Journal of Behavioral Economics for Policy, Vol. 9, Issue 1, 75-80, 2025

Money Illusion in Large Language Models: An Exploratory Replication Study

Milos Ciganovic[1], Nicole Macchitella[1], Luca Panaccione[1]

Abstract

As Large Language Models become increasingly integrated in daily life, understanding the extent to which they may reflect behavioural tendencies embedded in the training data becomes a relevant issue. In this paper, we conduct an exploratory replication study on the occurrence of "money illusion" across six Large Language Models: GPT-5, GPT-5.1, DeepSeek-V3.2, Grok 4.1, Sonnet 4.5, and Gemini-3-pro-preview. Specifically, we replicate the survey questions of Shafir, Diamond, and Tversky (1997) and report on the models' responses across twelve scenarios covering earnings, transactions, contracts, and mental accounting. Our exploratory evidence tends to suggest that Large Language Models generally base their responses on real terms rather than nominal ones, particularly in contexts that require objective evaluations, even though there are some exceptions. However, it also tends to suggest that their responses more frequently aligned with those of the human participants in the original study when more subjective evaluations were involved. While a more systematic investigation based on a larger dataset is required to reach more definitive conclusion, we nonetheless hope that our results could be useful for future research analyses.

JEL classification: D14; D91; G41

Keywords

money illusion — large language models — cognitive bias

[1] *Sapienza University of Rome, Italy*

*__Corresponding author__: luca.panaccione@uniroma1.it

Introduction

Large language models (LLMs) are trained using vast amounts of human-generated text data, which can reflect human behavioral patterns (Durt and Fuchs 2024, Schramowski, Turan, Andersen, Rothkopf, and Kersting 2022). Therefore, it is important to investigate whether and how LLMs could display similar patterns too. Indeed, while LLMs are often assumed to behave like rational agents, recent studies have shown that they exhibit biases similar to those observed in humans (Macmillan-Scott and Musolesi 2024, Salecha, Ireland, Subrahmanya, Sedoc, Ungar, and Eichstaedt 2024, Hu, Kyrychenko, Rathje, Collier, van der Linden, and Roozenbeek 2025).

In this paper, we contribute to the literature on LLMs' behavioural bias in economic decision-making, a relevant issue also because of the growing use of AI tools to support financial decision-making[1] and enhance financial literacy[2] (Ross 2024, Corazzini, Deriu, and Guerzoni 2025, Kotek, Dockum, and Sun 2023). Specifically, we conduct an exploratory replication analysis of "money illusion", an issue which, to the best of our knowledge, has not been examined in previous studies. The implications of money illusion, which usually refers to the tendency to consider monetary amounts in nominal rather than real terms, affect many aspects of economic decision-making, from wage negotiations to consumption and investment behaviour.

Using survey questions, Shafir, Diamond, and Tversky (1997) have extensively analysed money illusion, which they interpret as "a bias in the assessment of the real value of economic transactions, induced by a nominal evaluation" (Shafir, Diamond, and Tversky 1997, p. 348). Their analysis has been replicated across several geographical and cultural contexts (de Moraes Ferreira, Santiago, Bastos, Fatori, Borborema, Seda, and Batistuzzo 2024, Ziano, Li, Tsun, Lei, Kamath, Cheng, and Feldman 2021, Mees and Franses 2014).

We used the survey questions proposed by Shafir, Diamond, and Tversky (1997) to investigate whether six LLMs exhibit instances of money illusion similar to those observed in the original human subjects. Our exploratory evidence tends to suggest that LLMs generally base their responses on real terms rather than nominal ones, particularly in contexts that require objective evaluations, even though there are some exceptions. However, it also tends to suggest that their responses more frequently aligned with those of the human participants in the original study when more subjective

[1] See, e.g., T Elbæk, Uzodinma, Ismagilova, and Mitkidis (2022). See also Upravitelev (2024) for a related discussion.

[2] See, e.g., Choi and Kim (2023). For recent analyses and discussions on the role of financial literacy on financial decision-making, see Conte, De Santis, Melissari, Paiardini, and Temperini (2024) and Mortillaro and Signore (2024).

evaluations were involved.

The remainder of the paper is organized as follows: Section 2 presents the methodology, Section 3 reports the results, and Section 4 concludes.

Methodology

We evaluated six LLMs: GPT-5, GPT-5.1, DeepSeek-V3.2, Grok 4.1, Sonnet 4.5, and Gemini-3 Pro. The models were accessed via their free online chatbot interfaces, using default parameters to ensure that our findings reflect the models' standard behaviour.

Because of the exploratory nature of our analysis, we confronted the models with seven problems from Shafir, Diamond, and Tversky (1997), maintaining the original wording[3] and most of their variations,[4] for a total of twelve scenarios related to earnings (Problem 1), transactions (Problems 2 and 3), contracts (Problem 4), mental accounting (Problems 5 and 6), fairness and morale (Problem 7).

Each scenario was presented in a separate chat session to ensure independence between responses and prevent prior interactions from influencing subsequent answers. Since preliminary testing revealed that models often gave indecisive responses or avoided providing direct answers, we added the instruction "answer the question in the requested format" to each prompt, ensuring transparent and comparable responses across all models. To account for the potential variability of model reasoning across iterations, each model was presented with each scenario 10 times, for a total of $N = 60$ observations per scenario (10 repetitions $\times$ 6 models).[5] All responses were systematically recorded and organized for subsequent analysis.

As a measure of response variability, we report the sample variance (s^2) of the responses, which were suitably coded with integer values starting from zero. Therefore, $s^2 = 0$ corresponds to identical responses across all repetitions and models, whereas larger values reflect greater response variability. A binomial chi-square (χ^2) test is conducted for each condition to determine whether the LLM's choices deviated significantly from a random uniform distribution. In addition, and where applicable, a χ^2 test of independence was calculated to assess whether the framing of the problem significantly altered the distribution of responses across conditions.

Results

While we refer to the original article by Shafir, Diamond, and Tversky (1997) for the complete statement of the problems, which are presented in the original order for ease of reference, we reproduce here the original wording of the questions, each of which is preceded by a brief summary of the scenario considered. To enable comparison, we also report the participants' responses from the original study.

Problem 1 considers salary increases across three different frames related to economic terms, happiness, and job attractiveness.[6] It refers to Ann and Barbara, who graduated from the same college one year apart and took similar jobs upon graduation, both starting with an annual salary of $30,000. After one year with no inflation, Ann received a 2% raise ($600), while Barbara received a 5% raise ($1,500) after a year with 4% inflation. The following questions were posed in relation to this problem:

***Economic terms**: As they entered their second year on the job, who was doing better in economic terms?*

***Happiness**: As they entered their second year on the job, who do you think was happier?*

***Job attractiveness**: As they entered their second year on the job, each received a job offer from another firm. Who do you think was more likely to leave her present position for another job?*

	Economic terms	Happiness	Job attractiveness
Ann	98% (71%)	47% (36%)	18% (65%)
Barbara	2% (29%)	53% (64%)	82% (35%)
χ^{2*}	56.07	0.27	24.07
p-value	< 0.001	$= 0.606$	< 0.001

Note: LLM percentages aggregated across all responses for each scenario ($N = 60$). $s^2 = 0.02$ (Economic terms), $s^2 = 0.25$ (Happiness) and $s^2 = 0.15$ (Job attractiveness). Results from Shafir, Diamond, and Tversky (1997) in parentheses. *Binomial χ^2 test against 50% ($df = 1$).
χ^2 test of independence (condition $\times$ person): $\chi^2(2) = 79.61, p < 0.001$.

When the question is posed in economic terms, the vast majority of LLMs correctly evaluate Ann as better off in real rather than nominal terms. When the framing shifts to well-being, around half of the responses (47%) indicate that Ann is happier, and the majority (82%) indicate that Barbara is more likely to leave her current position. Thus, while LLMs evaluate the scenario in terms of real value when the question is framed in economic and job-attraction terms, they seem to exhibit some money illusion when evaluating happiness. Compared to Shafir, Diamond, and Tversky (1997), our results tend to align the human responses mainly in the happiness scenario.

Problem 2 considers real estate transactions, asking participants to rank three individuals, Adam, Ben, and Carl, based on the success of their transactions.[7] Each individual used an inheritance of $200,000 to buy a house that was sold one

[3] This implies, in particular, that monetary values and dates are as in Shafir, Diamond, and Tversky (1997). In Problem 4, a date is cited that was in the future relative to the respondents. This may make the problem less straightforward to interpret. However, our results do not suggest any particular issue with the responses to this problem.

[4] For parsimony, we omitted some variants of the problems which did not lead to substantial differences in the original study's findings.

[5] Some variability across repeated responses was observed, which is, however, not explored further in this analysis.

[6] Shafir, Diamond, and Tversky (1997, p. 351)

[7] Shafir, Diamond, and Tversky (1997, p. 353)

year later. Adam, who experienced 25% deflation during that year, sold the house for $154,000 (23% less than the price paid). Ben, who experienced no inflation, sold it for $198,000 (1% less than the price paid). Carl, who experienced 25% inflation, sold it for $246,000 (23% more than the price paid). The following question was posed in relation to this problem:

Please rank Adam, Ben, and Carl in terms of the success of their house transactions. Assign '1' to the person who made the best deal, and '3' to the person who made the worst deal.

	Adam	**Ben**	**Carl**
1	98% (37%)	0% (17%)	2% (48%)
2	2% (10%)	98% (73%)	0% (16%)
3	0% (53%)	2% (10%)	98% (36%)
χ^{2*}	114.10	114.10	114.10
p-value	< 0.001	< 0.001	< 0.001

Note: LLM percentages are aggregated across all responses ($N = 60$). $s^2 = 0.02$. Results from Shafir, Diamond, and Tversky (1997) are in parentheses. *Goodness-of-fit χ^2 test against uniform distribution (33.3% each, $df = 2$). χ^2 test of independence (person × rank): $\chi^2(4) = 342.30, p < 0.001$.

Adam sells the house with a 23% nominal loss and 2% real gain; Ben with a 1% nominal and real loss; and Carl with a 23% nominal gain but incurred a 2% real loss. LLMs consistently identify Adam's transaction as the most successful, followed by Ben's and then Carl's, thus correctly evaluating the real, inflation-adjusted outcomes. This departs from the results of Shafir, Diamond, and Tversky (1997), in which participants exhibit money illusion, with evaluations primarily driven by nominal rather than real price changes.

Problem 3 considers nominal price changes in the context of buying and selling decisions.[8] It refers to a hypothetical scenario involving unusually high inflation, where all benefits and salaries, as well as all prices, increase by 25% within six months, so that earnings and expenses are both 25% higher than before. The following questions were posed in relation to this problem:

Six months ago, you were planning to buy a leather armchair whose price during the 6 months went up from $400 to $500. Would you be more or less likely to buy the armchair now?

Six months ago, you were also planning to sell an antique desk you own, whose price during the 6 months went up from $400 to $500. Would you be more or less likely to sell your desk now?

[8] Shafir, Diamond, and Tversky (1997, p. 355)

	Buying armchair	**Selling desk**
More likely	0% (7%)	30% (43%)
Same	67% (55%)	65% (42%)
Less likely	33% (38%)	5% (15%)
χ^{2*}	40.00	32.70
p-value	< 0.001	< 0.001

Note: LLM percentages are aggregated across all responses for each scenario ($N = 60$). $s^2 = 0.23$ (Buying armchair) and $s^2 = 0.84$ (Selling desk). Results from Shafir, Diamond, and Tversky (1997) are in parentheses. *Goodness-of-fit χ^2 test against uniform distribution (33.3% each, $df = 2$). χ^2 test of independence (decision × likelihood): $\chi^2(2) = 30.58, p < 0.001$.

The majority of LLMs recognize that the likelihood of buying or selling the items should remain unchanged, as real prices have not changed despite the higher nominal prices. However, around 30% of the responses report being "less likely" to buy the armchair and "more likely" to sell the desk, indicating that nominal prices prompt a greater willingness to sell and a diminished tendency to buy. This is in line with the direction of the bias found by Shafir, Diamond, and Tversky (1997), albeit at a lower rate.

Problem 4 considers contracting decisions in scenario with three different frames: nominal, real, or neutral.[9] In the hypothetical role of head of a corporate division in Singapore, one should consider choosing a contract to sell computer systems currently priced at $1,000, for delivered in January 1993, when payment is to be received (the problem was originally presented to subjects in the spring of 1991). Due to inflation, the prices in Singapore are expected to rise in the next two years before the expected delivery. The best estimates suggest an increase of approximately 20%, and a 10% decrease is presented just as likely as a 30% increase. The following alternatives were considered in relation to this problem:

(Contracts framed in real terms)

Contract A: You agree to sell the computer systems (in 1993) at $1,200 apiece, no matter what the price of computer systems is at that time. Thus, if inflation is below 20% you will be getting more than the 1993-price; whereas, if inflation exceeds 20% you will be getting less than the 1993-price. Because you have agreed on a fixed price, your profit level will depend on the rate of inflation.

Contract B: You agree to sell the computer systems at 1993's price. Thus, if inflation exceeds 20%, you will be paid more than $1,200, and if inflation is below 20%, you will be paid less than $1,200. Because both production costs and prices are tied to the rate of inflation, your "real" profit will remain essentially the same regardless of the rate of inflation.

(Contracts framed in nominal terms)

Contract C: You agree to sell the computer systems (in 1993) at $1,200 apiece, no matter what the price of computer systems is at that time.

[9] Shafir, Diamond, and Tversky (1997, p. 356)

Contract D: You agree to sell the computer systems at 1993's price. Thus, instead of selling at $1,200 for sure, you will be paid more if inflation exceeds 20%, and less if inflation is below 20%.

(Contracts under a neutral frame)

Contract E: You agree to sell the computer systems (in 1993) at $1,200 apiece, no matter what the price of computer systems is at that time.

Contract F: You agree to sell the computer systems at 1993's prices.

	Real (A/B)	**Nominal (C/D)**	**Neutral (E/F)**
First contract (A/C/E)	0% (19%)	40% (41%)	28% (46%)
Second contract (B/D/F)	100% (81%)	60% (59%)	72% (54%)
χ^{2*}	60.00	2.40	11.27
p-value	< 0.001	$= 0.121$	< 0.001

Note: LLM percentages are aggregated across all responses for each scenario ($N = 60$). $s^2 = 0.00$ (Real frame), $s^2 = 0.24$ (Nominal frame) and $s^2 = 0.21$ (Neutral frame). Results from Shafir, Diamond, and Tversky (1997) are in parentheses. *Binomial χ^2 test against 50% ($df = 1$).
χ^2 test of independence (frame × contract): $\chi^2(2) = 28.87, p < 0.001$.

Although LLMs predominantly choose the second contract, which is riskless in real terms, the response pattern reveals some noticeable differences across frames. Specifically, this contract is always chosen in the real frame, but less frequently in the other two frames. Their response distribution is in line with to the one reported by Shafir, Diamond, and Tversky (1997) in the nominal frame, and is qualitatively similar, albeit at a somewhat different rate, in the neutral frame.

Problems 5 and 6 consider mental accounting by evaluating the effect of past nominal values on current decisions.[10] Problem 5 concerns the evaluation of drinking a bottle of 1982 Bordeaux from a case purchased in the futures market for $20 per bottle, which now sells at auction for $75 per bottle. The following question was posed in relation to this problem:

Which of the following best captures your feeling of the cost to you of drinking this bottle?

Doesn't cost me anything	3% (30%)
$ 20	0% (18%)
$ 20 plus interest	0% (7%)
$75	97% (20%)
Feels like I saved $55	0% (25%)
χ^{2*}	220.67
p-value	< 0.001

Note: LLM percentages are aggregated across all responses ($N = 60$). $s^2 = 0.03$. Results from Shafir, Diamond, and Tversky (1997) and Shafir and Thaler (2006) are in parentheses. *Goodness-of-fit χ^2 test against uniform distribution (20% each, $df = 4$).

The vast majority of responses indicate the replacement value as the cost of drinking the bottle of wine, with only a minority opting for the "costs nothing" option. This contrasts with the results discussed by Shafir, Diamond, and Tversky (1997), who found that only 20% of participants selected the replacement value.

Problem 6 considers two competing bookstores having in stock identical copies of Oscar Wilde's collected writings. Store A purchased the copies for $20 each, and its employee Tom sold 100 copies to a local high school for $44 each. Because of 10% yearly inflation, one year later Store B purchased the copies for $22 each, its employee Joe sold 100 copies to another school for $45 each. The following question was posed in relation to this problem:

Who do you think made the better deal selling the books, Tom or Joe?

Tom	77% (87%)
Joe	23% (13%)
χ^{2*}	17.07
p-value	< 0.001

Note: LLM percentages are aggregated across all responses ($N = 60$). $s^2 = 0.18$. Results from Shafir, Diamond, and Tversky (1997) are in parentheses. *Binomial χ^2 test against 50% ($df = 1$).

The large majority of responses (77%) identify Tom as having made the better deal, focusing on a higher nominal profit margin. Therefore, in this case LLMs aligned with the human tendency to prioritize nominal over real profit.

Problem 7 considers the relationship between fairness and money illusion.[11] It compares Ablex and Booklink, two small publishing firms with a dozen editors each. In a year with no inflation, Ablex gave a 6% raise to half its editors and a 1% raise the other half. In the following year with 9% inflation, Booklink gave a 15% raise to half its editors and a 10% raise to the other half. The following questions were posed in relation to this problem:

[10] For Problem 5, see Shafir, Diamond, and Tversky (1997, p. 361), where the problem and the results are cited from concurrent research work by Shafir and Thaler. We used the complete set of questions and results as reported in Shafir and Thaler (2006, p. 697). For Problem 6, see Shafir, Diamond, and Tversky (1997, p. 362).

[11] Shafir, Diamond, and Tversky (1997, p. 364).

Morale: *In which firm were there likely to be more morale problems?*

Job decision: *Which editor was more likely to leave for another job?*

	Morale
Ablex	60% (49%)
Booklink	7% (8%)
Same in both	33% (43%)
χ^{2*}	25.60
p-value	< 0.001
	Job decision
The editor who received the lower raise in Ablex	58% (57%)
The editor who received the lower in Booklink	17% (5%)
The two were equally likely	25% (38%)
χ^{2*}	17.50
p-value	< 0.001

Note: LLM percentages are aggregated across all responses for each scenario ($N = 60$). $s^2 = 0.88$ (Morale) and $s^2 = 0.73$ (Job decision). Results from Shafir, Diamond, and Tversky (1997) are in parentheses. *Goodness-of-fit χ^2 test against uniform distribution (33.3% each, $df = 2$).
χ^2 test of independence (scenario × response): $\chi^2(2) = 3.30, p = 0.192$.

When asked about morale, the majority of LLMs' responses (60%) predict greater discontent at Ablex, the firm offering smaller nominal pay rises. A third of responses (33%) claim that morale issues would be the same in both firms. For the job decision scenario, the LLMs' responses are more varied. While 58% identify the Ablex editor as more likely to leave, 17% select the Booklink editor, and 25% claim that both editors faced equal incentives to leave. These results are consistent with human responses relative to Ablex, although there are differences in preferences for the other two options. Notably, human responses in Shafir, Diamond, and Tversky (1997) are more likely than LLMs to claim that outcomes should be equivalent in both scenarios.

Conclusion

We conducted an exploratory replication study on money illusion in LLMs, using survey questions from Shafir, Diamond, and Tversky (1997). For some problems, our results tend to suggest that LLMs are less susceptible to money illusion than the human participants in the original study. This is particularly evident in the first scenario of Problem 1 and in Problems 2 and 5. However, we also found that LLMs exhibited qualitative response patterns similar to those of human participants, favouring the use of nominal terms to evaluate the proposed scenarios. This mainly occurred when individual characteristics, social judgements, or fairness assessments were more relevant to response selection, as in the last two scenarios of Problem 1 and in Problems 4 and 7. Notably, this also occurred in Problem 6, despite the scenario involving an explicit economic evaluation. Interestingly, some LLM responses correctly identified specific scenarios as instances of the money illusion and occasionally referenced the Shafir, Diamond, and Tversky (1997) study directly.

Overall, our exploratory evidence tends to suggest that LLMs can avoid the money illusion. However, it also tends to suggest that, when dealing with questions regarding money illusion, they exhibit decision-making patterns which can be related to those in the human-generated text data on which they are trained. While a systematic analysis based on a larger dataset is required for more definitive conclusions, we nonetheless hope that our results can be useful for future research analyses.

Acknowledgments

This research has benefited from the financial support of the Italian Ministry of University and Research (PRIN 20229-LRAHK).

References

Choi, I. and W. C. Kim (2023). Enhancing financial literacy in South Korea: Integrating AI and data visualization to understand financial instruments' interdependencies. *Societal Impacts 1*(1-2), 100024.

Conte, A., G. De Santis, G. Melissari, P. Paiardini, and J. Temperini (2024). Chronicles of choice. survey insights into intertemporal preferences. *Journal of Behavioral Economics for Policy 8*(S2), 19–27.

Corazzini, L., E. Deriu, and M. Guerzoni (2025). Decision and gender biases in large language models: A behavioral economic perspective. *arXiv preprint arXiv:2511.12319*.

de Moraes Ferreira, M., M. Y. T. Santiago, R. Bastos, D. Fatori, R. S. Borborema, L. Seda, and M. C. Batistuzzo (2024). Replication: The money illusion effect in a Brazilian sample and meta-analyses. *Journal of Economic Psychology 104*, 102744.

Durt, C. and T. Fuchs (2024). Large language models and the patterns of human language use. In M. Cavallaro and N. de Warren (Eds.), *Phenomenologies of the Digital Age*, pp. 106–121. New York: Routledge.

Hu, T., Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, and J. Roozenbeek (2025). Generative language models exhibit social identity biases. *Nature Computational Science 5*(1), 65–75.

Kotek, H., R. Dockum, and D. Sun (2023). Gender bias and stereotypes in large language models. In M. Bernstein, S. Savage, and A. Bozzon (Eds.), *Proceedings of the ACM collective intelligence conference*, pp. 12–24. Association for Computing Machinery.

Macmillan-Scott, O. and M. Musolesi (2024). (Ir)rationality and cognitive biases in large language models. *Royal Society Open Science 11*(6), 240255.

Mees, H. and P. H. Franses (2014). Are individuals in China prone to money illusion? *Journal of Behavioral and Experimental Economics 51*, 38–46.

Mortillaro, I. and M. L. Signore (2024). Financial trading is not just a gender-based difference issue. A critical investigation across market mechanisms. *Journal of Behavioral Economics for Policy 8*(2), 61–67.

Ross, J. A. (2024). Examining LLMs in economic settings. Master's thesis, Massachusetts Institute of Technology.

Salecha, A., M. E. Ireland, S. Subrahmanya, J. Sedoc, L. H. Ungar, and J. C. Eichstaedt (2024). Large language models show human-like social desirability biases in survey responses. *arXiv preprint arXiv:2405.06058*.

Schramowski, P., C. Turan, N. Andersen, C. A. Rothkopf, and K. Kersting (2022). Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence 4*(3), 258–268.

Shafir, E., P. Diamond, and A. Tversky (1997). Money illusion. *The Quarterly Journal of Economics 112*(2), 341–374.

Shafir, E. and R. H. Thaler (2006). Invest now, drink later, spend never: On the mental accounting of delayed consumption. *Journal of Economic Psychology 27*(5), 694–712.

T Elbæk, C., I. Uzodinma, Z. Ismagilova, and P. Mitkidis (2022). Suppetia ex machina: How can AI technologies aid financial decision-making of people with low socioeconomic status? *Journal of Behavioral Economics for Policy 6*(S1), 49–57.

Upravitelev, A. (2024). Fintech revolution's impact on financial behavior and education. *Journal of Behavioral Economics for Policy 8*(S2), 29–33.

Ziano, I., J. Li, S. M. Tsun, H. C. Lei, A. A. Kamath, B. L. Cheng, and G. Feldman (2021). Revisiting "money illusion": Replication and extension of Shafir, Diamond, and Tversky (1997). *Journal of Economic Psychology 83*, 102349.