LLMs as Proxy Survey Participants with RAG

Anonymous ACL submission

Abstract

We explore how LLMs can be employed as proxies for humans in surveys, by encoding personas via individuals' chat data to simulate their response patterns. We observe promising mimicking capabilities of LLMs, and although performance varies by survey and subject, we suspect it depends on the specificity of the survey and available chat data. Our study suggests that some LLMs can replicate survey participants more precisely than our naive guessing methods, when leveraging chat data via RAG.

We test on two surveys covering notably different thematic scopes (broad and narrow): an OCEAN personality test and a Kano model survey about video game preferences. In the OCEAN survey, the LLMs consistently perform better than naive guessing benchmarks, meanwhile results are inconclusive for the Kano survey.

1 Introduction

Are Large Language Models (LLMs) only able to impersonate generic personas, or can they adopt detailed consumer profiles? We investigate whether LLM's role-playing capabilities make them a useful source of synthetic survey data. The objective of our experiment is to see if LLMs can be employed as proxy humans by market researchers, and thereby disrupt conventional survey-based marketing's speed, cost, and exhaustive sample limitation.

The human element is inherently a limitation for survey facilitators. Although costs and speed can be trivial to larger operations, the inability to recreate initial impressions is a resource-independent constraint. The utility would be an inexhaustive way to test which exposure resonates best with the target in order to provoke a desired reaction. Naturally, market research would also become faster and cheaper.

Leveraging LLMs with a Retrieval Augmented Generation (RAG) memory system offers a compelling solution to the challenges addressed. LLM's stateless inference ensures a "reset" of previous exposures to provoke genuine reactions to subsequent feature implementations, and RAG allows researchers to "introspectively" analyze these reactions. This ability to control and analyze the LLM's "memory" provides a level of experimental control impossible with human subjects. In other words, it allows researchers to get a second chance at first impressions. We provide an LLM with chat data of a given subject, to see how much the LLM's response resembles the actual responses of the subject.

1.1 Research Question and Hypothesis

Research Question: "How well can an LLM replicate the survey responses of an individual when induced with their chat data?"

Hypothesis: Providing an LLM with an individual's chat data; an LLM can replicate the survey answers of that individual better than naive guessing methods (always pick middle and base model w/o persona encoding).

2 Related Work

LLMs' role playing abilities has been explored within various domains. One of the most remarkable pioneering studies is Park et al.'s *Interactive simulacra of human behavior*, where backstories and an advanced RAG memory system allow the LLMs to adopt social agency (Lewis et al., 2020).

In subsequent work, Brand et al. (2023) show that LLMs adhere to economic theory about willingness to pay – a well-established property of consumer demand (Varian, 2010). Dillion et al. (2023) showcase a 0.95 correlation with humans on moral judgements across 364 publicly available scenarios. Aher et al. (2023) reproduce human behavior in classic experiments as a Turing test with pass rates of 51-99.5%. Horton (2023) successfully achieve

qualitatively similar results to that of humans in economic experiments. In Wang et al.'s (2024b) psychological interviews, role playing LLMs had up to 80.7% alignment with the human-perceived personalities of widely-known characters (provided with a description of them as system prompt).

More native to market research, scholars replicate preexisting experiments conducted with real humans, by introducing basic persona (Rind, n.d.) characteristics of a generic target subject. For example: When providing age, gender, and income, Li et al. (2023) finds agreement rates over 75% with humans in a consumer perceptual analysis replication (Keller, 1993); Wang et al. (2024a) discover that LLMs can misrepresent in-group heterogeneity more than real humans (providing four demographic axes); and Argyle et al. (2022) create backstories based on the five demographic axes (politics, race, gender, age, social class) and finds LLMs to be "efficient" proxies of varied sub-populations for social science research.

Surveys like OCEAN (Johnson, 2014) is used in plenty of the research. Some use it for evaluation: Serapio-García et al. (2023) use it in conjunction with another psychometric test to assess the consistency of a model's perceived personality. Lu et al. (2024) argue personality is determined by prompting, however Li et al. (2024) manipulate personality traits on a token-level in the decoding phase. Jiang et al. (2023) employ personality prompting by translating an OCEAN dimension to a description of person. Our experiment goes by the reverse order, using the text messages to indirectly induce the personality of the subject (Brown et al., 2020), which we afterwards test in the OCEAN survey. In contrast, we did not find academic literature where Kano surveys (Noriaki Kano, 1984) is used with LLMs.

Prompting literature that we apply: Making the LLM simulate instead of classify, as Aher et al. (2023) discovered; Wang et al.'s (2024b) experience on using *expert rating*; And avoiding misleading clues that can summon intrinsic character knowledge associated with names (Lu et al., 2024).

Evaluation is, by us, centered around alignment (others also propose [internal] consistency measures). While the most common quantitative approach is to benchmark against existing datasets with humans (Jiang et al., 2023; Aher et al., 2023; Argyle et al., 2022; Brand et al., 2023; Li et al., 2023; Horton, 2023), it is limited to the participants in the original survey and how well their personas has been described (e.g., the amount of demographic axes). The subjects data is used with their survey answers, thus persona encoding is inseparable from evaluation for creating any insights. Alternatively Jiang et al. (2023) qualitatively evaluate how well OCEAN psychometrics is induced with a human vignette test, and Lu et al. (2024) even use LLMs as judges of quality. Since evaluation precedence has yet to be established we suggest a new quantitative method in the experiment section.

In summary, most of the work resembling real humans use relatively shallow personas with five or less explicit demographic axes. We explore the gap between the nuanced characters of Park et al. (2023) and the generic ones present in more commercially oriented work.

3 Experiment

This section introduce the variables, RAG memory system, and evaluation metrics used to assess the LLM's ability to impersonate survey respondents.

3.1 Variables, Values, and Configurations

Our configurations range over the following variables:

- 2 Surveys: OCEAN personality, and Kano video game preferences (Barsalou, 2023).
- 2 Subjects: Authors L and S providing 900,000 and 60,000 tokens of chat data, respectively.
- 2 Retrieval Methods: "Dynamic" (query per question), "static" (fixed query per survey).
- 3 Context sizes: 1-chunk, 4000-, 7500 tokens.
- 3 LLMs: Llamma3-70b, -8b, Mixtral8x22b.

It should be noted that the author with most chat data also, anecdotally, is more engaged with gaming; presumably including more clues to video game preferences in their chat data.

We construct a total of 24 unique prompts that each LLM is running inference on (72 variable combinations). We also include six "base" configurations ($2surveys * 3LLMs$) without persona encoding for comparison. All 78 configuration's measured performance is the average of three simulations, for a total of 234 simulations.

```
systemMsg("You are participating in a survey. You will be presented with a series of questions about your {SURVEY}.", f"\nYou must choose answer to the question below with one of the five options: {', '.join(surv.POSSIBLE_ANSWERS)}. The answer must only contain the chosen option. "),
assistantMsg('Understood. I will answer the question below with one of the given options.'),
userMsg(question,f"\nYour choice: ")
```

Figure 1: Prompt Template w/o Chat Data (N_{conf}=6)

```
systemMsg("\\n".join([f"You are an expert actor, specializing in impersonation of non-famous people. You will be presented to the subject through explicit datapoints of their digital footprint. In addition, you will deduct their implicit {SURVEY} by shadowing chats between the subject and friends. You will be asked to fully immerse yourself in the role, and answer questions from the point of view of the persona. \\n#Context \\n##Chat conversations between the subject and their friends:\\n", "\\n\\nNEW CONVERSATION:\\n".join(chunks_most_similar)])),
assistantMsg("Understood. I will answer from the point of view of the persona, based on what I could the deduct from the text provided."),
userMsg("\\n".join([f"Persona is questioned about their {SURVEY} in {METHOD}. The persona must choose an appropriate answer to the question below with one of these five given options: {', '.join(surv.POSSIBLE_ANSWERS)}. Persona's answer must only contain the chosen option, without any elaboration, nor introduction.\\n\\n**Your question is:**\\n", question, "\\nThe persona chooses:"]))
```

Figure 2: Prompt Template w/ Chat Data N_{conf}=72

3.2 RAG Memory System

As clues to how the subject would respond to a given query, we provide 1-on-1 text communication between the subject and multiple different friends to give a general persona-portrait. Retrieval Augmented Generation (RAG) enable the subject's chat data to be used as in-context examples (Brown et al., 2020; Radford et al., 2018), without surpassing context length restrictions (Lewis et al., 2020; Hsieh et al., 2024).

Messages are sequentially grouped into coherent chunks (size: 75, overlap: 3) to preserve context and minimize noise. These chunks are then embedded into high-dimensional vectors using the "nomic-embed-text" model (768 dimensions).

During inference, the LLM retrieves relevant chunks based on cosine similarity between the vector of the search query (either the survey question or a fixed query, depending on the retrieval method) and the chunked chat data.

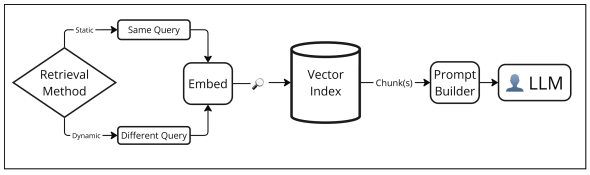

Figure 3: Retrieval Search Mechanism in RAG

3.3 Evaluation Metrics – Alignment

We evaluate each configuration by calculating the Mean Absolute Error (MAE) between the LLM's responses and those of the subject. Each survey answer is mapped to an ordinal integer value for this calculation.

To evaluate the effectiveness of persona encoding, we compare the MAE against two naive guessing methods as control variables and sanity checks: MAE_{Guess} of guessing the neutral options in each survey, and MAE_{Base} of running the LLM without retrieving persona data (figure 4). As seen in

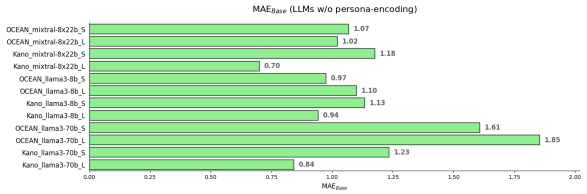

Figure 4: The 12 MAE_{Base}

figure 4, MAE_{Base} varies between 0.7 and 1.85 depending on configuration. Our other control variable, MAE_{Guess}, is 1.21 for OCEAN, and 0.9125 for Kano (average between subjects).

We calculate ΔMAE, representing the directional performance change in MAE from a control variable, to quantify the effect of persona encoding.

4 Results

OCEAN Personality Survey (Big Five)

Configuration	ΔMAE_{Base}	ΔMAE_{Guess}
L70-S	-0.46	-0.07
L70-L	**-0.78**	-0.14
L8-S	+0.03	-0.21
L8-L	-0.09	-0.20
Mixtral-S	-0.11	**-0.25**
Mixtral-L	+0.13	-0.06

Kano Survey on Video Game Preferences

Configuration	ΔMAE_{Base}	ΔMAE_{Guess}
L70-S	**-0.06**	+0.27
L70-L	-0.02	-0.09
L8-S	+0.08	+0.30
L8-L	+0.04	+0.07
Mixtral-S	-0.04	+0.23
Mixtral-L	0	**-0.21**

Table 1: Change over control variables when providing chat data of the subjects (the lower the better)

We observe in the table that Llama3-70b

is achieving remarkably higher alignment gain (ΔMAE_{Base}) from the subject's chat data than any of the other models in the OCEAN survey. We also notice that Llama3-8b is performing worse when given chat data in three out of four of the configurations. In addition, not a single configuration of subject *S* outperformed MAE_{Guess} in Kano. Finally, while Llama3-70b clearly is superior at utilizing provided chat data, Mixtral-8x22b somehow achieve the lowest MAE configurations (with subject *S* in OCEAN, and subject *L* in Kano).

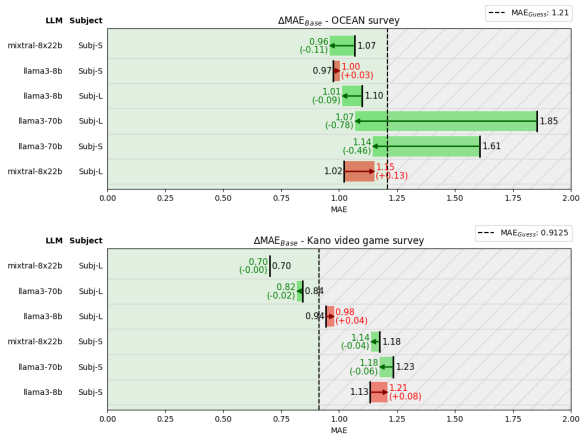

Figure 5: N_{conf}=18 in each row

We should also point out that meanwhile the base-personality of Llama3-70b is more aligned with subject *S* by $(1.85 - 1.61)$ 0.24 points, it actually becomes more aligned with subject *L* when provided with chat data $(1.14 - 1.07 = 0.07$ points).

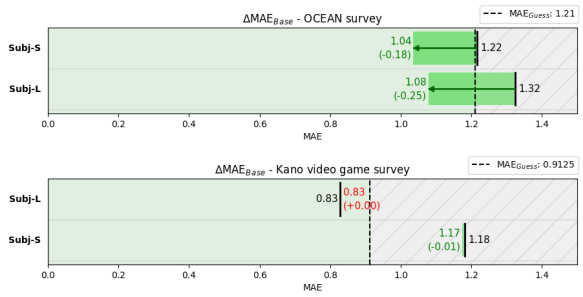

Figure 6: Aggregated MAE (N_{conf}=54 in each row)

When we aggregate the configurations the alignment improvement is remarkable in OCEAN for both subjects, meanwhile it is almost unnoticeable in the Kano survey. Another thing to notice, is the MAE_{Base} values – depicted as the vertical black lines of the floating bar chart in figure 6 – showing that the average base-personality is more aligned with *S* in OCEAN, but closer to *L* in Kano. This could indicate that *S* has more niche gaming preferences, and *L* is diverging more from average personality traits.

Regardless of the naive guessing method, the model performs better with the subject's chat data in three out of four cases, albeit marginally in Kano (-0.01 ΔMAE_{Base}, or -0.0825 ΔMAE_{Guess}). However, the objective is to outperform both control variables, and we can therefore only confidently say that chat data improves alignment in the OCEAN survey.

5 Limitations

We do acknowledge that the study is of an inadequate sample size to sufficiently generalize, but that is the premises of our experiment's required data. Therefore, we consider this study an initial exploration on the feasibility of LLM-proxy respondents.

Like touched upon by related work, the alignment is not the only metric determining an LLMs imitation abilities. Our study did not consider the internal consistency of the LLM, nor of the subjects. While the former is relatively straight forward, the latter invites many more questions: Are humans consistent at self reporting over time (Wang et al., 2024b; Jiang et al., 2023)? If no, should we readjust the "gold standard" of perfect alignment to match the subject's internal deviation – or is the objective only to capture a snapshot of the subject at a given moment?

6 Conclusion

The answer of our research question depends on the configuration of our experiment. We conclude that providing Llama3-70b with an individual's chat data, it can better replicate the OCEAN survey answers of the individual than our naive guessing methods. Alignment is at 1.07-1.14 MAE per question on a five point scale, and the error reduction from adding the subject's chat data is at 0.07-0.14 points relative to guessing the middle, and 0.46-0.78 points compared to the base configurations. That is 29-42% improvement with chat data, and is equivalent to 6-12% less errors than middle guess.

The results are more ambiguous for Kano than OCEAN, and only Llama3-70b showed consistent improvements in both surveys. When aggregating all configurations, we find evidence of better alignment than naive guessing in 2 out of 4 cases. Separating the model and subject variables, 6 out of 12 (4 of 6 OCEAN, and 2 of 6 Kano) outperform naive guessing – thus persona-encoding via RAG improves alignment in half of our experiments.

References

G. V. Aher, R. I. Arriaga, and A. T. Kalai. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In *Proceedings of the 40th International Conference on Machine Learning*, pages 337–371.

L. P. Argyle, E. C. Busby, N. Fulda, J. R. Gubler, C. Rytting, and D. Wingate. 2022. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Matthew Barsalou. 2023. Kano model video game data. [Dataset].

J. Brand, A. Israeli, and D. Ngwe. 2023. Using gpt for market research. *SSRN Electronic Journal*.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, and D. Amodei. 2020. Language models are few-shot learners.

D. Dillion, N. Tandon, Y. Gu, and K. Gray. 2023. Can ai language models replace human participants? *Trends in Cognitive Sciences*, 27:597–600.

J. J. Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *SSRN Electronic Journal*.

C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. 2024. Ruler: What's the real context size of your long-context language models? No. arXiv:2404.06654.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 10622–10643. Curran Associates, Inc.

J. A. Johnson. 2014. Measuring thirty facets of the five factor model with a 120-item public domain inventory: Development of the ipip-neo-120. *Journal of Research in Personality*, 51:78–89.

K. L. Keller. 1993. Conceptualizing, measuring, and managing customer-based brand equity. *Journal of Marketing*, 57(1):1–22.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks.

P. Li, N. Castelo, Z. Katona, and M. Sarvary. 2023. Determining the validity of large language models for automated perceptual analysis. *SSRN Electronic Journal*.

Tianlong Li, Shihan Dou, Changze Lv, Wenhao Liu, Jianhan Xu, Muling Wu, Zixuan Ling, Xiaoqing Zheng, and Xuanjing Huang. 2024. Tailoring personality traits in large language models via unsupervisedly-built personalized lexicons.

Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. 2024. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. *Preprint*, arXiv:2401.12474.

Fumio Takahashi Shin-ichi Tsuji Noriaki Kano, Nobuhiko Seraku. 1984. Attractive quality and must-be quality. *Journal of The Japanese Society for Quality Control*, 14(2):39–48.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery (ACM).

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Bonnie Rind. n.d. The power of the persona [a guide]. Pragmatic Institute - Resources, Retrieved May 15, 2024.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality traits in large language models. *Preprint*, arXiv:2307.00184.

H. R. Varian. 2010. *Intermediate Microeconomics: A Modern Approach*, 8 edition. W. W. Norton Co.

A. Wang, J. Morgenstern, and J. P. Dickerson. 2024a. Large language models cannot replace human participants because they cannot portray identity groups. No. arXiv:2402.01908.

Xintao Wang, Yunze Xiao, Jen tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024b. Incharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. *Preprint*, arXiv:2310.17976.