Do Open-Source LLMs Have a Home Bias?

Investment Advice from Chinese, American, European and Thai Models

Thanatchon Chongmankhong

Research note · November 2026

DOI: 10.5281/zenodo.23861180

Abstract

Do open-weight language models allocate more of an equity portfolio to the stock market of the country where they were built? We ask 12 models from China, the United States, Europe, Thailand and Korea, run locally, to split a portfolio across eight markets in 240 factorial prompts that vary the investor's residence, the prompt language, risk tolerance and horizon. In the pre-registered regression a model's home market receives -0.42 pp (95% CI -0.59 to -0.25) relative to what other models give that market: no home tilt, and if anything a small negative one. A post-hoc check that allows for each model's dispersion around equal weight puts the effect at +0.12 pp. The investor's stated residence is associated with a far larger shift of +1.88 pp towards that market, and the prompt language with +0.46 pp. Thai fine-tuning moves the weight on Thailand by -1.04 pp in one pair of models and by +1.26 pp in the other. All prompts, raw replies and code are public.

Keywords: large language models; home bias; portfolio allocation; financial advice; pre-registered audit.
JEL: G11, G41, C45.

1 Why this matters

Language models are entering retail financial advice and research workflows, and central banks have begun to map how AI changes intermediation and asset management [1, 2]. If a model systematically allocated more to one country's assets, that tilt would reach every user who asks it, much as investor home bias [3–5] shapes portfolios one investor at a time. Language models also carry geographic and cultural slants: their default answers resemble the values of English-speaking and Protestant European countries [6, 7], and their ratings of places are systematically lower for regions with weaker socioeconomic conditions [8]. In finance, a recent audit of 32 models finds advice consistent with investor home bias [9], three commercial chat assistants put over 90% of equity portfolios in the United States [10], and whether such advice can meet a fiduciary standard is an open question [11].

Those audits vary the investor, not the model's origin. This note asks the complementary question with open-weight models whose developers sit in different countries: *holding the prompt fixed, does a model allocate more to its own home market?* Because the prompt also states the investor's residence and is written in four languages, the design separates three channels that are confounded in ordinary use: the model's origin (β), the investor's residence (γ) and the prompt language (λ). Two Thai models, fine-tuned from an American and a Chinese base model that are themselves in the sample, let us ask whether fine-tuning moves the allocation towards Thailand. Section 2 describes the design, Section 3 the results, Section 4 the pre-registered robustness checks and Section 5 what the evidence does not show.

2 Design

Models. Twelve open-weight instruction-tuned models, run locally through Ollama [12] at 4-bit quantisation (Q4_K_M; the Ollama default builds of GLM-4-9B and Mistral-NeMo-12B are Q4_0) on one consumer GPU (Table 3). Origin is the lead developer's country: China (Qwen3-8B, Qwen2.5-7B, GLM-4-9B), the United States (Llama-3.1-8B, Gemma-3-12B, Phi-4), Europe (Mistral-7B, Mistral-NeMo-12B, EuroLLM-9B), Thailand (Typhoon2-8B [13], OpenThaiGPT-1.5-7B) and Korea (EXAONE-3.5-7.8B). The home market of a Chinese model is China, of a US model the United States, of a European model the Eurozone, of a Thai model Thailand and of the Korean model South Korea. The two Thai models are fine-tunes of Llama-3.1-8B [14] and Qwen2.5-7B [15], which are also in the sample; these pairs identify whether fine-tuning moves the allocation.

Prompts. Each model answers the same 240 prompts, the full factorial of investor residence (United States, China, France, Thailand, unspecified) $\times$ prompt language (English, Chinese, French, Thai) $\times$ risk tolerance (conservative, balanced, aggressive) $\times$ horizon (3 or 20 years) $\times$ two paraphrases. The investor asks how to split an international equity portfolio across a fixed menu of eight markets: the United States, China, the Eurozone, Japan, the United Kingdom, India, South Korea and Thailand. Japan, the United Kingdom and India are nobody's home market. Menu order is randomised in Latin-square blocks so that every market appears in every position exactly 30 times. The reply is constrained to a JSON schema: eight integer weights, a cash currency and a reason of at most 30 words. Decoding is greedy (temperature 0, seed 42).

Parsing. Weights are rescaled to sum to 100. A reply whose weights do not sum to 95–105 gets one stricter retry; if that also misses, the earliest valid allocation is renormalised (amendment A1, Appendix A). Refusals and invalid replies are dropped. A model with more than 30% refusals plus parse failures would be excluded from the regressions; 0 of the 12 models is (Section 3).

Specification. On the model $\times$ prompt $\times$ market panel of weights $w_{m,p,c}$ (pp) we estimate

$$w_{m,p,c} = \alpha_m + \alpha_c + \beta\, Home_{m,c} + \gamma\, InvestorHome_{p,c} + \lambda\, LanguageHome_{p,c} + \sum_k \theta_k\, Pos_{p,c,k} + \varepsilon_{m,p,c}, \tag{1}$$

where *Home* marks the model's home market, *InvestorHome* the investor's country of residence (France maps to the Eurozone), *LanguageHome* the market whose main language the prompt is written in (English maps to the United States and the United Kingdom), and *Pos* the menu position. Market fixed effects absorb each market's average appeal across all models, so β is the extra weight a model gives its own home market relative to how other models treat that market. Standard errors are clustered by prompt (240 clusters). Because the eight language $\times$ paraphrase templates are the independent units of wording, we also report a wild cluster bootstrap over templates [16, 17].

Hypotheses. H1: $\beta > 0$. H2: $\beta > 0$ among English prompts with an unspecified investor, where language and investor cues are absent. H3: each Thai fine-tune allocates more to Thailand than its base model, estimated with prompt fixed effects. A hypothesis is supported if the 95% CI excludes zero in the predicted direction. The choice of cash currency is analysed with the same right-hand side in a linear probability model for choosing the home currency.

Pre-registration. The pre-analysis plan (`docs/PAP.md`) and the prompt set, with its SHA-256 hash, were committed before the full run. Amendment A1 was made after a 20-prompt pilot

and before any full run, without comparing allocations across models. The robustness checks in Section 4 are the pre-registered list.

3 Results

Instrument. Every model returned a usable allocation for almost every prompt: refusals are at most 0.4% and parse failures at most 0.8% per model, so no model is excluded. Many replies do not sum to 100: the median model has 15.6% of replies outside 95–105 after one retry, and one model 82.5% (Fig. 5, Appendix B). These are renormalised under amendment A1, and the original rule is reported as a robustness check.

Home tilt (H1). Table 1 reports Eq. (1). A model's home market receives $\beta = -0.42$ pp (SE 0.09, 95% CI -0.59 to -0.25; wild cluster bootstrap over templates -0.75 to -0.09) relative to what other models give that market. The estimate is small against an equal weight of 12.5 pp and has the opposite sign to H1, which is *not supported*. Per model (Fig. 1), the home tilt ranges from -3.5 to +1.9 pp: 5 models allocate significantly more to their home market than other models do and 6 allocate significantly less, with no consistent pattern within origin groups. By group, Chinese models allocate +0.60 pp more to China, European models -2.07 pp to the Eurozone, US models -0.17 pp to the United States, Thai models +0.07 pp to Thailand and the Korean model -0.40 pp to South Korea. Among English prompts with no investor nationality (H2), $\beta = -1.02$ pp (SE 0.29; only 12 prompts per model), so H2 is not supported either.

Table 1. Primary and secondary estimates. Weights in percentage points; SEs clustered by prompt except where noted. N counts model × prompt × market observations.

	Estimate (pp)	SE	p	N	Models
Pooled regression, Eq. (1); SE clustered by prompt					
β Home (model origin)	−0.42	0.09	<0.001	23,008	12
γ InvestorHome	+1.88	0.21	<0.001	23,008	12
λ LanguageHome	+0.46	0.13	0.001	23,008	12
β Home, wild cluster bootstrap over 8 templates (95% CI)		[−0.75, −0.09]		23,008	12
Home effect by origin group ($\beta\times$ group, same specification)					
Chinese models	+0.60	0.15	<0.001	23,008	12
US models	−0.17	0.26	0.513	23,008	12
European models	−2.07	0.13	<0.001	23,008	12
Thai models	+0.07	0.12	0.573	23,008	12
Korean model	−0.40	0.16	0.014	23,008	12
H2: English prompts with an unspecified investor					
β Home	−1.02	0.29	<0.001	1,152	12
H3: Thai fine-tune minus its base model, weight on Thailand (prompt fixed effects)					
δ Typhoon2-8B vs Llama-3.1-8B	−1.04	0.26	<0.001	480	2
δ OpenThaiGPT-1.5-7B vs Qwen2.5-7B	+1.26	0.27	<0.001	480	2
Cash currency: home currency chosen (linear probability model, pp)					
β Home	−2.18	0.48	<0.001	23,008	12

Origin, investor and language. The prompt moves allocations more than the model's origin does (Fig. 2). When the investor states a country of residence, that market receives $\gamma = +1.88$ pp more (SE 0.21), 4.4 times the absolute size of the home effect; writing the prompt in a market's main language adds $\lambda = +0.46$ pp (SE 0.13). Levels are far from both market-capitalisation

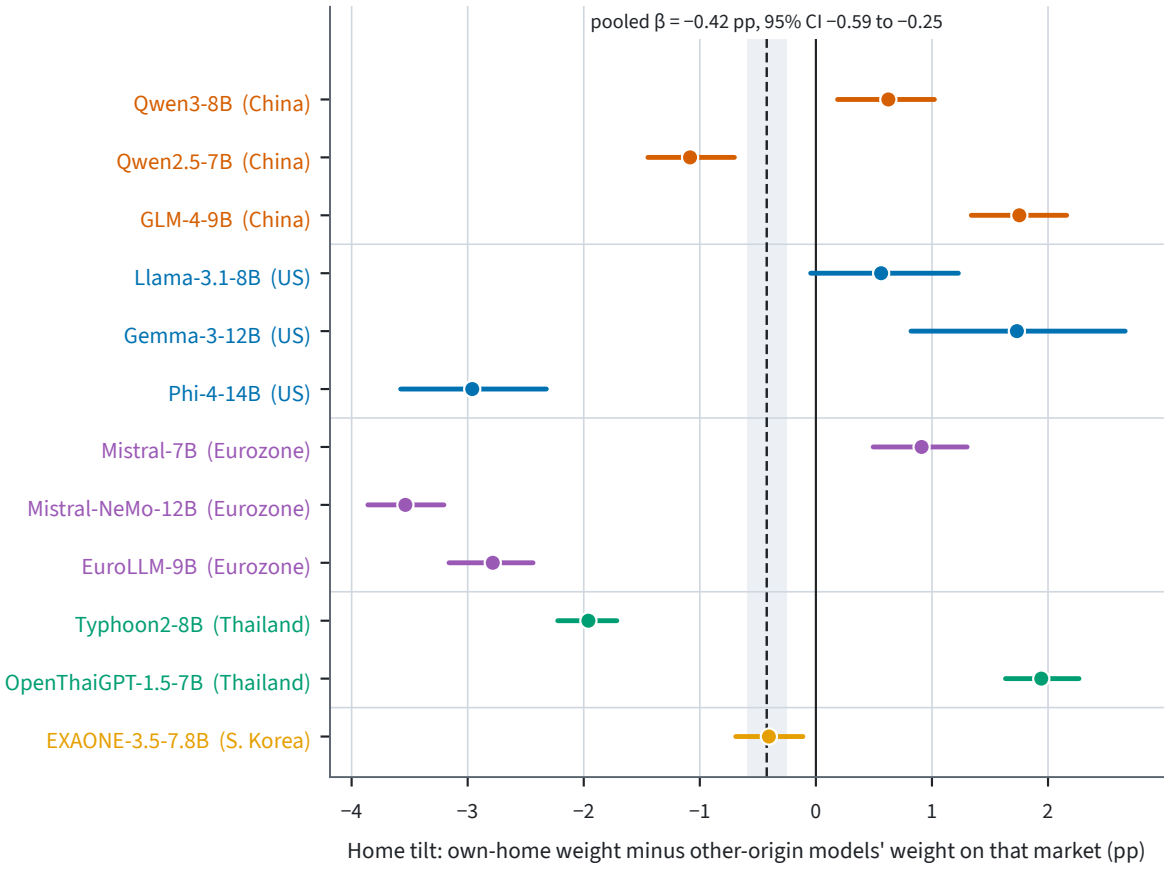


Figure 1. Home tilt by model. Dots: mean over the 240 prompts of the model's weight on its home market minus the mean weight that models of other origins give that market on the same prompt; bars: 95% CI from a cluster bootstrap over prompts (999 draws). Dashed line and band: the pooled β from Eq. (1) with its 95% CI (SE clustered by prompt). Colours mark the origin group; the home market is in parentheses.

and GDP weights (Fig. 3): every model gives the United States 7.9–22.7% of the portfolio and China 13.7–22.5%, against 75.5% and 2.5% in MSCI ACWI and 37.0% and 23.5% by GDP, each renormalised over the eight markets. All models sit much closer to equal weight than to any capitalisation benchmark.

Fine-tuning (H3). The two Thai fine-tunes move in opposite directions (Fig. 4). OpenThaiGPT allocates +1.26 pp more to Thailand than its base Qwen2.5-7B (9.2% against 8.0%), whereas Typhoon2 allocates −1.04 pp relative to its base Llama-3.1-8B (5.3% against 6.4%). H3 is supported for one pair only, so fine-tuning on Thai data is not associated with a consistent shift towards Thailand.

Cash currency. Every model chooses US dollars for cash in 80.4–96.7% of replies. The home-currency coefficient in the linear probability model is −2.18 pp (SE 0.48), so no model favours its own currency beyond what the currency effects explain; the high home-currency share of the US models is the common dollar default, not a home effect.

4 Robustness

Table 2 reports β under each pre-registered check. The negative sign survives keeping only replies that sum to exactly 100 (−0.67 pp), the original 95–105 rule (−0.33 pp, with 4 models

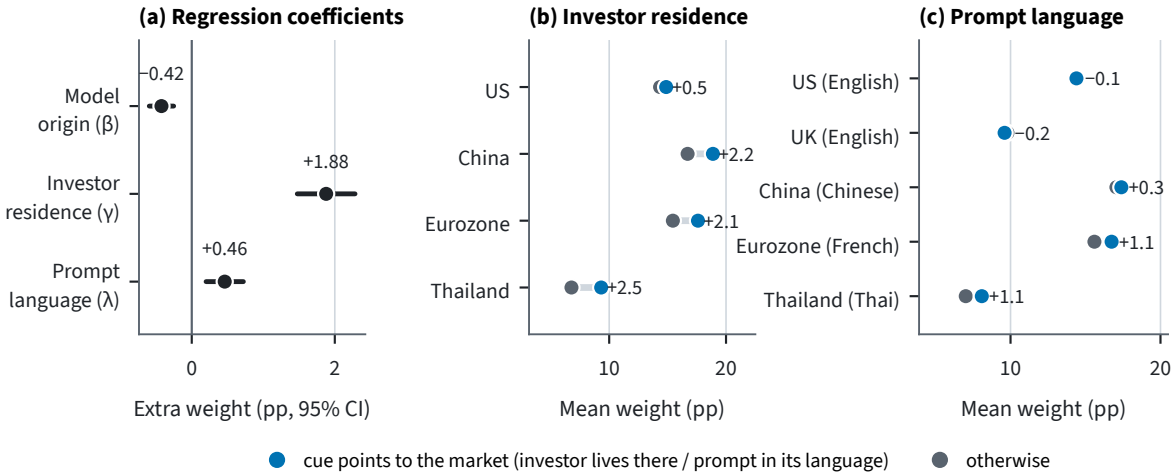


Figure 2. What moves the allocation. (a) The three channels in Eq. (1) with 95% CIs (SE clustered by prompt). (b) Mean weight on a market when the investor states that they live there (blue) and otherwise (grey), over all models and prompts. (c) The same for the prompt language: English maps to the United States and the United Kingdom, Chinese to China, French to the Eurozone and Thai to Thailand. Labels give the difference in pp.

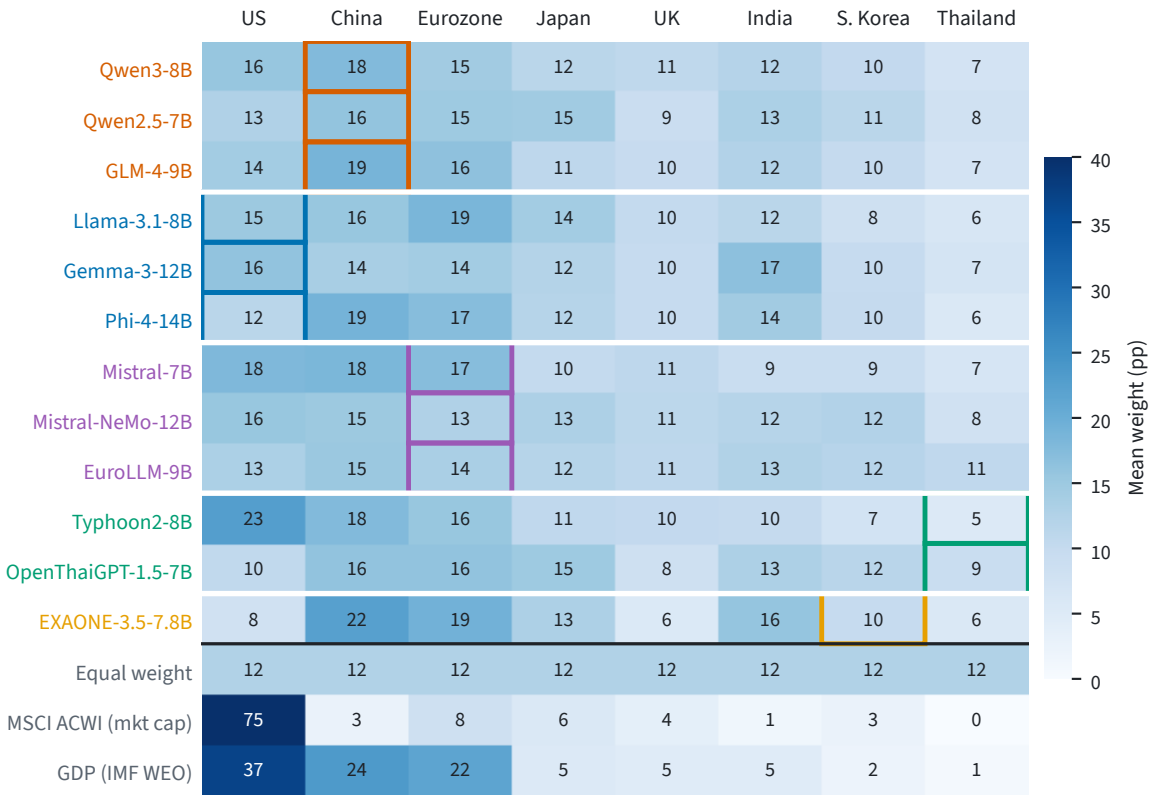


Figure 3. Mean allocation by model and market (pp, over 240 prompts, weights renormalised to 100). Outlined cells: the model's home market. The three reference rows are context, not a benchmark of correctness: equal weight, MSCI ACWI country weights (proxied by an ACWI-tracking ETF's holdings on 29 October 2026) and 2025 GDP (IMF WEO, October 2026), each renormalised over the eight markets.

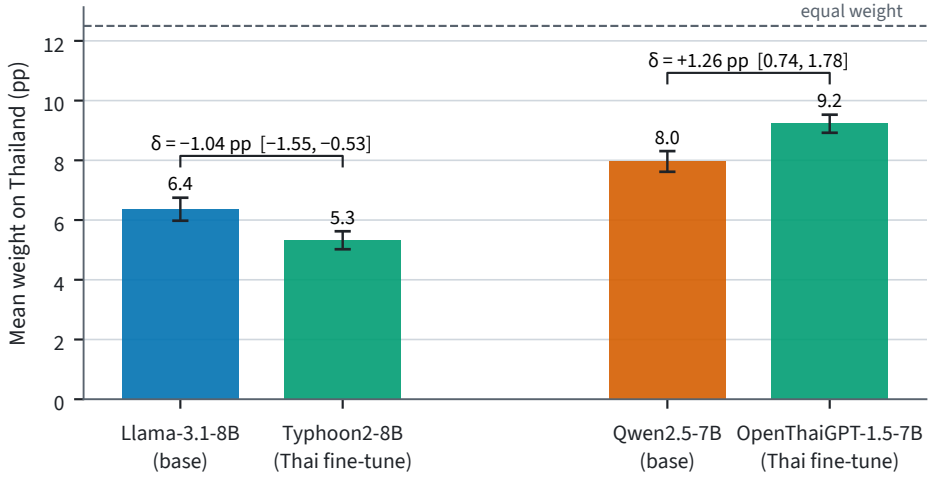


Figure 4. Thai fine-tuning and the weight on Thailand. Bars: mean weight on Thailand over the same 240 prompts, with ±1.96 standard errors; dashed line: equal weight (12.5 pp). δ: fine-tune minus base model from a regression with prompt fixed effects, with its 95% CI (SE clustered by prompt).

excluded), English prompts only (−0.70 pp), unspecified investors only (−0.64 pp) and a log-ratio outcome (−0.026 log points). Leaving out one model at a time gives −0.82 to −0.01 pp; the estimate loses significance without Mistral-NeMo-12B, Phi-4-14B or Typhoon2-8B, so the negative sign rests on a few models. On the 60-prompt robustness subset, greedy decoding gives −0.37 pp, temperature-0.7 sampling with two seeds −0.17 pp (95% CI −0.48 to 0.13) and dropping the system prompt −0.65 pp (95% CI −0.98 to −0.33). The post-hoc specification that allows each model its own dispersion around equal weight gives +0.12 pp (95% CI −0.08 to 0.32). Across every specification, including the leave-one-out fits and the post-hoc check, the highest upper bound of any 95% CI is +0.32 pp: a positive home tilt of one percentage point or more is excluded everywhere.

5 What this does not show

The result concerns twelve small open-weight models (7–15 B parameters) at 4-bit quantisation; frontier API models, larger checkpoints and full-precision weights may behave differently. The eight-market menu, the templates and their translations are one instrument among many: another menu, a request for a single-country recommendation, or native-speaker templates could change the picture. Decoding is greedy, so each estimate describes a model's modal reply, not the distribution a user would sample. Origin is the lead developer's country, which is coarse: Mistral-NeMo was co-developed with NVIDIA, and fine-tunes inherit most of their base model. The outputs say nothing about any developer's intent or training data, and nothing about whether the advice is good: an allocation can be free of home tilt and still be far from any sensible benchmark, as these are.

6 Conclusion

Open-weight LLMs from five origins do not allocate more to their home equity market when the prompt is held fixed: the pre-registered home effect is −0.42 pp, and a post-hoc check for dispersion puts it near zero. What moves the allocation is what the user says. The investor's stated residence is associated with +1.88 pp more for that market, 4.4 times the absolute size of

Table 2. β under the pre-registered robustness checks and the leave-one-model-out fits; the last row is post hoc. Percentage points except the log-ratio row (log points); SEs clustered by prompt.

Check	β (pp)	SE	95% CI	p	N	Models
Pre-registered checks						
Primary specification (greedy, 240 prompts)	−0.42	0.09	[−0.59, −0.25]	<0.001	23,008	12
Temperature 0.7, two seeds (60-prompt subset)	−0.17	0.16	[−0.48, 0.13]	0.262	11,472	12
Greedy decoding, same 60-prompt subset	−0.37	0.17	[−0.71, −0.03]	0.032	5,752	12
No system prompt, greedy (60-prompt subset)	−0.65	0.16	[−0.98, −0.33]	<0.001	5,712	12
Only replies summing to exactly 100	−0.67	0.14	[−0.95, −0.40]	<0.001	12,784	12
Original 95–105 rule (before amendment A1)	−0.33	0.12	[−0.58, −0.09]	0.007	13,496	8
Failures imputed as equal weights	−0.42	0.09	[−0.59, −0.25]	<0.001	23,040	12
English prompts only	−0.70	0.14	[−0.98, −0.42]	<0.001	5,760	12
Unspecified investor only	−0.64	0.17	[−0.98, −0.30]	<0.001	4,608	12
Log-ratio outcome $\log(w_c/w_{\text{Japan}})$	−0.026	0.007	[−0.040, −0.012]	<0.001	20,132	12
Leave one model out						
Without EuroLLM-9B	−0.25	0.10	[−0.44, −0.06]	0.009	21,096	11
Without EXAONE-3.5-7.8B	−0.39	0.09	[−0.58, −0.21]	<0.001	21,096	11
Without Gemma-3-12B	−0.81	0.07	[−0.95, −0.66]	<0.001	21,088	11
Without GLM-4-9B	−0.61	0.09	[−0.79, −0.44]	<0.001	21,088	11
Without Llama-3.1-8B	−0.52	0.09	[−0.70, −0.35]	<0.001	21,088	11
Without Mistral-7B	−0.46	0.09	[−0.64, −0.28]	<0.001	21,088	11
Without Mistral-NeMo-12B	−0.16	0.09	[−0.34, 0.02]	0.081	21,088	11
Without OpenThaiGPT-1.5-7B	−0.82	0.09	[−1.00, −0.64]	<0.001	21,088	11
Without Phi-4-14B	−0.11	0.09	[−0.29, 0.07]	0.216	21,104	11
Without Qwen2.5-7B	−0.41	0.09	[−0.59, −0.23]	<0.001	21,088	11
Without Qwen3-8B	−0.52	0.09	[−0.70, −0.34]	<0.001	21,088	11
Without Typhoon2-8B	−0.01	0.10	[−0.20, 0.17]	0.885	21,088	11
Post hoc, not pre-registered						
Model-specific dispersion around equal weight	+0.12	0.10	[−0.08, 0.32]	0.243	23,008	12

the pooled home effect, and every model holds cash mostly in US dollars (80.4–96.7% of replies). For users and supervisors, the practical tilt to watch in these models is towards the user's own framing and towards equal weighting, not towards the developer's country.

A Deviations from the pre-analysis plan

A1 (before the full run). The pilot showed that unusable replies were arithmetic errors, not refusals; under the original 95–105 rule five of the twelve models would have been excluded, removing both fine-tune pairs. Replies still outside 95–105 after one retry are therefore renormalised, and the original rule is reported as a robustness check (it excludes 4 models). The temperature-0.7 check uses two seeds instead of three to stay within the GPU budget. *Post hoc.* The dispersion-adjusted specification in Table 2 was added after the primary results were seen and is labelled as such.

B Models and failure rates

Table 3 lists the models with their throughput on one RTX 4070 Ti and their reply status in the primary run; Fig. 5 breaks the status down by model and prompt language.

Table 3. Models, throughput on one RTX 4070 Ti (generated tokens per second) and reply status in the primary run (240 greedy prompts per model).

Model	Developer	Origin	Params (B)	Quant.	tok/s	Refusal (%)	Parse fail (%)
Qwen3-8B	Alibaba	China	8.2	Q4_K_M	81	0.0	0.0
Qwen2.5-7B	Alibaba	China	7.6	Q4_K_M	91	0.0	0.0
GLM-4-9B	Zhipu AI / Tsinghua	China	9.4	Q4_0	80	0.0	0.0
Llama-3.1-8B	Meta	US	8.0	Q4_K_M	86	0.0	0.0
Gemma-3-12B	Google	US	12.2	Q4_K_M	52	0.0	0.0
Phi-4-14B	Microsoft	US	14.7	Q4_K_M	49	0.0	0.8
Mistral-7B	Mistral AI	Europe	7.2	Q4_K_M	94	0.0	0.0
Mistral-NeMo-12B	Mistral AI with NVIDIA	Europe	12.2	Q4_0	63	0.0	0.0
EuroLLM-9B	UTTER consortium	Europe	9.2	Q4_K_M	77	0.4	0.0
Typhoon2-8B	SCB 10X	Thailand	8.0	Q4_K_M	86	0.0	0.0
OpenThaiGPT-1.5-7B	OpenThaiGPT	Thailand	7.6	Q4_K_M	91	0.0	0.0
EXAONE-3.5-7.8B	LG AI Research	Korea	7.8	Q4_K_M	89	0.0	0.4

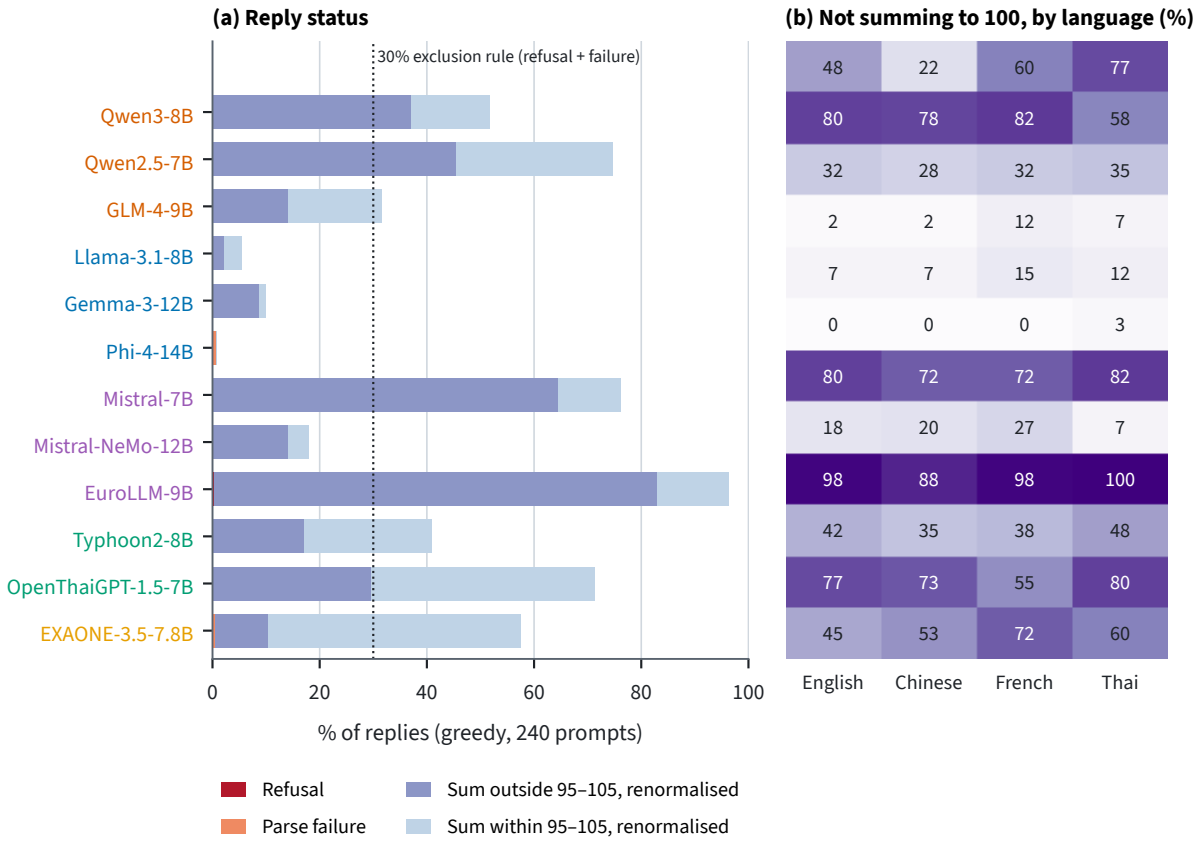


Figure 5. Reply status in the primary run. (a) Share of the 240 greedy replies per model by status: a reply gets one stricter retry if it is invalid or sums outside 95–105, and sums still outside 95–105 after the retry are renormalised (amendment A1). The dotted line is the 30% exclusion rule for refusals plus parse failures. (b) Share of replies not summing to exactly 100, by prompt language.

C Reproduction and use of generative AI

Code, prompts, all raw replies and the parsed panel are at `https://github.com/thanatchon36/llm-home-bias`. `make all` rebuilds every number, table and figure from the cached replies in under a minute without a GPU, and `make paper` typesets this note; `make run` and `make robust` regenerate the replies with Ollama. The study was carried out with Claude (Anthropic), operated through the Claude Code agent, from the author's written brief, which fixed the question, hypotheses, design and decision rules. The agent wrote the pre-analysis plan and the code, ran the models, estimated the results and drafted this note; the author directed the work, reviewed the outputs and takes sole responsibility for the content. Every number in the text is generated from versioned result files.

References

[1] I. Aldasoro, L. Gambacorta, A. Korinek, V. Shreeti, and M. Stein, "Intelligent financial system: How AI is transforming finance," Bank for International Settlements, BIS Working Papers 1194, Jun. 2024. [Online]. Available: https://www.bis.org/publ/work1194.htm

[2] Bank for International Settlements, "Artificial intelligence and the economy: Implications for central banks," in *BIS Annual Economic Report 2024*. Basel: Bank for International Settlements, Jun. 2024, ch. III. [Online]. Available: https://www.bis.org/publ/arpdf/ar2024e3.htm

[3] K. R. French and J. M. Poterba, "Investor diversification and international equity markets," *American Economic Review*, vol. 81, no. 2, pp. 222–226, May 1991.

[4] K. K. Lewis, "Trying to explain home bias in equities and consumption," *Journal of Economic Literature*, vol. 37, no. 2, pp. 571–608, Jun. 1999.

[5] J. D. Coval and T. J. Moskowitz, "Home bias at home: Local equity preference in domestic portfolios," *Journal of Finance*, vol. 54, no. 6, pp. 2045–2073, Dec. 1999.

[6] Y. Tao, O. Viberg, R. S. Baker, and R. F. Kizilcec, "Cultural bias and cultural alignment of large language models," *PNAS Nexus*, vol. 3, no. 9, p. pgae346, Sep. 2024.

[7] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, L. Lovitt, S. McCandlish, O. Sikder, A. Tamkin, J. Thamkul, J. Kaplan, J. Clark, and D. Ganguli, "Towards measuring the representation of subjective global opinions in language models," in *Proceedings of the First Conference on Language Modeling (COLM)*, 2024, arXiv:2306.16388. [Online]. Available: https://openreview.net/forum?id=zl16jLb91v

[8] R. Manvi, S. Khanna, M. Burke, D. B. Lobell, and S. Ermon, "Large language models are geographically biased," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 235. PMLR, 2024, pp. 34 654–34 669. [Online]. Available: https://proceedings.mlr.press/v235/manvi24a.html

[9] C. Fieberg, L. Hornuf, M. Meiler, and D. J. Streich, "Using large language models for financial advice," CESifo, Munich, CESifo Working Paper 11666, Jan. 2025. [Online]. Available: https://www.ifo.de/en/cesifo/publications/2025/working-paper/using-large-language-models-financial-advice

[10] P. Winder, C. Hildebrand, and J. Hartmann, "Biased echoes: Large language models reinforce investment biases and increase portfolio risks of private investors," *PLOS One*, vol. 20, no. 6, p. e0325459, Jun. 2025.

[11] A. W. Lo and J. Ross, "Can ChatGPT plan your retirement?: Generative AI and financial advice," *Harvard Data Science Review*, no. Special Issue 5, Aug. 2024.

[12] Ollama, "Ollama," GitHub repository, 2026, software. Accessed 2026-11-01. [Online]. Available: https://github.com/ollama/ollama

[13] K. Pipatanakul, P. Manakul, N. Nitarach, W. Sirichotedumrong, S. Nonesung, T. Jaknamon, P. Pengpun, P. Taveekitworachai, A. Na-Thalang, S. Sripaisarnmongkol, K. Jirayoot, and K. Tharnpipitchai, "Typhoon 2: A family of open text and multimodal Thai large language models," arXiv:2412.13702, 2024. [Online]. Available: https://arxiv.org/abs/2412.13702

[14] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey *et al.*, "The Llama 3 herd of models," arXiv:2407.21783, 2024. [Online]. Available: https://arxiv.org/abs/2407.21783

[15] Qwen, A. Yang, B. Yang, B. Zhang *et al.*, "Qwen2.5 technical report," arXiv:2412.15115, 2024. [Online]. Available: https://arxiv.org/abs/2412.15115

[16] A. C. Cameron, J. B. Gelbach, and D. L. Miller, "Bootstrap-based improvements for inference with clustered errors," *Review of Economics and Statistics*, vol. 90, no. 3, pp. 414–427, Aug. 2008.

[17] M. D. Webb, "Reworking wild bootstrap-based inference for clustered errors," *Canadian Journal of Economics*, vol. 56, no. 3, pp. 839–858, Aug. 2023, earlier version: Queen's Economics Department Working Paper No. 1315.