Evaluating Generative Agent Simulation Performance

Tim Schopf

1 Introduction and Objective

This report introduces an evaluation approach to quantify how closely AI agents (used as "synthetic respondents") reproduce real human responses in a market-research setting. To this end, an agent approach [6] is mimicking the traits and behaviors of real persons.

The central objective is not merely to check whether an agent produces an answer that looks plausible, but to measure whether the agent reaches the *same outcomes* as humans *for the same reasons*. This distinction matters for market research: outcomes (e.g., purchase intent) determine top-line conclusions, while reasons (values, constraints, and trade-offs) drive actionable insights for product positioning and messaging.

2 Related Work and Design Rationale

Lexical overlap metrics such as BLEU [5] or ROUGE [1] count surface-form matches and are ill-suited for open-ended responses where paraphrases are common. Pure semantic similarity via embeddings (e.g., BERTScore [7]) is more robust to paraphrasing, but it tends to be too coarse-grained: it collapses multiple aspects into a single similarity score and often does not reflect whether two answers match on the *decision* and the *underlying justification*. Therefore, the evaluation uses an *LLM-as-a-judge* approach: a separate judge model scores alignment between human and simulation outputs along explicitly defined criteria.

3 Evaluation Method

3.1 Evaluation dimensions: *How* vs. *Why*

The evaluation is designed around two complementary dimensions:

- **Answer conclusion (*How* humans and simulations behave):** whether simulated and human answers reach the same final conclusion/outcome for a given question.
- **Answer arguments (*Why* humans and simulations behave that way):** whether the underlying reasons and trade-offs match, i.e., whether the simulation behaves similarly because it relies on the same rationale rather than matching the conclusion by chance.

3.2 Metrics

We evaluate N paired answers (same question, one human answer, one simulated answer). An LLM is used to (i) extract a short *conclusion* from each answer and (ii) decompose the reasoning rationale of each answer into a set of *atomic arguments*. All judging is performed with the GPT-5-nano model [4], which provides a good trade-off between performance and cost-efficiency.

Average Conclusion Alignment (ACA). For each pair, a G-Eval [2] judge assigns a semantic alignment score (0-1) between the simulated and human *conclusions* (higher is better). ACA is the mean of these scores across all N answers. *This metric evaluates whether simulated and human answers reach the same conclusion.*

Average Argument Recall (AAR). Adapting the FActScore [3] approach, we define per-answer recall as $r_i = \frac{s_i}{h_i}$ (where s_i is the number of human arguments included in the simulated answer and h_i is the total number of arguments in the human answer) and report $\text{AAR} = \frac{1}{N}\sum_{i=1}^{N} r_i$. *This metric evaluates whether the human arguments are also included in the simulated arguments.*

4 Results

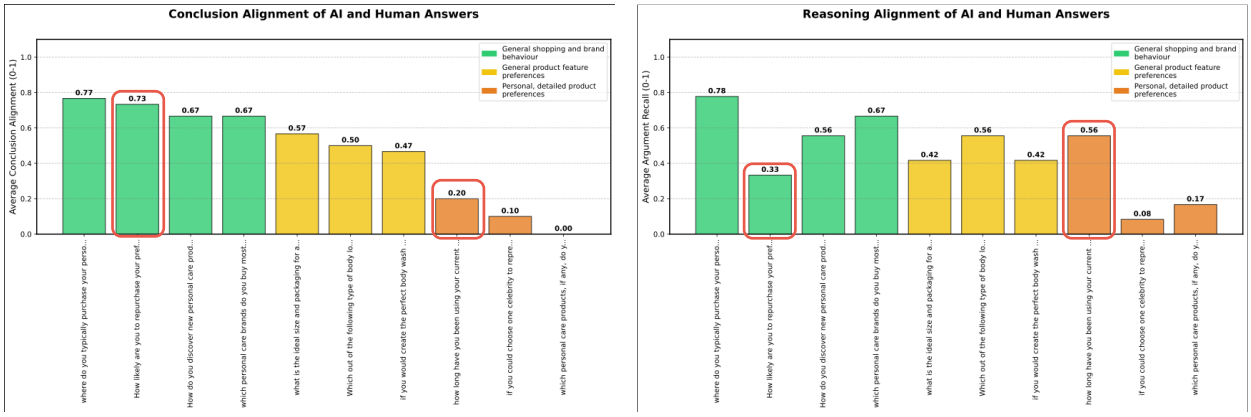

Figure 1: Overview of ACA and AAR scores between simulated and human answers.

Across respondents, scores are comparable (ACA/AAR: c_human 0.53/0.56; g_human 0.46/0.40; s_human 0.41/0.40), which suggests that AI agents can simulate different humans with similar accuracy. Using the predefined *question categories* (brand loyalty, influences, product preferences, shopping behavior), ACA is highest for *brand loyalty* (0.73), moderate for *product preferences* (0.48) and *shopping behavior* (0.48), and lowest for *influences* (0.10). AAR is highest for *product preferences* (0.52) and *shopping behavior* (0.50), moderate for *brand loyalty* (0.33), and lowest for *influences* (0.08). In Figure 1, we additionally regrouped questions into broader qualitative groups (color-coded) to highlight systematic patterns across related questions. We observe high conclusion alignment between humans and AI agents for general shopping and brand-preference questions (green/yellow), but reduced alignment for personal, highly specific product preferences (orange). The red-circled cases highlight divergences between conclusion and argument alignment for the same questions in both directions (high ACA with low AAR and low ACA with high AAR), i.e., conclusions and reasons are not consistently aligned. Based on the observed ACA–AAR divergences, simulations may reach the same conclusion as humans with mismatched motivations. Therefore, the use of ACA alone is not sufficient and should be combined with AAR for comprehensive evaluation.

5 Conclusion

Overall, the results suggest that AI agents can approximate human answer *conclusions* for broad, high-level questions (e.g., general shopping and brand preferences), but they are less reliable for simulating highly personal, detailed preferences. Importantly, AAR reveals that alignment of human and simulated answers on the final conclusion (high ACA) does not necessarily imply alignment on the underlying reasons (low AAR), and that strong reason overlap can occur even when conclusions differ. Hence, relying on ACA alone can misrepresent simulated consumer behaviour. In practice, simulations may be used to validate hypotheses about general shopping and brand preferences. However, in-depth market analysis and motivation-sensitive decisions should use human data.

References

[1] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[2] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[3] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100, Singapore, December 2023. Association for Computational Linguistics.

[4] OpenAI. GPT-5 nano (model documentation), 2026. Accessed: 2026-02-03.

[5] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.

[6] Joon Sung Park, Carolyn Q. Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S. Bernstein. Generative agent simulations of 1,000 people, 2024.

[7] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.