	0.14	[0.83]	0.00
Interest rates and inflation									
Log policy rate	−0.88	[0.37]	0.07	6.33	[0.01]	0.62	2.41	[0.01]	0.44
Log ten-year yield	−0.40	[0.52]	0.01	6.59	[0.00]	0.67	2.31	[0.00]	0.41
Log mortgage rate	−0.19	[0.81]	0.00	6.61	[0.00]	0.67	2.04	[0.00]	0.32
Log inflation	0.18	[0.69]	0.00	5.39	[0.00]	0.45	1.51	[0.01]	0.17
Real interest rate	−0.09	[0.81]	0.00	0.67	[0.84]	0.01	0.55	[0.51]	0.02
Institutions									
Pension replacement rate	−1.70	[0.02]	0.26	0.37	[0.88]	0.00	−0.53	[0.62]	0.02
Mortgage interest deduction	0.48	[0.42]	0.02	−0.93	[0.46]	0.01	−1.05	[0.19]	0.08
Normative benchmark									
Normative benchmark for the question	1.22	[0.32]	0.13	0.47	[0.86]	0.00			

Notes: Each cell reports a separate regression, across the twenty non-U.S. countries, of the country fixed effects on one country characteristic. The dependent variables are the country effects of Appendix Tables A8 and A9 (Version B, age 40, United States omitted), in percentage points. Continuous characteristics are standardized within the sample, so a slope is the change in the country effect per standard deviation of the characteristic. Interest rates, inflation, and GDP per capita enter in logs. The English-prompt indicator is a dummy. The home-index rows cover eighteen countries due to a lack of data on the MSCI country series for China and Mexico. The benchmark row uses the benchmark share for each question, Benchmark S for Question 1 (Section 4.1) and α_c^{Q2} for Question 2 (Section 4.2), in deviation from the U.S. value. Question 3 has no benchmark. p in brackets is the wild bootstrap p-value over countries (Webb weights, 9,999 replications). Holm-adjusted p-values within each question are reported in the text. The rows on GPS risk-taking and financial literacy are the pre-registered H7 and H8 slopes in standardized units, and the Question 1 benchmark row is the calibration slope κ in standardized units. All other cells are post-filing exploratory analyses.

Table A15: Collection-robustness cells: temperature bookends and numeral magnitude

Panel A: Country contrasts vs. the U.S. at the temperature bookends

	$T = 0$		$T = 0.7$ (design)		
Country	Estimate	95% CI	Estimate	95% CI	Es
Brazil	−11.1	[−16.2, −6.0]	−12.5	[−17.1, −8.0]	
Germany	−3.7	[−8.7, 1.4]	−4.4	[−9.4, 0.5]	
Sweden	−0.8	[−5.2, 3.7]	−0.7	[−5.3, 4.1]	
Japan	3.8	[−1.7, 9.4]	1.8	[−1.8, 5.9]	
Mean within-cell SD, 10 models	0.036		0.039		
Mean within-cell SD, 3 models without temperature	0.033		0.032		

Panel B: Stating the amount as $50,000 instead of the native-currency figure

Country	Mean (Version B)	Mean (B-USD)	Shift	95% CI	p (wild)
Germany	0.53	0.54	1.2	[−0.4, 2.7]	0.130
Japan	0.58	0.55	−2.8	[−5.4, −0.4]	0.010
Sweden	0.57	0.58	1.2	[−0.8, 3.4]	0.255

Notes: The table reports the two pre-registered collection-robustness cells (Section 5.6), re-collected on July 6, 2026 alongside the main corpus (Section 3.4). Panel A re-estimates the country fixed effects of equation (3) separately at each sampling temperature on the five bookend countries (Question 1, Version B, age 40, United States omitted), in percentage points of the recommended equity share. The $T = 0.7$ column uses the primary-collection cells and the bookend columns use the re-collected cells, so the $T = 0.7$ estimates differ from Appendix Table A8 only in the estimation sample, which is restricted to these five countries. GPT-5-mini, GPT-5.5 and Claude Opus 4.8 do not accept a temperature parameter (Section 3.4), so their bookend cells re-collect the provider default, and Panel A estimates the contrasts on the other ten models ($G = 10$). The dispersion rows report the mean across country–model cells of the within-cell standard deviation of the recommended share, in fraction units, separately for the ten models and for the three without a temperature parameter. Panel B compares Version B, whose stake is a round native-currency amount, with an otherwise identical prompt stating the amount as $50,000, for three pre-registered countries with country and language held fixed. The mean columns are recommended equity shares, and the shift is the B-USD minus Version-B difference in percentage points, estimated with model fixed effects. All p-values and confidence intervals come from the wild cluster bootstrap (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$ in Panel B and $G = 10$ in Panel A). The pooled estimates quoted in Section 5.6 average over the countries shown here.

Table A16: Country contrasts in the exploratory re-collections: stake at purchasing-power parity, translated instructions, and the elicitation probe

Panel A: Country contrasts vs. the U.S., percentage points

Country	Stake ratio	Main collection (July)	Same-day re-collection		PPP − filed	Translated instructions	Elicitation probe
			Filed stake	PPP stake			
Brazil	2.00	−11.0†	−11.5†	−9.7†	1.8	−10.5†	−8.5†
Italy	1.50	−8.6†	−8.9†	−7.6	1.2	−11.1†	−4.4†
Poland	2.00	−7.4†	−8.4†	−6.1†	2.3	−9.6†	−4.5
France	1.32	−5.4†	−6.1†	−5.6†	0.5	−6.3†	−3.6†
China	1.94	−4.4	−6.3†	−4.4	1.9	−8.4	−3.3
Spain	1.61	−4.4†	−6.0†	−3.9†	2.1	−5.3†	−3.5†
Switzerland	0.93	−4.3	−4.3	−3.7	0.6	−5.1	−3.1
Netherlands	1.23	−4.2†	−4.4†	−4.1†	0.3	−5.8†	−1.9
Germany	1.29	−4.0	−5.7	−5.6	0.1	−5.3	−2.4
Turkey	3.33	−3.1	−1.7	−0.8	0.9	−4.4	1.2
Mexico	2.00	−2.7†	−3.2	−2.2	1.1	−6.9†	−4.1
Finland	1.20	−2.4	−3.7	−2.1	1.5	−7.4†	−0.4
India	4.00	−1.9	−2.6†	−0.9	1.7	−3.4	−2.1
Canada	1.22	−1.1	−1.5	−1.2	0.2	−3.0	−1.8
South Africa	2.40	−0.9	−0.2	2.2	2.4	−1.6	−0.7
Sweden	1.18	−0.2	0.7	−0.0	−0.8	−0.8	0.8
Australia	1.10	0.6	0.5	1.6	1.0	−1.0	0.9
United Kingdom	1.21	0.7	0.9	0.5	−0.4	−0.2	0.3
Japan	1.58	1.2	1.5	2.9	1.5	−0.3	−4.8†
South Korea	1.75	1.4	1.3	1.4	0.1	−0.2	−0.3

Panel B: Pooled shifts, percentage points

	Estimate	95% CI	p (wild)
Stake restated at purchasing-power parity			
Pooled shift, PPP minus filed stake (20 countries)	1.0	[0.5, 1.7]	<0.001
Per doubling of the real stake	−1.2	[−2.0, −0.6]	<0.001
net of a common shift	−0.9	[−1.4, −0.4]	0.003
Translated minus English instructions (October minus July)			
All twenty-one countries	−1.2	[−2.7, 0.2]	0.095
English-prompting countries (identical prompts)	−0.6	[−1.8, 0.4]	0.257
Non-English countries	−1.4	[−3.2, 0.1]	0.084
Language component (non-English minus English)	−0.8	[−1.8, 0.2]	0.130

Notes: Panel A reports the country fixed effects of equation (3) (Question 1, age 40, United States omitted, model fixed effects) in five samples. The main collection reproduces Table A8 (July 2026, fifty repetitions per cell). The same-day columns come from a re-collection on October 29, 2026 (twenty-five repetitions per cell) with the filed stake and with the non-U.S. stakes restated at purchasing-power parity (World Bank, 2024), both against the same-day U.S. rows, and the stake ratio is the filed over the restated stake. The translated-instructions column re-collects the primary cell on the same day (twenty repetitions per cell) with the Instructions block in the local language for the fifteen non-English countries, excluding 45 truncated responses, all from the two DeepSeek models. The elicitation-probe column reports the contrasts on the provisional share that the model gives when it may first ask questions (Section 5.6, October 28, 2026, ten repetitions per cell). † marks a contrast significant at five percent after Holm correction across the twenty countries. In Panel B, the doubling rows are the slope on the log real stake without and with an indicator for the restated prompt, and the English-prompting row measures model drift, because those prompts are identical in the two collections. Inference is the wild cluster bootstrap (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$). Post-filing exploratory deviations recorded in the design lock.

Table A17: Model-tracking waves: stability of the advice across model generations

	All models		New models	Carried-over models	
	July	September	September	July	September
Panel A: Level and composition of the variance					
Models	13	21	14	7	7
Responses kept	13,645	11,011	7,339	7,347	3,672
Repetitions per cell	50	25	25	50	25
Mean equity share (%)	54.9	49.4	45.8	55.4	56.6
Variance share, model (%)	78.0	70.5	42.9	73.9	76.7
Variance share, country (%)	2.8	4.7	14.8	1.5	1.5
Variance share, model × country (%)	4.3	6.7	10.6	2.1	5.1
Variance share, within cell (%)	14.9	18.1	31.7	22.5	16.7
Range of country means, pooled over models (pp)	9.4	10.3	13.1	7.0	6.5
SD of country means net of noise, average over models (pp)	3.2	3.5	3.8	2.2	2.9
Non-U.S. cells below the model's own U.S. cell (%)	67	77	79	62	74
Slope of country means on α_c^S	0.118	0.081	0.080	0.073	0.084
(robust standard error)	(0.091)	(0.113)	(0.143)	(0.070)	(0.054)

Panel B: Carried-over models, same identifier

	Mean, July	Mean, September	Change (se)	Δ SD of country means	Largest cell change
DeepSeek V4 Flash	46.3	54.8	+8.46 (1.08)	+2.47	15.2
Grok 4.3	54.5	57.0	+2.43 (0.37)	−0.70	5.4
GPT-5-mini	37.2	37.5	+0.22 (0.26)	+0.09	2.2
Claude Haiku 4.5	72.5	72.6	+0.06 (0.12)	+0.02	1.8
Mistral Medium 3.5	70.1	70.0	−0.07 (0.05)	−0.08	0.6
Gemini 3 Flash	59.4	59.0	−0.44 (0.57)	+2.53	9.2
DeepSeek V4 Pro	47.4	45.0	−2.31 (0.52)	+0.44	6.6

Panel C: Slope on release date, percentage points per year (27 models, 34 points)

	Estimate	Cluster se	p (wild)
SD of country means net of noise	+0.98	0.53	0.19
with wave fixed effects	+0.92	0.59	0.30

Notes: Question 1, age 40, one English prompt per country (Version B in the six English-prompting countries, Version B-cross in the other fifteen), temperature 0.7. The July columns are the main collection of July 2–4, 2026 cut to this slice, and the September columns the tracking wave of September 19–20, 2026. New models are the fourteen September models absent from the paper panel. Carried-over models are the seven paper-panel models re-queried under the same identifier. Panel A: variance shares are sums of squares relative to the total on response-level data, as in Section 5.1. With 25 rather than 50 repetitions per cell, sampling noise raises the between-cell shares by a negligible amount. The SD of country means net of noise subtracts the average sampling variance of the cell means, $\overline{s^2_{mc}/n_{mc}}$, from the variance of a model's twenty-one country means. The benchmark slope regresses the twenty-one country means, pooled over the models in the column, on the standalone benchmark of Appendix B.1. It carries no model fixed effects, so it is a descriptive analog of κ and its changes across columns are

Table A18: Stated-investor arm: cell means, pre-registered effects, and exploratory balance-sheet probes

Panel A: Recommended equity share by treatment cell

Treatment cell	Mean	SD	p10	Median	p90	N
Risk-averse, limited pension	0.30	0.08	0.20	0.30	0.40	13646
Risk-averse, generous pension	0.29	0.07	0.20	0.30	0.40	13647
Risk-tolerant, limited pension	0.77	0.10	0.60	0.80	0.85	13643
Risk-tolerant, generous pension	0.79	0.09	0.70	0.80	0.90	13648
Horizon 1 year	0.36	0.24	0.00	0.30	0.70	13645
Horizon 20 years	0.76	0.05	0.70	0.75	0.80	13647

Panel B: Pre-registered effects, percentage points

	Contrast	Estimate	95% CI	p (wild)
H11	Risk-tolerant vs. risk-averse	46.3	[41.1, 51.4]	<0.001
H12	Generous vs. limited pension	−1.0	[−2.3, 0.2]	0.107
–	Risk × pension interaction	3.6	[2.0, 5.2]	<0.001
H13	20-year vs. 1-year horizon	39.9	[25.7, 53.4]	<0.001

Panel C: Exploratory balance-sheet probes, percentage points

Contrast	Estimate	95% CI	p (wild)
Pension bond-like annuity (Merton: +)	4.8	[0.1, 9.6]	0.042
Pension equity-invested (Merton: −)	−16.9	[−21.8, −12.1]	<0.001
Fund currency-hedged (at mean σ_c)	0.4	[−1.5, 2.2]	0.659
Fund unhedged (at mean σ_c)	−1.3	[−3.2, 0.4]	0.149
Hedged $\times\sigma_c$ (Merton: +)	23.8	[6.1, 42.1]	0.010
Unhedged $\times\sigma_c$ (Merton: −)	36.0	[15.8, 55.5]	0.002
Stake = all liquid wealth (US only)	−23.3	[−30.4, −16.1]	<0.001
Stake = half of wealth (US only)	1.4	[−0.4, 3.3]	0.109
Stake = a third of wealth (US only)	2.5	[−0.1, 4.9]	0.057
Linear B/L slope, per step (US only)	7.6	[5.4, 9.8]	<0.001

Notes: Panel A reports the distribution of the recommended equity share in the stated-investor arm (Version B$\gamma\rho\tau$, Question 1, age 40, English), pooled across the twenty-one countries and thirteen models, fifty repetitions per model, country, and cell. The upper block states the investor's risk attitude and pension generosity at a common five-year horizon, and the lower block states only the horizon. Panel B reports the pre-registered regressions with country and model fixed effects. β_R, β_P, and the descriptive risk–pension interaction come from equation (6) on the four crossed cells, with the risk-averse, limited-pension cell as baseline, and β_H from the horizon analog with the one-year horizon as baseline. These are the estimates behind H11–H13 in Table A7. Panel C reports the three balance-sheet probes added after filing (Section 6.3), each estimated against the no-statement Version B and B-cross reference rows with country and model fixed effects. The pension-composition and hedging probes pool all twenty-one countries (N = 37,447 and 37,565, fifty repetitions per cell in the United States and twenty-five elsewhere), and the wealth ladder is U.S.-only (N = 2,595). The two σ_c-interaction rows are per unit of local-currency volatility σ_c of the world fund, which spans roughly 0.18 to 0.29 across the panel, so panel-range effects are about one tenth of the printed coefficients. The hedged and unhedged rows are cell effects at the mean σ_c. The linear B/L row is the per-rung slope over $D/W = 0, 1, 2$ (the stake stated as all, half, or a third of liquid wealth). Row labels carry the benchmark-implied sign where the benchmark predicts one. Estimates and confidence intervals are in percentage points, and p (wild) and the intervals come from the wild cluster bootstrap (Webb six-point weights, 9,999 replications, clustered by model, G = 13). The probes are post-filing exploratory deviations in no Holm family, collected in August 2026 against reference cells from the July 2026 main collection.

Table A19: Stated-investor effects by country

	Pre-registered cells			Exploratory probes		
Country	Risk tolerance	Horizon	Generous pension	Pension in stocks	Pension annuity	Hedged − unhedged
Japan	48.9	36.9	0.5	−20.0	1.7	0.5*
Turkey	49.7	35.2	0.3	−16.3†	6.3*	−2.4*
South Korea	48.2	34.6	0.6	−19.9	3.0	0.3
Spain	48.0	41.8	1.0	−16.2†	4.9	1.3
Poland	48.2	44.5*	0.9	−13.5†	6.1	3.1
Italy	47.8	42.5	1.4	−14.2†	5.7	3.0
China	48.8	34.1	0.6	−19.0	5.3	3.1
Brazil	48.3	37.6	1.0	−13.8†	7.5*	4.0
Mexico	49.9	34.3	0.9	−19.7*	3.9	1.2
France	47.1	44.9*	1.2	−15.4†	5.1	2.0
Germany	47.7	43.8*	0.5	−14.8†	5.1	0.9*
India	47.5	38.0	0.8	−18.9*	4.1	2.7
South Africa	46.7*	38.5	−0.1	−19.0	4.9	2.1
Switzerland	47.4	44.2†	0.6	−16.5†	5.8*	2.4
Netherlands	47.1	41.5*	0.9	−15.0†	5.2	1.9*
Canada	47.3*	41.1	1.2	−20.6	3.9	0.9*
United States	48.2	35.8	0.9	−21.7	3.5	3.1
United Kingdom	48.3	43.1*	0.9	−15.4†	3.4	1.2*
Australia	47.9	36.5	0.9	−19.9	3.4	1.2*
Finland	48.2	42.2*	1.0	−14.5†	4.7	2.7
Sweden	48.5	46.0†	0.6	−10.8†	3.9	0.8*
Mean of country effects	48.1	39.9	0.8	−16.9	4.6	1.7
Range	46.7 to 49.9	34.1 to 46.0	−0.1 to 1.4	−21.7 to −10.8	1.7 to 7.5	−2.4 to 4.0
Median CI half-width	±5.4	±14.4	±1.4	±5.3	±5.0	±1.6
Countries differing from U.S. (Holm)	0 of 20	2 of 20	0 of 20	12 of 20	0 of 20	0 of 20

Notes: Effect of each stated-investor treatment on the recommended equity share, in percentage points, estimated separately for each country (Question 1, age 40, statements in English). Risk tolerance is the risk-tolerant minus risk-averse difference pooled over the two pension statements, and generous pension the generous minus limited difference pooled over the two risk statements, both from the four crossed cells of equation (6). Horizon is the twenty-year minus one-year difference on the two horizon cells. Pension in stocks and pension annuity are the probe cells of Section 6.3 against the country's own no-statement reference, and hedged − unhedged is the difference between the two hedging cells, estimated on those cells alone. Each column comes from one regression that interacts the treatment with country indicators and absorbs model fixed effects, so each entry is the treatment effect within that country. † marks a country whose effect differs from the U.S. effect at the five percent level after Holm correction within the column's family of twenty contrasts, and * a difference significant at five percent before correction only. The summary rows give the unweighted mean and range of the twenty-one country effects, the median half-width of the per-country 95 percent confidence intervals, and the number of Holm-significant contrasts. All inference is the wild cluster bootstrap (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$). No joint test of equal effects across countries is reported, because the model-clustered covariance of twenty interaction coefficients has rank at most twelve with thirteen clusters. The per-country breakdown is a post-filing exploratory deviation recorded in the design lock.

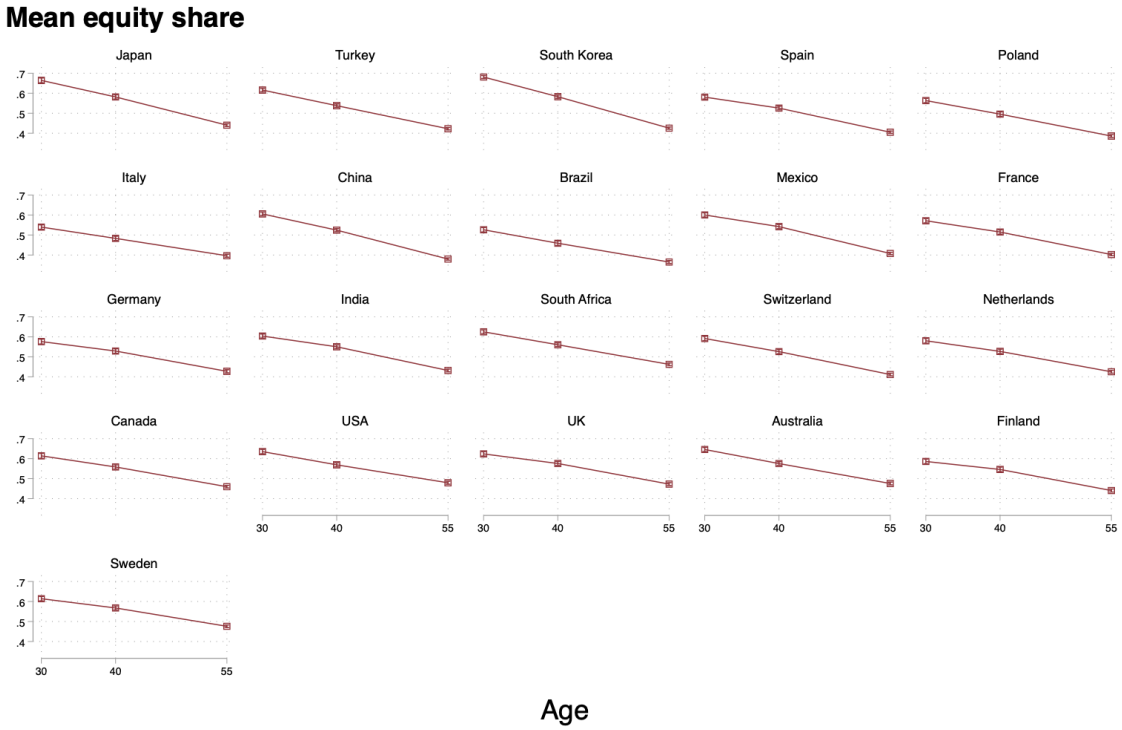
Table A20: The age gradient and its interaction with pension generosity (Question 1, Version B)

	(1) No interaction	(2) Continuous $\tilde{\rho}$	(3) Median split
Age 30 (vs. 40)	6.24	6.24	6.24
	(0.57)	(0.57)	(0.57)
Age 55 (vs. 40)	-11.13	-11.13	-11.34
	(1.90)	(1.90)	(1.96)
$\times$ std. replacement rate (ψ)		-0.02	
		(0.33)	
$\times$ above-median pension			0.44
			(0.61)
Country fixed effects	Yes	Yes	Yes
Model fixed effects	Yes	Yes	Yes
Wild p: Age 30	0.000	0.000	0.000
Wild p: Age 55	0.000	0.000	0.000
Wild p: interaction		0.964	0.489
Observations	40,910	40,910	40,910
Model clusters (G)	13	13	13

Notes: Estimates of equation (7) on Question 1, Version B, pooled across the three prompted ages (30, 40, 55), the twenty-one countries, and the thirteen models. The outcome is the recommended equity share in percentage points, and every column includes country and model fixed effects with the United States and age 40 as the omitted baselines. Column (1) fits the age effects alone. Column (2) adds the pre-registered interaction of the age-55 indicator with the standardized OECD net pension replacement rate $\tilde{\rho}_c$ (the coefficient ψ of hypothesis H6). Column (3) replaces it with the coarser interaction of age 55 with an above-median replacement-rate indicator. Parentheses report CR1 cluster-robust standard errors clustered by model. The wild cluster bootstrap p-values (Webb six-point weights, 9,999 replications, clustered by model, $G = 13$) are the inference of record and are reported in the lower panel.

C.2 Figures

Figure A1: Age glide path by country (Question 1, Version B)



Notes: Mean recommended equity share by stated investor age (30, 40, or 55) within each country, Version B of Question 1, with 95 percent confidence intervals on the cell means. The downward glide path has nearly the same level and slope in every country, the raw pattern behind the null interaction of age with pension generosity in equation (7).