Energy social surveys replicated with Large Language Model agents

Michael J. Fell, UCL Energy Institute, University College London, London, UK. michael.fell@ucl.ac.uk.

Abstract

Large Language Models (LLMs) are artificial intelligence systems trained to understand and predict human language. In this study I programmatically create numerous LLM agents with population-representative characteristics, and prompt them provide survey responses with the aim of replicating existing energy social survey findings. Three studies are replicated, yielding moderate to high degrees of fidelity to the original results. Potentially significant contributions of the approach include improving the efficiency of research by identifying most promising interventions before conducting human studies, and simulating input from harder-to-access populations. However, there are also important practical and ethical challenges requiring of careful consideration.

Keywords: large language models; artificial intelligence; energy; social surveys; replication

This is a non-peer-reviewed preprint uploaded on 6 January 2024.

Introduction

The potential of Large Language Models (LLMs) in contributing to scientific research has been widely discussed and demonstrated [1–3]. In the social sciences, potential contributions range from supporting analysis of various types of qualitative data [4], conducting interviews [5], informing hypothesis generation [6], and contributing to survey research [7]. However, their application in energy social studies has so far been extremely limited. With this study I attempt to replicate some of the findings of three survey-based studies using LLM-generated agents in place of human respondents. I draw on the experience and findings to suggest this and similar approaches could have significant impacts on the efficiency and nature of social research in energy, although it also comes with risks.

LLMs are a form of artificial intelligence (AI) model based on neural networks. Trained on large bodies of text of different kinds, they are effective natural language processors and can also generate coherent text. While the roots of LLMs in natural language processing stretch back decades, it is only in the last few years that their diverse capabilities have brought them to wider public attention. Prominent examples such as OpenAI's ChatGPT and Google's Bard now have hundreds of millions of users [8]. They are widely predicted to have substantial impacts on many sectors, including scientific research – both regarding productivity and the nature of work that is undertaken. Their application in the context of social surveys serves to illustrate this potential. Possible uses in this area include informing instrument design, sampling, response, data cleaning and analysis, and reporting [7].

Social surveys are a very widely used tool in energy research to collect structured quantitative and qualitative data from population samples, usually with the aim of making some inferences about the population itself. They permit easy experimental variation, as well as measurement of variables which are hard to capture in other ways, such as beliefs and opinions of large samples of individuals.

They are also often used to infer possible behaviour when this would be impossible or impractical through other means. For example, if the cost of directly observing a behaviour would be prohibitive, or if the behaviour relates to a product or service that is not available in the market – a common situation in an energy sector addressing low-carbon transition. Surveys are often an important prior step to more resource intensive field trialling of products, services, and communication approaches.

Previous work has explored the potential of LLMs to act as a sample of multiple individuals for the purpose of survey research. Argyle et al. (2023) [9] describe this as "silicon sampling", and argue that because it allows for agent characteristics to be represented in proportion with the general population, it can avoid some of the problems associated with algorithmic bias that may otherwise be present in the LLM. Using this approach to investigate political questions such as vote prediction, Argyle et al. [9] find strong alignment in patterns of association between human and GPT-3-generated responses. Similar approaches have been used to replicate well-established findings in psychology [10–12], management studies [13], and economics [14], amongst others. Aher et al. [10] introduce the idea of a Turing Experiment which, instead of focusing on the possibility of distinguishing between AI and individual responses, tests whether AI is able to faithfully simulate human behaviour as revealed at the group level in experimental settings. In the context of commercial user research, a variety of services are available which offer AI-user-based input, although these do not appear to include survey response [15].

Where LLMs have been used by energy researchers, they have largely focused in conceptually related areas such as systems modelling and technical data analysis [16,17]. While there are examples with a social science element, such the generation of building occupant personas [18], as yet their application in published social science research has been limited. To help assess their suitability and consider potential applications and challenges, I attempted to replicate parts of three studies. Two of these I originally co-authored, one on peer-to-peer (P2P) energy trading [19], the on energy retail models based on multiple suppliers [20]. The other [21] is an independent study investigating the role of defaults and trust in stated uptake of energy efficient appliances, and was selected for it recency, rigor, and difference in topic and setting.

Method

The LLM used was the OpenAI model GPT-3.5-turbo-1106, accessed via the OpenAI API. While not the most recent model (which is gpt-4-1106-preview at time of writing) it is still extremely powerful while being a tenth of the cost to use. Prompts submitted through the API consist of two parts, "system" and "user". The system text sets the behaviour of the LLM, while the user text provides the actual prompt to which the LLM will respond by way of an output. The system setting can be used to endow the LLM with a range a characteristics such as might be found in an individual survey respondent – for example, demographic, attitudinal, and personality characteristics. The user prompt can then be used to submit the survey question to the "respondent", and the response returned as an output and recorded. This is analogous to a response to a single survey question by a single respondent. To collect a response to the same question by a different respondent, a new API call is submitted with a newly generated combination of respondent characteristics in the system setting.

Prompt length is a significant consideration because it has implications for cost and speed of API calls. LLMs work based on tokens, which represent words or parts of words. The more tokens are sent in a prompt or received in an output, the longer the time required and the higher the cost. While the costs and time requirements are generally low for the basic exploratory work presented here, they could become significant for future possible uses discussed later. As such, I attempted to strike a reasonable balance between providing sufficient system/user prompt and output detail, and time/cost.

To create the sample and run the data collection I created a Python program called PolyPersona (available open source at https://github.com/mikefsway/polypersona). First, the program accesses a spreadsheet holding a range of respondent characteristic variables and their possible states. As an exploratory study based on trial and error, I included a range of standard demographic and attitudinal variables often or sometimes measured in surveys in energy, including age, gender, household income, household size, education, tenure, occupancy patterns, environmental concern, economic rationality, risk aversion, social trust, political orientation, and place attachment. I also included the “big 5” personality traits of openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. While these traits are not typically measured in energy-related surveys, they have been shown to have significant explanatory power in a range of areas of life [22,23] and are a way of enriching the variety of responses. Finally, I included a variable for the speed at which the survey was being completed. Respondents give differing levels of attention to surveys, so to simulate this a variable was included for either slow response, giving careful thought to the topic and drawing on all characteristics, and quick response prioritising personality traits and stressing the need for an instinctive reaction.

Demographic characteristics could take specific states, such as an age of 20, 30, 40, etc. or a household size of 4 or 5, and a probability for each state was assigned according to UK census data where available. Most attitudinal and personality characteristics were assigned states of either high, medium, or low, which could occur with equal probability. When PolyPersona generates a specific respondent persona, it probabilistically assigns a state of each characteristic to the respondent. This simplistic approach has the limitation of ignoring consistent associations between characteristics. For example, home ownership shows a clear tendency to increase with age [24], and there are likely many associations between different attitudinal and personality characteristics. Such associations could be explored and, if shown to be impactful, captured in future iterations. Having generated a range of characteristics, these are composed into a system prompt (see Box 1).

Box 1: Example system prompt generated by PolyPersona.

You are a UK householder completing a survey in a hurry, so rely more on fast, intuitive judgement. You are male, aged 70, with a medium household income, 3 people in your household, you highest level of education is 'further', you are at home most of time, and you are a social tenant. You have these attitudes: high environmental concern, low risk aversion, medium social trust, left of centre politics, high place attachment, high economic rationality, and your innovation adoption status is 'late majority'. Your personality has traits of low extraversion, medium agreeableness, high conscientiousness, high neuroticism, and medium openness to new experience.

The program then composes a user prompt, which consists of the survey question and instructions for the LLM on how to respond. The survey question is replicated as closely as possible from the original study, along with any necessary preliminary information. Experimental variations are included in a separate spreadsheet and selected probabilistically in the same way as for the respondent characteristics. This is then concatenated with the instruction on how to respond. LLMs work by sequentially generating the most likely next word – they "talk by talking". I therefore prompt the model to first provide a brief explanation of its reasoning relating to the question, followed by a consistent answer to the survey question (usually in the form of an integer). See the supplementary material for example user prompts for each replication.

Responses are returned in JSON format and added to a spreadsheet along with a record of the respondent characteristics. The program then loops and produces a new combination of characteristics and question variation, and so on until the desired number of responses has been obtained. This raw data can then be analysed according to the approach taken in the original study. Here, this usually consisted of a simple graphical representation, with the addition of some statistical analysis as described. The data generated for each of the replications is available at OSF (https://osf.io/g7ua8/).

Results

Study 1

The first study I attempted to replicate was Fell et al. (2019) [19], which explores consumer demand to participate in peer-to-peer (P2P) energy trading, a model whereby prosumers and consumers can sell and buy electricity directly between themselves. A UK nationally representative sample of 2064 respondents were shown details of a P2P scheme, with experimental variation of the scale (neighbourhood, city/region, national) and proportion of consumption supplied (25%, 50%, 75%), and asked if they would choose to participate. The study subsequently investigates the effect that scheme organiser and other characteristics have on demand, but I did not include these in this exploratory replication. This was the replication I used in development of the PolyPersona program, charts of preliminary runs are available in the supplementary materials.

Figure 1 shows the results of the replication based on 1798 PolyPersona responses. There appears to be moderate agreement with the original study. Demand is highest for P2P energy at the regional level in most consumption conditions, similar to the original study. Overall magnitude of demand is broadly similar for regional and neighbourhood conditions, but substantially lower in the replication for country-level supply. Some findings, such as the drop in demand for neighbourhood energy at 75% consumption, are not evident.

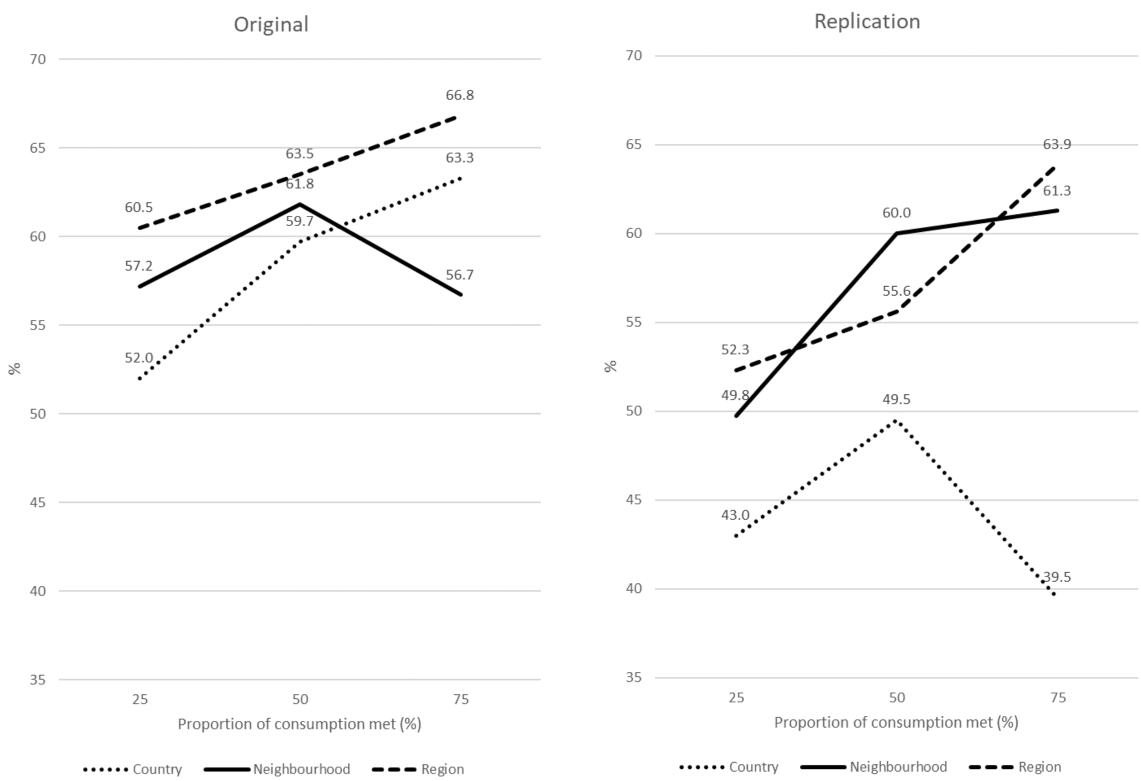


Figure 1: Percentage of respondents indicating willingness to participate in the P2P energy scheme, for the original study (left) and replication (right). Original data based on Fell et al., 2019.

The original study reported some statistically significant differences between the experimental groups. Binary logistic regression revealed significantly higher uptake at 75% consumption level compared to 50% (odds ratio [OR] = 1.318, $p = 0.02$). No such effect was evident for the replication, although uptake at 25% consumption was significantly less than at 50% (OR = 0.763, $p = 0.02$). For spatial scale, the original study found significantly less uptake at the neighbourhood (OR = 0.766, $p = 0.03$) and country (OR = 0.732, $p = 0.01$) levels as compared to region. For the replication, neighbourhood uptake was not significantly lower than regional, but country was (OR = 0.587, $p < 0.01$). The effects observed in the original study were therefore partially replicated.

As described in the methods, the LLM-based responses required the generation of an explanation prior to giving a participation decision. This has the co-benefit of making it possible to view a qualitative explanation of the rationale of each decision. Table 2 illustrates this with examples of a positive and negative participation decision, and shows how responses can vary depending on whether the LLM was instructed to respond at leisure or in a hurry.

Table 2: Examples of the explanations the LLM generated to inform its decisions.

Response speed	Decision to participate	Explanation
At leisure	Positive	As a homeowner with medium environmental concern and a general willingness to adopt new innovations, I would be open to the idea of buying energy directly from local homes and businesses with solar panels. I have a medium place attachment, so supporting local energy production aligns with my values.
	Negative	As a homeowner with medium risk aversion and low environmental concern, I prefer to stick with my current supplier for reliability and stability. Despite the cost savings, I value the consistency of my current energy supply over the potential benefits of a cheaper, decentralized option.
In a hurry	Positive	I prefer to support local energy initiatives and as an early adopter, I am open to trying new energy solutions. The slight cost saving and the community aspect appeal to me.
	Negative	I don't have high environmental concern, and I prefer to stick with what I know. Plus, I'm not usually at home to engage with my neighbours in this way.

Study 2

The second study I replicated was Watson et al., (2020) [20], which explores consumer demand for electricity retail models involving contracting with more than one supplier. Such models could unlock the ability of specialist suppliers to supply electricity only from, from example, local sources or for specific appliances. A representative sample of 1200 respondents were shown letters from energy suppliers asking them if they would like to sign up to a multiple supplier arrangement, meaning they could choose to be supplied with local electricity (when available) by a local supplier, with back-up provided by the original supplier. There were three conditions. The first framed the single original supplier option as the default, with respondents having the option to add the local supply. The second framed the multiple supplier option as the default, with respondents having the option to opt back to the single supplier. The third requested respondents to make an active choice between the two options.

Figure 2 shows the results of the replication based on 600 PolyPersona responses. The general trend across the conditions is the same, with most respondents choosing the multiple supplier option when it is the default, followed by the active choice, followed by the single supplier option default. The replication slightly underestimates multiple supplier take-up in the active choice and multiple supplier default conditions, and is almost exactly the same for the single supplier default.

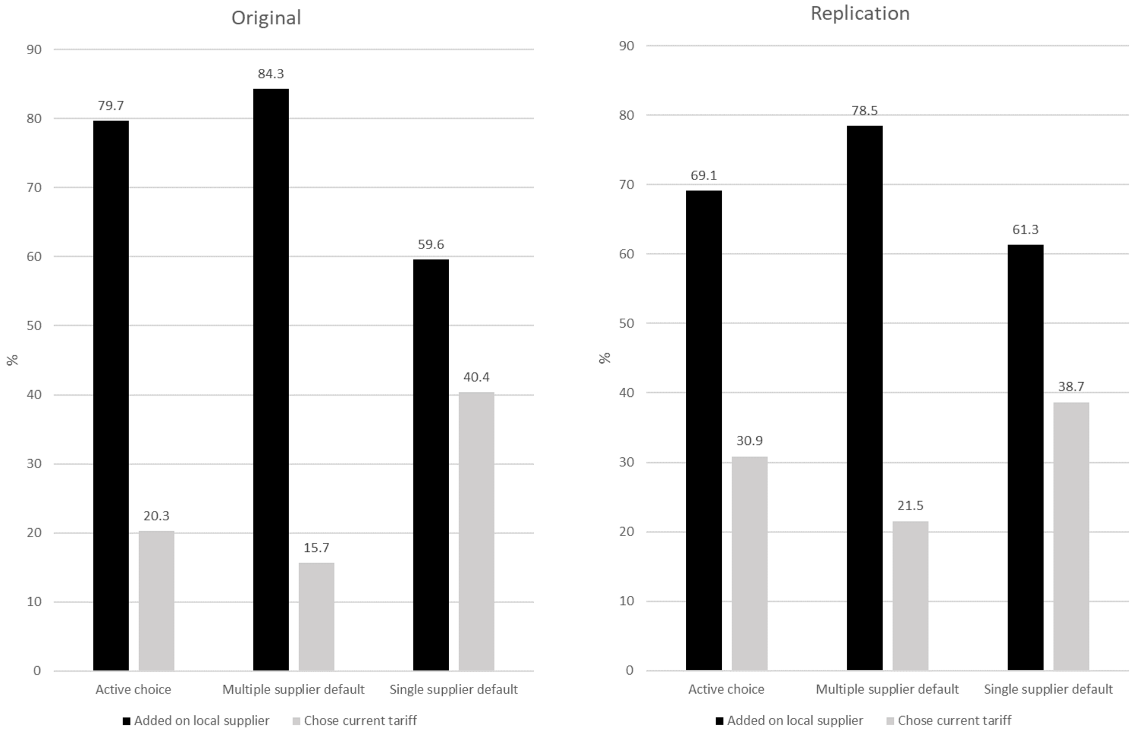


Figure 2: Percentage of respondents choosing a local supplier (multiple suppliers), or stick with their single (current) tariff, for the original study (left) and replication (right). Original data based on Watson et al., 2020.

A 2x3 Chi-squared test for the original study showed a statistically significant effect of condition on respondents' choices ($\chi^2 = 67.6$, $p < 0.01$). This was also the case the replication, albeit with a smaller effect size ($\chi^2 = 14.87$, $p < 0.01$). Pairwise Chi-squared tests in the original study found that the single supplier default group responses differed significantly from those of the other groups. This was partially reproduced in the replication, where the single and multiple supplier default groups differed significantly ($\chi^2 = 14.06$, $p < 0.01$), but the single supplier and active choice groups did not ($\chi^2 = 3.55$, $p = 0.06$). This may also have been a function of the smaller sample size.

I conducted a sensitivity analysis for this study, to explore the effect of including or excluding different sets of respondent characteristics (i.e. demographic, attitudes, personality), as well as survey completion speed. I ran ten variants in total. The first specified only that the respondent was "a UK householder completing a survey". Seven subsequent runs included all combinations of characteristic sets (i.e. demographics only, attitudes only, personality traits only, demographics and attitudes, etc.). The final two runs included all characteristic sets, with one run specifying that all were completed fast ("in a hurry, so rely more on fast, intuitive judgement"), and the other that all were completed slow ("at your leisure, so rely more on calculated, rational thought"). Figure 3 shows the results of the sensitivity analysis.

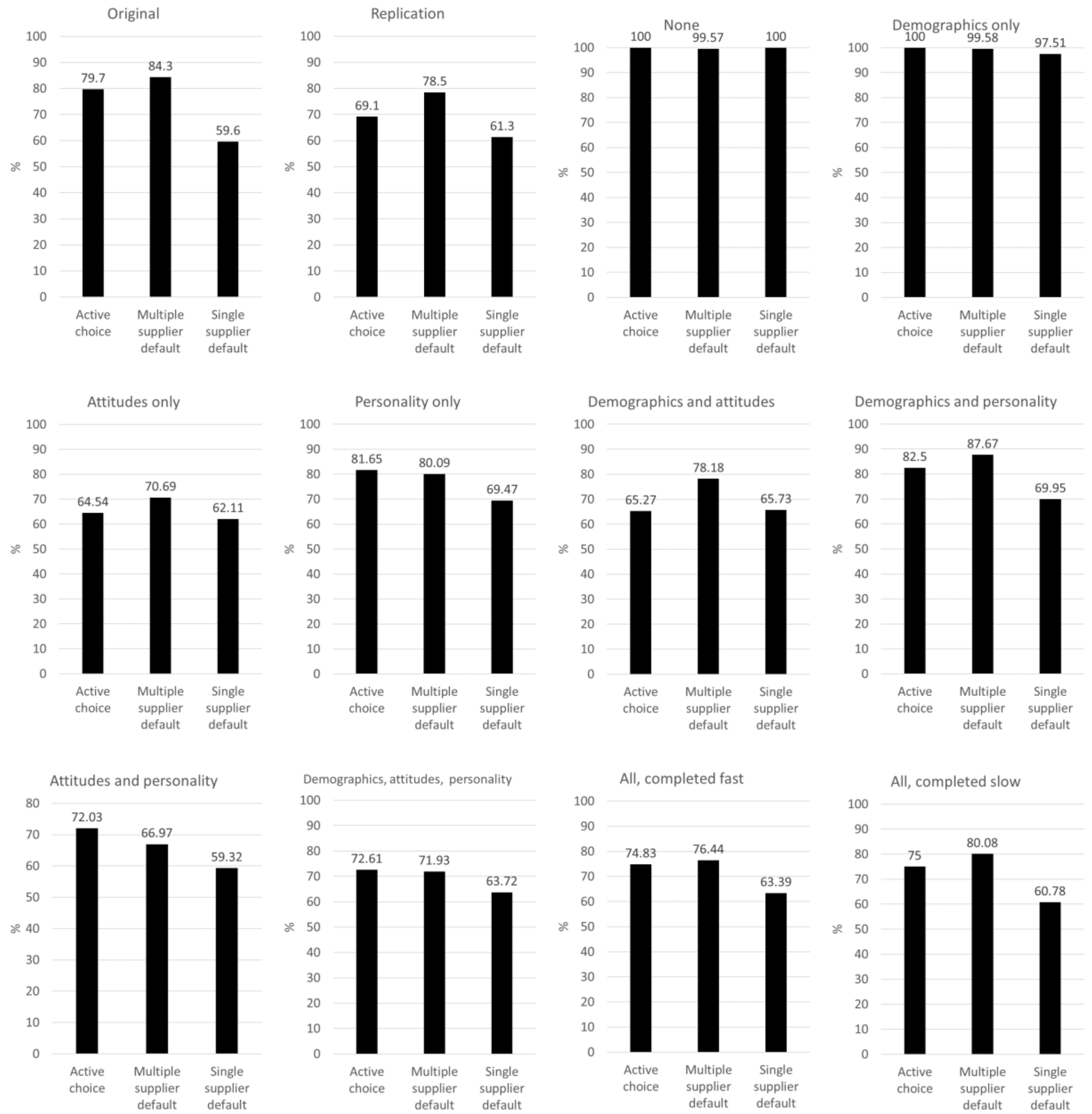


Figure 3: Sensitivity analysis of study 2. Bars indicate those respondents selecting the multiple supplier option in each case. The first two graphs show the original results and my main replication, for comparative purposes.

Providing the LLM with no varying characteristics, or varying only demographic factors, results in negligible outcome variation with practically all respondents selecting the multiple supplier option. All other combinations result in some level of variation. The combination with the closest pattern of resemblance to the original study includes all characteristics and slow completion, suggesting this could be a preferable option to the mix of fast (33%) and slow (67%) used in the main replications.

The original study yielded the unexpected finding that the preference for adding the local supplier was strong enough to override the default effect. It is striking that this finding is replicated here. For this study, to test the effectiveness of using a multiple-agent approach as compared to simply asking ChatGPT for a prediction, I submitted all three scenario descriptions directly to ChatGPT-4 in the form of a zero shot prompt, and requested it to predict the results of a nationally representative survey. Its prediction is shown in Figure 4 (see supplementary material for full prompt and response). It differs substantially from the actual findings, and includes the expected default effect.

While it is feasible that prompt chaining approaches could improve the likelihood of achieving a more faithful replication (see [6]), this illustrates the value of the multiple LLM agent approach.

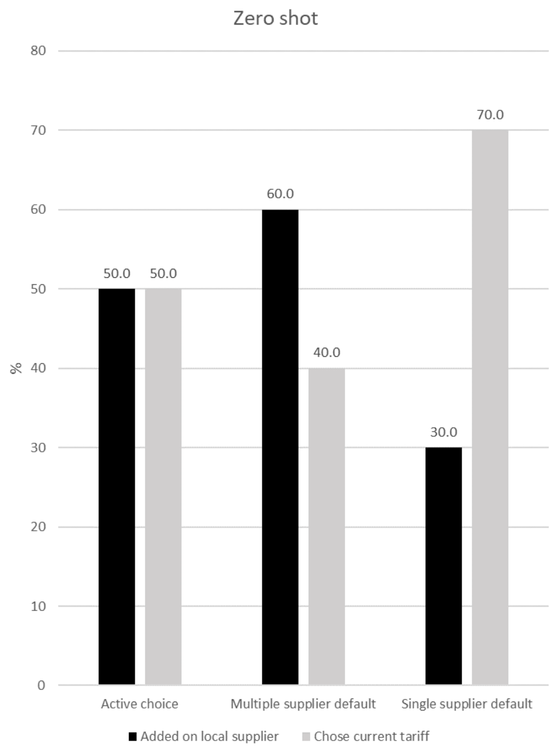


Figure 4: ChatGPT (GPT-4) prediction for the results of Watson et al. (2020) when prompted directly with a zero-shot prompt.

Study 3

The third and final study I selected for replication was Kuhn et al. (2023) [21]. I wanted to use a recent, rigorous study (which would not appear in the LLM training data), on a topic distinct from the previous two, conducted outside of the UK, and selected this paper based on searches on a convenience basis. It focuses on the role of defaults and trust in a housing provider to explore the effect of these factors on stated uptake of a range of energy efficient appliances. The 956 participants in the US branch of the survey were experimentally shown either positive or negative news headlines regarding real estate, and then presented with a choice task where either standard or efficient appliances were the default condition, with the outcome being the number of efficient appliances chosen.

Figure 5 shows the replication results for the replication of [21] based on 942 PolyPersona runs. Again, a broadly similar pattern of responses can be observed between the original and replication. A greater number of efficient appliances are chosen where efficient appliances are the default in both case, while promoting mistrust of the housing provider results in lower efficient appliance choice. The default effect is less pronounced in the replication, perhaps because it was not possible to use the pre-checked box approach as was the case in the original study.

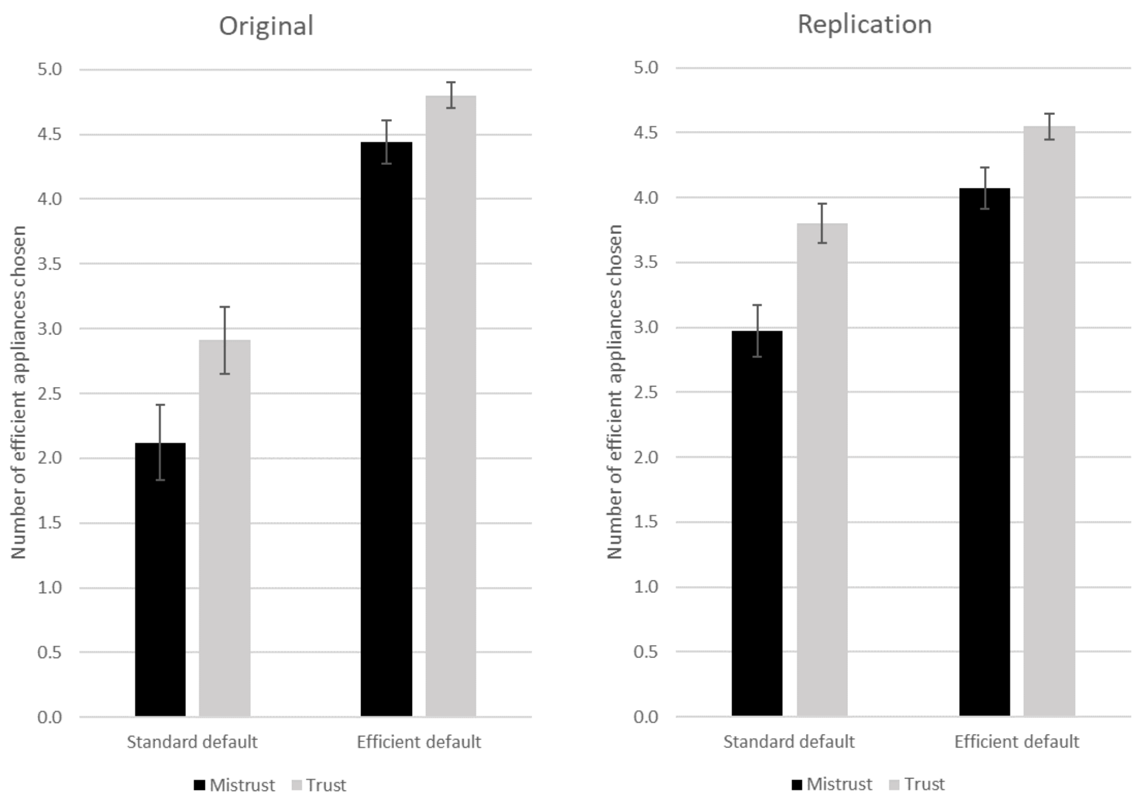


Figure 5: Number of energy efficient appliances chosen by respondents in each experimental condition, for the original study (left) and replication (right). Error bars show 95% confidence intervals. Original data based on Kuhn et al., 2023.

Multiple linear regression analysis with the dependent variable of number of efficient appliances chosen showed statistically significant positive main effects of default and trust, and a negative interaction effect, as detailed in Table 3. These statistically significant effects were also detected in the original study, although with a minor variation (the interaction effect was only significant when also controlling for covariates).

Table 3: Multiple regression results for replication of Kuhn et al., 2023. Adjusted R^2=.182.

Variable	B	Std Error	p
Efficient appliance default	0.923	0.077	< .001
Trust-promoting headlines	0.658	0.077	< .001
Interaction	-0.345	0.155	0.026

Discussion and conclusions

In this study, I used LLM-based agents to replicate the findings of previous social surveys in energy. The replication of [19] showed some similar trends to the original, such as a generally high acceptability of P2P schemes and a tendency to higher uptake of schemes operating at a regional level. Statistical test results were partially replicated. However, some key observations from the original study were not apparent. The replications of [20] and [21] showed results that were much more consistent with those of the original study, including capturing unanticipated results. Most statistically significant findings were replicated. Overall, similar or the same conclusions could have been drawn from the replications as the original studies. Viewed as a Turing Experiment [10], this attempt to replicate the behaviour of human samples using AI could therefore be considered reasonably successful.

Nevertheless, given evident variability and the fact that this work is still at an early developmental stage, it is far from possible to conclude that LLM-based agents could reliably replace human respondents at this time (nor would this be desirable in many cases – see discussion below). However, the results are promising enough to suggest that further exploration of possible applications of the approach in energy research is warranted (users can experiment with the PolyPersona tool available open source at https://github.com/mikefsway/polypersona).

First, it is important to stress that LLMs do not model human thought processes, but rather replicate a human-like response based on a vast body of training data. The explanation provided by the LLM, and its subsequent decision, is probabilistically selected from all possible responses based on the words produced so far and given the constraints presented [25]. It represents a highly sophisticated black-box method of text sampling. None of my analyses were preregistered, although I report the first and only replication attempt unless otherwise stated (i.e. replication of [19], and sensitivity analysis).

It is also necessary to highlight important ethical considerations surrounding the use of LLMs in this kind of research context. A key concern is around algorithmic bias, whereby biases present in training data are replicated in LLM outputs [2]. This is a nuanced problem. In many cases, surveys are explicitly attempting to understand the existence and implications of bias present in the population, so in this respect the presence of bias (if appropriately controlled) can be helpful – what Argyle et al. [9] refer to as algorithmic fidelity. It is striking that sensitivity analysis showed that varying only demographic characteristics led to negligible variation in results across condition in study. This may be due to developer efforts to mitigate bias based on demographic factors [26]. However, it still likely that systematic underrepresentation of certain viewpoints in training data could lead to misleading outputs. Similarly, certain biases of interest to researchers may not manifest in survey outputs unless they are explicitly endowed with them [14], as may have been the case for risk aversion as used here. A further ethical risk is the possibility that LLMs are used to produce fake survey responses, either by malicious respondents or by researchers themselves. However, the approach could also bring a range of benefits, including ethical ones.

The major advantage of "silicon sampling" is its cost- and time-effectiveness compared to human surveys. The full replications included here cost a US dollar or less each using the GPT-3.5 API – several orders of magnitude less than a comparable human survey – and depending on the number of participants could be run in a timeframe of minutes to hours. Using PolyPersona, it is quick and easy to set up and modify a variety of respondent characteristics, survey questions, and experimental conditions. As such, it is reasonable to expect the approach to be used at the scoping stage of research, for example to help identify conditions that have the biggest effects, and what the sizes of these effects are. Such approaches, if they provide reliable, could help improve research efficiency by focusing it in areas more likely to yield the biggest impacts. This has benefits to researchers, research funders, and wider society. By analysing the explanations returned, it is also possible to check if questions are being misinterpreted, and thereby make them clearer (if necessary) for a human audience. This could help act as a supplement to human pilot testing. The approach could also be helpful in producing pseudo data to help plan analyses before human data are collected.

It is possible that LLM-augmented approaches could be useful where samples of interest may be harder to access using standard survey techniques. For example, online surveys cannot represent the input of non-Internet users. While LLM training data is drawn from the Internet, this does not mean it is only representative of Internet users as a lot of content available online was previously available in non-Internet formats. Especially if combined with fine-tuning approaches (such as training on oral history databases), the approach could provide a powerful way to provide insight into underrepresented viewpoints. This is a nuanced question which opens an important debate, as there is also a risk that the approach could be used to substitute for genuine engagement with underrepresented groups.

Related to the question of sampling, the approach raises the intriguing possibility of being able to explore the impact of respondent characteristics which are otherwise hard to measure or select for. Personality traits, for example, while commonly measured in some disciplinary areas, are not routinely captured in energy-related surveys and would be quite resource-intensive to do so using established measurement scales. But it is straightforward to endow LLM-based agents with them, and it is possible that these or other uncommonly measured characteristics could have unexpectedly large impacts on outcomes. Preliminary LLM-based work could help researchers decide which might be the most important characteristics to measure and select for, given limited resources.

There may be circumstances where the use of LLM-based respondents in place of a human survey could be justified, although extreme caution should be exercised here. An example could be where there is no data on a topic of interest, no resources to conduct a survey, and a “best guess” is needed to facilitate further work. In such a case, an LLM-based approach may provide a starting point for further sensitivity analysis. Use of such an approach should always be clearly highlighted and justified, given the risk of reading too much validity into LLM-generated datasets that can realistically resemble human-generated datasets.

The above discussion sets out some possible applications of an LLM-based survey approach in energy social studies, but there are several specific areas where the work presented here could be developed. Further work could explore what affects whether and why different characteristics are prioritised in formulating explanations and decisions. Combinations of characteristics could be identified that maximise general applicability with efficiency of token use. The impact of fine tuning could be investigated, especially in the case of harder-to-access samples (as mentioned above) or more niche topics. There is also a question as to whether simulating survey responses is the most useful end point. Surveys have well-known limitations, especially where the research interest is in behaviours where what is done (revealed preference) may differ significantly from what is said (stated preference). It would be fruitful to explore whether LLM agents could be used in such a way as to better reflect real behaviours in relation to energy.

Acknowledgements

This work was supported by funding from the UKRI Energy Demand Research Centre (grant number EP/Y010078/1). I gratefully acknowledge the authors and co-authors of the replicated studies. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence (where permitted by UKRI, ‘open government licence’ or ‘creative commons attribution no-derivatives (CC BY-ND) licence’ may be stated instead) to any author accepted manuscript version arising.

Data availability

Datasets for the replications are available at https://osf.io/g7ua8/.

References

1. Boyko J, Cohen J, Fox N, Veiga MH, Li JI-H, Liu J, et al. An Interdisciplinary Outlook on Large Language Models for Scientific Research. arXiv; 2023. doi:10.48550/arXiv.2311.04929

2. Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, et al. Using large language models in psychology. Nat Rev Psychol. 2023;2: 688–701. doi:10.1038/s44159-023-00241-5

3. Ziems C, Shaikh O, Zhang Z, Held W, Chen J, Yang D. Can Large Language Models Transform Computational Social Science? Comput Linguist. 2023; 1–53. doi:10.1162/coli_a_00502

4. Bano M, Hoda R, Zowghi D, Treude C. Large language models for qualitative research in software engineering: exploring opportunities and challenges. Autom Softw Eng. 2023;31: 8. doi:10.1007/s10515-023-00407-8

5. Chopra F, Haaland I. Conducting Qualitative Interviews with AI. Rochester, NY; 2023. doi:10.2139/ssrn.4583756

6. Rosenbusch H, Stevenson CE, van der Maas HLJ. How Accurate are GPT-3’s Hypotheses About Social Science Phenomena? Digit Soc. 2023;2: 26. doi:10.1007/s44206-023-00054-2

7. Jansen BJ, Jung S, Salminen J. Employing large language models in survey research. Nat Lang Process J. 2023;4: 100020. doi:10.1016/j.nlp.2023.100020

8. Carter R. Google Bard vs ChatGPT: Battle of the AI Chatbots. In: UC Today [Internet]. 29 Dec 2023 [cited 4 Jan 2024]. Available: https://www.uctoday.com/unified-communications/google-bard-vs-chatgpt-battle-of-the-ai-chatbots/

9. Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D. Out of One, Many: Using Language Models to Simulate Human Samples. Polit Anal. 2023;31: 337–351. doi:10.1017/pan.2023.2

10. Aher G, Arriaga RI, Kalai AT. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. arXiv; 2023. Available: http://arxiv.org/abs/2208.10264

11. Jiang H, Zhang X, Cao X, Kabbara J. PersonaLLM: Investigating the Ability of Large Language Models to Express Big Five Personality Traits. arXiv; 2023. doi:10.48550/arXiv.2305.02547

12. Park PS, Schoenegger P, Zhu C. Diminished Diversity-of-Thought in a Standard Large Language Model. arXiv; 2023. doi:10.48550/arXiv.2302.07267

13. Suguri Motoki FY, Monteiro J, Malagueño R, Rodrigues V. Charting New Territory: Using ChatGPT to Enhance Survey Instruments for Management Accounting Research. Rochester, NY; 2023. doi:10.2139/ssrn.4595896

14. Horton JJ. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? arXiv; 2023. doi:10.48550/arXiv.2301.07543

15. Arora A. 6 AI tools for user research. In: Bootcamp via Medium [Internet]. 2 Oct 2023 [cited 4 Jan 2024]. Available: https://bootcamp.uxdesign.cc/5-ai-tools-for-user-research-3365891a78c3

16. Buster G. Large Language Models (LLMs) for Energy Systems Research. National Renewable Energy Laboratory (NREL), Golden, CO (United States); 2023 Nov. Report No.: NREL/PR-6A20-87896. Available: https://www.osti.gov/biblio/2216925

17. Zhou M, Li F, Zhang F, Zheng J, Ma Q. Meta In-Context Learning: Harnessing Large Language Models for Electrical Data Classification. Energies. 2023;16: 6679. doi:10.3390/en16186679

18. Anik SMH, Gao X, Meng N. Towards automated occupant profile creation in smart buildings: A machine learning-enabled approach for user persona generation. Energy Build. 2023;297: 113485. doi:10.1016/j.enbuild.2023.113485

19. Fell MJ, Schneiders A, Shipworth D. Consumer Demand for Blockchain-Enabled Peer-to-Peer Electricity Trading in the United Kingdom: An Online Survey Experiment. Energies. 2019;12: 3913. doi:10.3390/en12203913

20. Watson NE, Huebner GM, Fell MJ, Shipworth D. Two energy suppliers are better than one: Survey experiments on consumer engagement with local energy in GB. Energy Policy. 2020;147: 111891. doi:10.1016/j.enpol.2020.111891

21. Kuhn S, Thøgersen J, Kutzner F. No trust in the choice architect? No problem! On the minor role of trust for the effectiveness of default interventions promoting the choice of energy-efficient appliances. J Environ Psychol. 2023;91: 102115. doi:10.1016/j.jenvp.2023.102115

22. Mammadov S. Big five personality traits and academic performance: A meta-analysis. J Pers. 2022;90: 222–255. doi:10.1111/jopy.12663

23. Zell E, Lesick TL. Big five personality traits and performance: A quantitative synthesis of 50+ meta-analyses. J Pers. 2022;90: 559–573. doi:10.1111/jopy.12683

24. Statista. England: homeowners by age 2022. In: Statista [Internet]. 2023 [cited 2 Jan 2024]. Available: https://www.statista.com/statistics/321065/uk-england-home-owners-age-groups/

25. Reynolds L, McDonell K. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. arXiv; 2021. doi:10.48550/arXiv.2102.07350

26. Ferrara E. Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. First Monday. 2023 [cited 13 Jan 2024]. doi:10.5210/fm.v28i11.13346

Supplementary material for “Energy social surveys replicated with Large Language Model agents”, Fell 2024

Preliminary runs of the replication of Fell et al., 2019

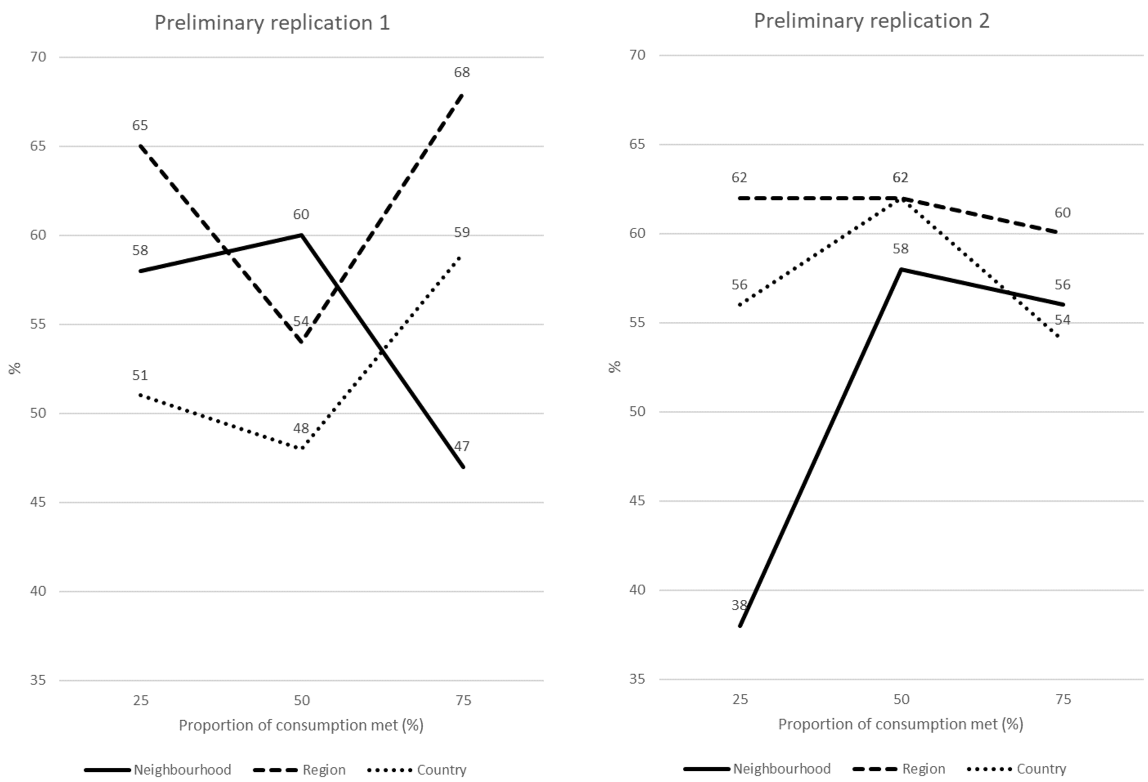


Figure S1: Test runs of the first replication. Preliminary replication 1 had 50 responses per experimental condition, while preliminary replication 2 had 100 responses.

Example user prompts

The following are example user prompts created for use in the replications. Italics are used to differentiate the response instructions from the survey question. For all experimental conditions, see https://osf.io/g7ua8/.

For the replication of Fell et al., 2019:

A new energy offer is available which allows you to buy energy directly from homes and businesses with their own solar panels. Energy bought this way is slightly cheaper than what you get from your usual supplier. If you participate, you could meet around a quarter of your household’s electricity needs through the offer. You would buy electricity directly from homes and businesses located on the streets in your neighbourhood. You would continue to buy the rest of your energy from your current supplier. Would you sign up to participate in this offer if it was available to you today? *Respond in JSON format. The first JSON object should be a short (<50 word) explanation of your reasoning, called 'explanation', drawing on your demographic, attitudinal, and personality* *characteristics. Then, output your decision on whether or not you would participate, in a JSON object called 'decision', with response options yes=1 or no=0 (return an integer). Your decision must be consistent with your explanation. Example output as follows: 'explanation': 'explanation text here', 'decision': integer*

For the replication of Watson et al., 2020:

Imagine that you receive the following letter from your current electricity supplier. Dear customer, We are working in partnership with LocalEnergy, a new local energy company. You are currently enrolled on our single-provider service. This means that all your electricity will continue to be provided by us. Alternatively, you can switch to the new combined-provider service, which is offered in partnership with LocalEnergy. Customers of the combined-provider service will get most of their electricity from locally produced, renewable sources, such as solar and wind farms. If there is not enough local energy to cover your needs, we will act as a back-up, so that you will always have the power you need. The price of each service will be the same as your current tariff. If you would like to stay with the single-provider service, you do not need to do anything. If you would rather change to the combined-provider service, you need to go to our website, log in to your account, and check a box indicating that you would like to switch. It will take a few days to switch. You can switch back at any time. What would you choose to do in this scenario? Stay with the single-provider service, or switch to the combined-provider service. *Respond in JSON format. The first JSON object should be a short (<50 word) explanation of your reasoning, called 'explanation', drawing on your demographic, attitudinal, and personality characteristics. Then, output your decision on whether you would choose the single or combined provider option, in a JSON object called 'decision', with response options combined=1 or single=0 (return an integer). Your decision must be consistent with your explanation. Example output as follows: 'explanation': 'explanation text here', 'decision': integer*

For the replication of Kuhn et al., 2023:

Imagine you consider renting a newly constructed apartment rented out by the Better Living Real Estate Group. Right now, this is your only available option. You have therefore decided to take it and signed the contract. Better Living has a programme that supports your move to the new apartment and is asking you to fill out an online form in which you choose amenities for your new home. Before you fill out the form you come across some recent headlines from news websites that could be relevant in this situation. Please read these headlines carefully: ‘Award Winners: Affordable high-quality housing and community project by Better Living wins the 2021 American Residential Real Estate’; ‘More transparency in real estate: Memorandum passed to combat corruption in U.S. housing sector’; ‘Buyers, Notice: Report shows real estate sector can be trusted’; ‘The rise of the housing market: What everyone gets wrong about housing in the U.S. Today, most public housing developments do, in fact, deliver on their promise of providing adequate, affordable housing to people in need, but they remain stigmatised because of the negative associations with distressed public housing’; ‘Americans deserve not just better, but the best: Administration is seeking largest ever investment in building resilient, quality houses’. Consider how interesting you perceive each article to be for your situation. Now read this letter from Better Living Real Estate Group: ‘Dear tenant, We are glad that you chose to rent with us. We want to make moving in to your new apartment as easy and satisfying as possible for you. We have therefore preinstalled the following amenities in your new home. They are standard amenities, but if you want we can replace them with energy-efficient alternatives. Compared to the standard amenities, the energy-efficient amenities cost more, but due to the savings on the electricity bill, the operating costs are lower. We estimate that the energy-efficient amenities save the equivalent of the initial higher costs in about three years. If you want to keep the standard amenities, you do not need to do anything. If you want us to replace one or more with energy-efficient alternatives, you can let us know.’ The appliances are as follows: freezer; refrigerator; washing machine; electric oven and stove; air conditioner. *Respond in JSON format. The first JSON object should be a short (<25 word) summary of the newspaper headlines, called 'summary'. Then, in a JSON object called 'better living', consider the headlines and state to what extent you trust that Better Living has the customers' best interests in mind, providing your answer as an integer on a scale from 0 to 10, where 0 means 'do not trust at all' and a 10 means 'completely trust'. The next object should be a short explanation (<50 word) explanation of your reasoning for your response to the Better Living Real Estate Group letter, called 'explanation', drawing on your demographic, attitudinal, and personality characteristics, and bearing in mind the headlines. Then, considering each amenity separately, output your decision on whether you would like the standard or efficient version, in JSON objects called ‘freezer’; ‘refrigerator’; ‘washing machine’; ‘electric oven and stove’; and ‘air conditioner’, with response options efficient=1 or standard=0 (return an integer). Example output as follows: 'summary': 'headline summary text here', 'better living': integer, 'explanation': 'explanation text here', ‘freezer’: integer, ‘refrigerator’: integer, ‘washing machine’: integer, ‘electric oven and stove’: integer, ‘air conditioner’: integer.*

Regression table for replication of Fell et al., 2019

Table S1: Results of binary logistic regression for replication of Fell et al., 2019.

Model	Deviance	AIC	BIC	df	X^2	p
H0	2487.213	2489.213	2494.708	1797		
H1	2452.985	2462.985	2490.457	1793	34.228	< .001

	Estimate	Standard Error	Odds Ratio	z	Wald Statistic	df	p
(Intercept)	0.382	0.106	1.465	3.588	12.872	1	<0.001
25% consumption	-0.27	0.117	0.763	-2.309	5.331	1	0.021
75% consumption	0	0.118	1	<0.001	<0.001	1	1
neighbourhood	-0.01	0.117	0.99	-0.084	0.007	1	0.933
country	-0.533	0.116	0.587	-4.595	21.113	1	<0.001

Direct prompting of ChatGPT as part of replication of Watson et al., 2020

Here are the GPT-4 prompt and response used to compare the multiple agent approach used in replications with a direct prompt.

Prompt: I am running a survey. It presents each participant with one of the following three scenarios at random: 1. "Imagine that you receive the following letter from your current electricity supplier. Dear customer, We are working in partnership with LocalEnergy, a new local energy company. You are currently enrolled on our single-provider service. This means that all your electricity will continue to be provided by us. Alternatively, you can switch to the new combined-provider service, which is offered in partnership with LocalEnergy. Customers of the combined-provider service will get most of their electricity from locally produced, renewable sources, such as solar and wind farms. If there is not enough local energy to cover your needs, we will act as a back-up, so that you will always have the power you need. The price of each service will be the same as your current tariff. If you would like to stay with the single-provider service, you do not need to do anything. If you would rather change to the combined-provider service, you need to go to our website, log in to your account, and check a box indicating that you would like to switch. It will take a few days to switch. You can switch back at any time. What would you choose to do in this scenario? Stay with the single-provider service, or switch to the combined-provider service." 2. "Imagine that you receive the following letter from your current electricity supplier. Dear customer, We are working in partnership with LocalEnergy, a new local electricity company. You have automatically been switched to our new combined-provider service, which is offered in partnership with LocalEnergy. Customers of the combined-provider service will get most of their electricity from locally produced, renewable sources, such as solar and wind farms. If there is not enough local energy to cover your needs, we will act as a back-up, so that you will always have the power you need. Alternatively, you can switch back to our single-provider service. This means that all your energy will continue to be provided by us. The price of each service will be the same as your current tariff. If you would like to stay with the combined-provider service, you do not need to do anything. If you would rather change back to the single-provider service, you need to go to our website, log in to your account, and check a box indicating that you would like to switch. It will take a few days to switch. You can switch back at any time. What would you choose to do in this scenario? Stay with the combined-provider service, or switch back to the single-provider service.". 3. "Imagine that you receive the following letter from your current electricity supplier. Dear customer, We are working in partnership with LocalEnergy, a new local energy company. You now have the choice of two services: 1) Single-provider: This means that all your electricity will be provided by us. 2) Combined-provider: This service is offered in partnership with LocalEnergy. Customers of the combined-provider service will get most of their electricity from locally produced, renewable sources, such as solar and wind farms. If there is not enough local energy to cover your needs, we will act as a back-up, so that you will always have the power you need. The price of each service will be the same as your current tariff. You need to go to our website, log in to your account, and check a box indicating which service you would like to choose. You can switch at any time. Which service would you choose in this scenario? The single-provider service or the combined-provider service." Imagine I administer this survey to a representative sample of 1200 UK citizens. Please consider what you think will affect the results, then predict what proportion of people will choose each outcome option in each scenario, and create a bar chart.

Response: To predict the outcomes of your survey with these three scenarios, we need to consider several factors that might influence the decisions of the participants:

1. Status Quo Bias: People have a tendency to prefer the current state of affairs. In scenarios where a default option is presented (Scenario 1 and 2), many might choose to stick with the default due to inertia or the perceived effort of changing.

2. Environmental Concerns: Given the emphasis on renewable energy in the scenarios, individuals who are more environmentally conscious may be more inclined to choose the combined-provider service.
3. Perceived Effort and Convenience: The effort required to switch services (or stay with the current one) might influence decisions. If switching seems cumbersome, people might avoid it.
4. Trust in New Services: People's trust in the new combined-provider service, especially regarding its reliability and the partnership with LocalEnergy, will play a role.
5. Price Sensitivity: Since the price remains the same for both services, this factor might not heavily influence the decision.

Based on these considerations, here are the predictions:

1. Scenario 1 (Default is single-provider): A significant number of people might stay with the single-provider service due to status quo bias. However, some environmentally conscious individuals might switch to the combined-provider service.
2. Scenario 2 (Default is combined-provider): A large portion might stay with the combined-provider service, again due to status quo bias. The effort to switch back to the single-provider might deter changes.
3. Scenario 3 (No default, active choice required): This scenario is likely to show a more balanced distribution, as people are required to make a deliberate choice without a default option.

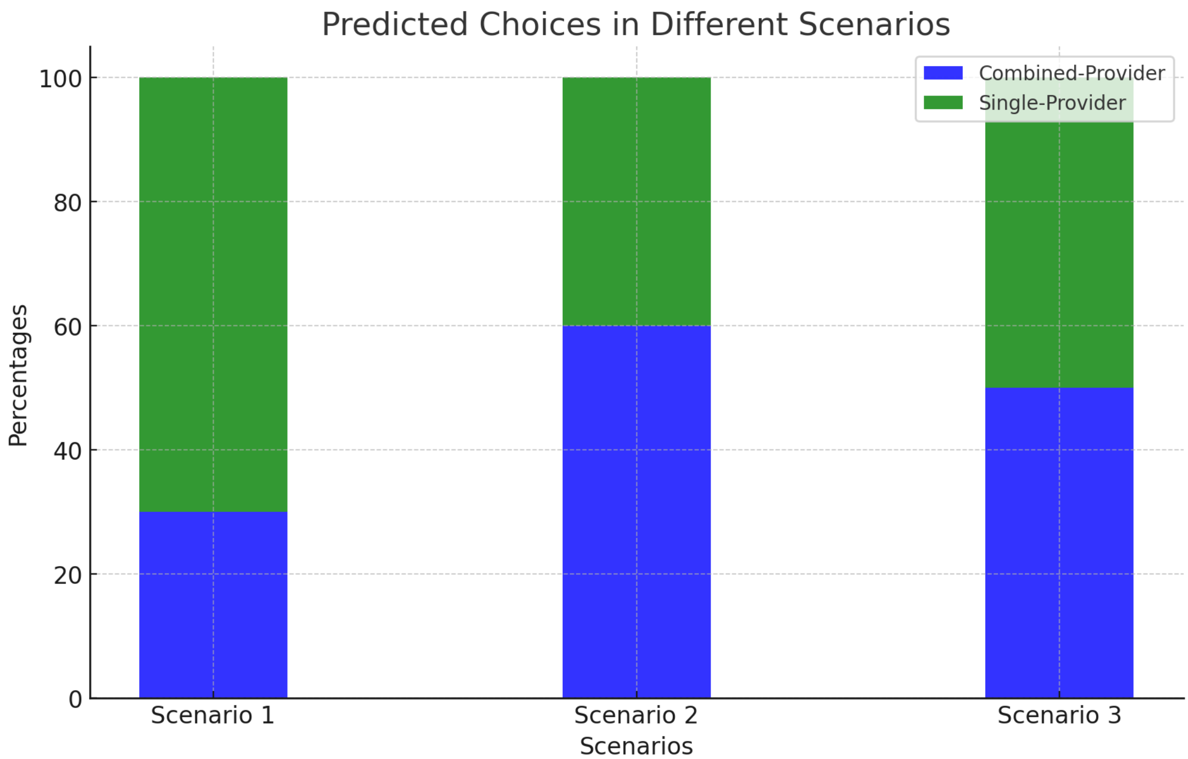
Now, I'll create a bar chart based on these predictions. Let's assume the following proportions for each scenario:

Scenario 1: 30% choose combined-provider, 70% stay with single-provider.

Scenario 2: 60% stay with combined-provider, 40% switch back to single-provider.

Scenario 3: 50% choose each service.

Let's generate the bar chart with these assumptions.



Here is the bar chart based on the predicted outcomes for each scenario in your survey:

Scenario 1: It shows that 30% of participants are predicted to choose the combined-provider service, while 70% would stay with the single-provider service.

Scenario 2: Here, 60% are expected to stay with the combined-provider service, and 40% might switch back to the single-provider service.

Scenario 3: This scenario predicts an even split, with 50% choosing each service.

These predictions are based on the factors discussed earlier, like status quo bias, environmental concerns, perceived effort, and trust in new services. Keep in mind that the actual results might vary based on specific attitudes and preferences of the surveyed UK citizens.