Pay What LLM Wants:
Can LLM Simulate Economics Experiment with 522 Real-human Persona?

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used to simulate human behavior, yet most studies rely on fictional personas and evaluate aggregate tendencies rather than predictions for specific individuals. We evaluate whether LLMs can predict individual economic decision-making through Pay-What-You-Want (PWYW) pricing experiments, using detailed personas of 522 real Korean participants. Three multimodal LLMs are tested across two interaction conditions, two persona formats, and four prompting methods, with predictions compared against the corresponding human responses. Across every condition, LLMs reproduce group-level tendencies to a reasonable degree but predict individual decisions poorly, with overall accuracy below 5%, and neither persona format nor prompting method closes this gap. Analyzing the response distributions, we find that LLMs produce far fewer distinct willingness-to-pay values than humans and assign similar amounts across different personas, indicating a failure to differentiate individuals rather than mere inaccuracy. These findings offer empirical guidance for persona-based simulation in computational social science.

1 Introduction

Recent advances in Large Language Models (LLMs) have sparked growing interest in their ability to reproduce social and psychological behaviors, leading to active research on simulating human-like judgment and choice patterns (Park et al., 2024, 2023; Wang et al., 2023; Cheng et al., 2023; Huang et al., 2024; Choi et al., 2025). A common approach injects personas into LLMs to mimic human attitudes, value judgments, and choice behaviors. Reliable simulation could complement behavioral research, where recruiting human participants is costly and slow. Most prior studies, however, evaluate general tendencies without precise behavioral comparisons against actual humans, employing methods such as fictional personas, creating personas from demographic information, or extracting personas from existing QA datasets (Horton, 2023; Hewitt et al., 2024; Leng and Yuan, 2023; Ross et al., 2024; Leng, 2024a; Xie et al., 2024b).

Beyond fictional personas, a line of work in the social sciences grounds LLM simulation in real human data, prompting LLMs with attributes of actual respondents and comparing the LLM responses against survey or experimental baselines (Argyle et al., 2023; Bisbee et al., 2024; Qu and Wang, 2024; Goli and Singh, 2024). The work supports LLM-based simulation under a narrow setting, where evaluation centers on aggregate agreement and a single model is tested under one prompting strategy and one text-only input format. Among the work using real human data, Park et al. (2024) reports 85% accuracy in reproducing general social survey responses, though the target is attitudinal surveys rather than economic decisions involving monetary costs. An attitudinal survey can elicit idealistic responses without realistic constraints. Economic decision-making that requires willingness to pay involves budget constraints, opportunity costs, and cultural values, providing a more demanding test of LLMs' human-like behavior reproduction.

The attempts in the social sciences show that LLMs can serve as behavioral simulators, but leave open how well current LLMs perform and under what conditions the performance holds, since a single-model, single-prompt setup does not separate the effect of model choice, prompting strategy, persona format, or input modality. Examining the capability along these axes calls for an experiment that places a precise demand on the simulation. We adopt the Pay-What-You-Want (Kim et al. (2009); PWYW) experiments of Lee and Jin (2022). PWYW asks consumers to set a payment amount of their own, so a prediction is correct only when the model reproduces a specific monetary decision rather than a directional attitude, and the amount re-

flects budget constraints, prior cultural experience, and social norms at once. The setting yields an individual-level target that is difficult to satisfy by aggregate tendency alone. The experimental design and human data collection were conducted by Lee and Jin (2022); we apply the same experimental structure to LLMs using detailed information from 522 actual participants, examining how precisely LLMs predict individual behaviors across various models and persona injection methods. In this paper, we try to answer following two questions.

RQ1. *Can LLMs predict economic decision of each person in sequential or human-guided way?*

The first question examines whether LLMs reproduce the decision-making of corresponding human participants when injected with individual persona information. Each PWYW scenario presents several questions in sequence, so an early error in the model's own answers propagates, and the observed inaccuracy no longer separates a failure to read the persona from the accumulation of prior mistakes (Lampinen et al., 2022; Wei et al., 2025). To isolate the two, the *Sequential Condition* lets the model keep its previous responses, while the *Human-guided Condition* replaces them with the actual human responses in the conversation history.

RQ2. *How do persona format and prompting method affect LLM simulation of human behavior?*

The second question analyzes how prediction performance varies with the way a persona is given to the model. Providing a persona involves two independent choices, the format that holds the same information and the prompting method that directs how the model reasons over it, so we organize the experiment along the two axes. For format, the Survey format gives the persona as structured key-value entries that follow the original questionnaire, while the Storytelling format presents the same information as a biographical narrative. For prompting, we test Chain-of-Thought (CoT) (Wei et al., 2022), Retrieval-Augmented Generation (RAG) (Lewis et al., 2020), and Few-shot prompting (Brown et al., 2020) against a base prompt, as the three are the most widely adopted techniques for complex reasoning tasks.

The study reports an empirical observation across three multimodal models, two persona formats, and four prompting methods. Current LLMs reproduce group-level tendencies to a reasonable degree but fail at individual-level prediction, with overall accuracy below 5% in every condition. Advanced prompting methods such as CoT, RAG, and Few-shot bring no consistent gain over a base prompt, and the structured survey format outperforms the storytelling format. The results indicate that the gap between group-level and individual-level fidelity is not closed by persona format or prompting choice.

2 Related Work

Research on LLM-based behavioral simulation falls into three groups by the type of persona, persona-free attempts, attempts with simulated personas, and attempts with real personas.

Persona-free Studies with Social Psychology: To assess LLM social behaviors, studies adopt social game paradigms including the Public Goods Game and the Ultimatum Game to evaluate cooperative tendencies, altruism, and strategic capabilities (Sreedhar et al., 2025a,b). The approaches assign roles or structure cooperative frameworks through prompts. Most analyses focus on aggregate group trends, showing that multi-agent systems approximate distributions of human cooperative behavior better than single models, and findings remain at the level of group statistics, leaving individual persona-level precision unexamined.

Evaluation Attempts with Simulated Personas: Simulated personas fall into three types by the information provided to the model. Personality-based personas from social psychology induce distinct behavioral patterns (Zhang et al., 2024; Ashery et al., 2025), such as agents with different traits engaging in negotiation. Demographic-based personas simulate variation in attitudes toward social issues (Piao et al., 2025; Kim et al., 2025); Piao et al. (2025) injected demographic profiles into over 10,000 LLM agents to analyze decisions on basic income and disaster response. Narrative-based personas with memories, goals, and routines simulate long-term interactions (Park et al., 2023; Xie et al., 2024a), where Park et al. (2023) tracked agent behavior within a virtual town. The approaches diversify LLM behavior but lack grounding in real human data, and so cannot establish one-to-one correspondence with actual human attitudes.

Evaluation Attempts with Real Personas: A line of work in the social sciences grounds simulation in real human data. In political behavior

and public opinion research, studies prompt LLMs with attributes of actual respondents and compare the outputs against survey baselines (Argyle et al., 2023; Bisbee et al., 2024; Qu and Wang, 2024), reporting that aggregate scores track the baseline while response variance is compressed. In consumer research, Goli and Singh (2024) examines whether LLMs reproduce human preferences and finds systematic departures from human choice. Park et al. (2024) constructs personas from interview transcripts and reports 85% accuracy on General Social Survey responses, with a focus on attitudinal surveys rather than economic decisions carrying monetary consequences.

The work above shares two properties. Evaluation centers on aggregate agreement, leaving individual-level accuracy largely unmeasured, and each study tests a single model under one prompting strategy and one text-only format, which does not separate the effect of model choice, prompting strategy, or persona format. The present study evaluates three multimodal models, measures both individual-level accuracy and group-level tendencies, and compares two persona formats across four prompting methods on an economic decision task. The next section describes the human experiment that provides the baseline data.

3 Human Experiment

This section describes the human experiment that provides the baseline data, based on the data of Lee and Jin (2022). We introduce the PWYW design, the collected persona items, and the decision-making questions.

3.1 Pay What You Want (PWYW)

PWYW is a participatory pricing mechanism where consumers determine payment amounts based on perceived value rather than a fixed price, widely used to measure willingness to pay in economics and consumer behavior research (Kim et al., 2009; Gerpott, 2017). Determining an amount requires a consumer to weigh preferences, economic constraints, social norms, and value systems at once. Lee and Jin (2022) applied PWYW to cultural product consumption to observe how cultural tastes, experience, knowledge, and educational background shape actual consumption decisions, which yields realistic economic choice data suitable for evaluating LLM behavioral simulation.

3.2 Persona Information

Lee and Jin (2022) collected personas from 522 Korean adults (254 male, 268 female; aged 30–39). As shown in Figure 1, each persona covers 65 items across seven categories, basic demographics, attitudes and preferences toward cultural arts, cultural arts knowledge, cultural education experience, institutional usage and satisfaction, social norms, and family background. The items together capture each individual's cultural capital and socioeconomic background. We obtained the data from the authors. For details, see Appendix A.

3.3 Question Form

Each PWYW scenario covers two cultural products, an art exhibition and a music performance. For each product, participants answer three questions that elicit economic decision-making in different response formats. The first question (`Q1`) asks whether the participant would attend at the regular price, and the second (`Q2`) asks whether the participant would attend when allowed to set their own price. Both are **binary** yes/no questions. The third question (`Q2.1`), given only to participants who agreed in `Q2`, is an **open-ended numerical** question on the exact amount they would pay. For participants unwilling to attend, a separate **multiple-choice** question (`Q3`) records the reason.

The three formats place different demands on prediction. The binary questions ask only for the direction of a decision, the multiple-choice question selects among a fixed set of reasons, and the open-ended question requires a specific amount and is therefore the most demanding. We report results by format throughout, so the difficulty of each format can be traced in the analysis. For the full questionnaire, see Appendix B.

3.4 Statistical Tests Conducted

Group-level human tendencies were estimated with the Heckman two-step model (Heckman, 1976). The first step is a probit regression with the yes/no item `Q2` as the dependent variable and persona attributes as predictors, estimating the probability of viewing intention. The second step is a linear regression on the willingness-to-pay amount `Q2.1`, restricted to participants who agreed in `Q2`, with the inverse Mills ratio from the first step added to correct for self-selection bias. For the detailed procedure, see Appendix C.

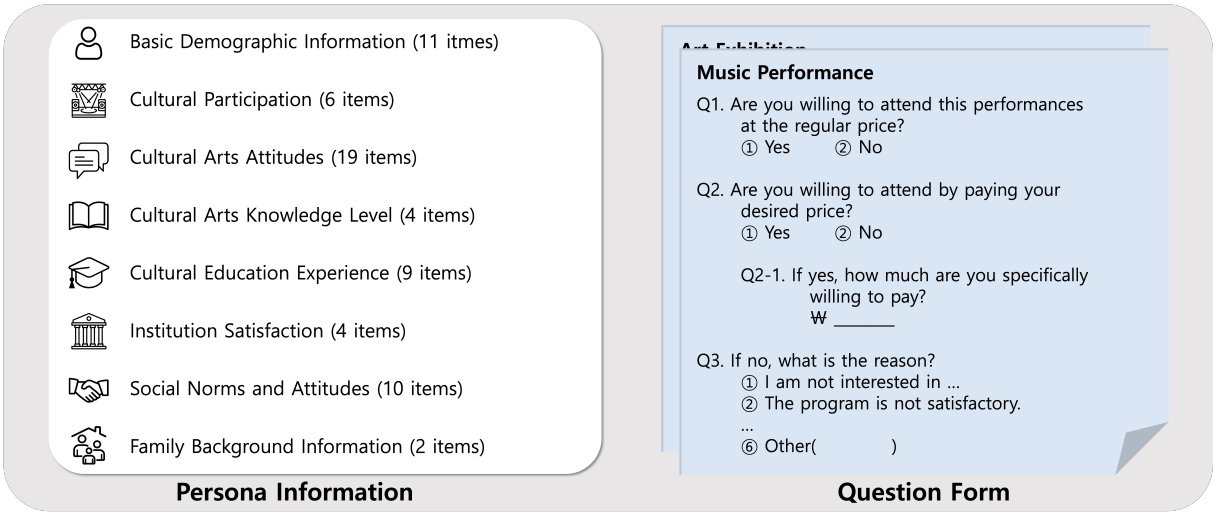

Figure 1: Overview of persona information categories and experimental question form.

4 RQ1: Sequential vs. Human-guided

This section investigates whether LLMs predict individual human decisions, and how prediction is affected by whether earlier answers in a scenario come from the model or from the human.

4.1 Experiment Setup

Procedure. We replicate the human experiment for each LLM, presenting the same materials in the same order shown to human participants. The persona of an actual participant is given as the system prompt with the Korean instruction *"Think of yourself as a person of the given persona and answer."*, followed by the survey introduction, the product descriptions, the accompanying images, and the questions, interleaving text and images exactly as in the original questionnaire. As all participants are Korean, and prior work shows that linguistic variation affects model behavior (Leng, 2024b; Verma et al., 2023), every item is given in the original Korean. The full prompt templates, the image inputs, and the response parsing rules are provided in Appendix D.

Conditions. Each scenario presents questions in sequence, so a model answers later questions while conditioning on earlier ones. We compare two conditions for the source of the earlier answers. In the *Sequential Condition*, the model keeps its own previous responses, so an early error propagates through the scenario. In the *Human-guided Condition*, the actual human responses replace the model's previous answers in the conversation log from the second question onward, so each prediction conditions on a correct prior context. Comparing the two isolates the effect of error accumulation from the effect of failing to read the persona. Appendix D shows how the conversation log is constructed under each condition.

Models. We evaluate three multimodal LLMs that process text and images jointly, GPT-4o (Hurst et al., 2024), Llama-3.2-90B-Vision-Instruct (Dubey et al., 2024), and Qwen2.5-VL-72B-Instruct (Bai et al., 2025). All three handle Korean reasonably well; on the Korean Hae-Rae benchmark (Son et al., 2023) they score 0.836, 0.738, and 0.636. GPT-4o is accessed through the OpenAI API and the other two through the OpenRouter API, with temperature fixed at 0 for consistency.

4.2 Evaluation

We evaluate prediction along two aspects, *individual-level accuracy* and *group-level tendency*. The exact metric definitions and the regression procedure behind them are given in Appendix C.

Individual-level accuracy. Individual-level accuracy is the proportion of cases where the LLM response exactly matches the human response. We report it at three granularities, per question, over the binary questions `Q1` and `Q2` together, and as an overall score that requires every item in a scenario to be correct. Exact match suits the binary and multiple-choice questions, but understates performance on the open-ended `Q2.1`, where a near-miss amount is counted as wrong. For `Q2.1` we therefore add the mean absolute error (MAE) and mean absolute percentage error (MAPE) between the predicted

Metric	Sequential GPT	Sequential Llama	Sequential Qwen	Human-guided GPT	Human-guided Llama	Human-guided Qwen
A. Accuracy by Question (Individual-level)						
M_Q1	62.26	60.15	61.11	62.07	63.60	62.26
M_Q2	57.28	78.54	75.10	70.50	80.84	76.82
M_Q2.1	5.43	22.25	9.78	15.65	26.41	33.74
MAE	2928	2429	2612	2495	2236	2308
MAPE	62.28	49.97	46.22	58.01	46.70	40.64
M_Q3	11.36	0.00	0.00	31.86	9.73	29.20
TF Total	44.06	53.83	46.17	45.98	55.36	45.79
M_total	6.90	11.49	4.79	6.70	11.11	15.13
A_Q1	64.75	37.55	66.28	65.13	55.17	65.71
A_Q2	67.82	62.84	64.75	76.44	73.37	65.71
A_Q2.1	16.23	9.60	4.28	28.44	23.24	24.16
MAE	9216	9191	10511	6867	7625	7203
MAPE	70.38	92.73	68.76	44.26	60.45	52.89
A_Q3	16.73	0.19	0.13	0.00	35.38	36.41
TF Total	45.59	32.95	46.55	44.44	42.72	40.04
A_total	12.45	5.94	1.72	17.05	16.09	13.79
Total Acc	4.60	0.77	0.77	4.02	0.96	4.41
B. Coefficient Similarity to Human (Group-level)						
CSA	0.62	0.55	0.66	0.72	0.86	0.76
SSA	0.86	0.83	0.72	0.93	0.90	0.86
Jaccard	0.43	0.00	0.27	0.67	0.57	0.20

Table 1: LLM performance across conditions for the music (M) and art (A) domains. `Q2.1` reports exact-match accuracy, with MAE and MAPE below it. TF Total is the accuracy on the binary questions `Q1` and `Q2`.

and the human amount.

Group-level tendency. Group-level tendency compares the persona-to-decision relationship estimated for humans and for each LLM, using the regression coefficients of the two Heckman models. We use three metrics. *Coefficient Sign Agreement* (CSA) is the proportion of coefficients with matching signs, measuring agreement on the direction of each variable's influence. *Statistical Significance Agreement* (SSA) is the proportion of variables that are significant in both models or non-significant in both, measuring agreement on which variables matter. *Jaccard Index* (Jaccard, 1908) is the overlap between the two sets of significant variables, measuring how much the important variables themselves coincide.

4.3 Results

Table 1 reports the two conditions. We read the gap between them as a measurement of error accumulation, the share of inaccuracy that comes from carrying over earlier mistakes rather than from misreading the persona.

The two conditions isolate error accumulation, which surfaces only on the open-ended item. The Human-guided condition is more accurate than the Sequential condition across models, but the gap

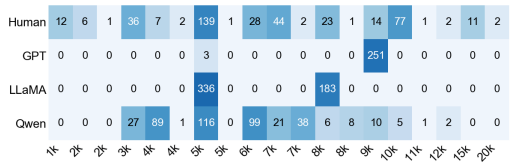

Figure 2: Response distribution in Sequential Condition

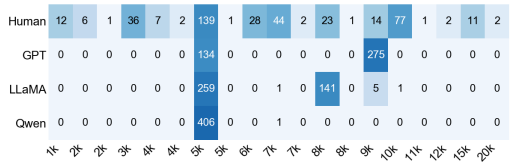

Figure 3: Response distribution in Human-Guided

is not uniform. The binary questions move little, while the open-ended `Q2.1` shifts sharply, for Qwen from 9.78 to 33.74 in music and from 4.28 to 24.16 in art. Because the only change between conditions is the source of the earlier answers, this selective gap lets us read the Sequential–Human-guided difference as a measurement of error accumulation, and it shows the accumulation is confined to the open-ended format. The improvement, however, is one of exact matches rather than of precision. For the same `Q2.1`, Qwen's MAE falls only from 2612 to 2308 while exact-match accuracy more than triples, so a corrected context lands more answers but leaves the rest far from the human amount.

Group-level agreement does not imply individual-level accuracy. A model can track human tendencies at the group level while failing to reproduce individual choices. Across both conditions, CSA and SSA stay between 0.72 and 0.93, yet Overall Accuracy never exceeds 4.6%. Llama is the clearest case, with the highest sign agreement in the Human-guided condition (CSA 0.86) but an Overall Accuracy of 0.96%. The model identifies which persona attributes push a decision up or down, yet does not convert that into a correct prediction for a specific person. Figures 2 and 3 show the same split from the response side, human amounts are widely spread while model amounts cluster at a few price points, and the clustering tightens under the Human-guided condition even as accuracy rises.

Models disagree on which attributes matter. While CSA and SSA are similar across models, the Jaccard index ranges from 0.00 to 0.67, so the models agree on the sign and significance of coefficients but not on the set of attributes they treat as important. The agreement on tendencies is there-

fore reached through different internal weightings rather than a shared reading of the persona.

4.4 Discussion

Taken together, the three observations describe a single pattern. LLMs recover human tendencies at the group level, yet fail at the individual level, and the failure is not stable, it shifts with question format, with the source of the conversation history, and across models. The Human-guided condition narrows the gap but does not close it, and the residual error shows that supplying a correct context is not enough for a model to reproduce a specific person's decision.

The pattern leaves one question open. RQ1 fixes a single way of presenting the persona, the structured Survey format with a base prompt, so it cannot tell whether the individual-level failure is a fixed limit or an artifact of how the persona is supplied. If a different format or a stronger prompting method changed the outcome, the failure would be a matter of input design; if it did not, the limit would lie deeper. RQ2 examines the question by varying persona format and prompting method.

5 RQ2: Persona Format and Prompting Method

RQ1 leaves open whether the individual-level failure is a fixed limit or an artifact of how the persona is supplied, as it fixes a single presentation, the Survey format with a base prompt. RQ2 addresses this by varying the persona along the points at which it reaches the model: the surface form that holds the information, the amount of information supplied, the reasoning applied to it, and the external examples placed beside it. If the failure were an artifact of input design, at least one of these variations would close the individual-level gap; if it persisted across all of them, the limit would lie deeper than persona presentation. We keep the models and metrics of RQ1 and adopt the Human-guided condition for all variations, since it fixes the earlier answers to actual human responses, so a change in performance can be attributed to the persona variation rather than to accumulated error from earlier turns.

5.1 Experiment Setup

RQ2 keeps the procedure, models, and metrics of RQ1 and uses the Human-guided condition as the baseline, organizing the conditions along two axes, persona format and prompting method. The full prompt templates, generation, and retrieval settings are in Appendix D.

5.1.1 Persona Format

Users typically supply a persona either as structured fields or as a natural-language description, and the two formats follow these practices. The Survey format is the structured key-value form of the 65 items used in RQ1. The Storytelling format reconstructs the same items into a biographical narrative with GPT-4o; each narrative is checked by keyword matching and regenerated until all 65 items are present.

5.1.2 Prompting Method

Each method intervenes at a different point of how the persona reaches the model: CoT changes the reasoning applied to it, RAG changes the amount supplied, and Few-shot adds an external signal beside it. Together with the base prompt, they let us test whether varying any of these points closes the individual-level gap.

Base and CoT The base prompt places the full persona in the system prompt and asks for an answer directly, as in RQ1; CoT adds an instruction to reason step by step, so only the reasoning differs.

RAG For each question, the question text retrieves the most relevant items from a per-persona index, and these replace the full persona, so the model sees only the part relevant to the current question. We retrieve the top $k{=}10$ items by cosine similarity, embedding both the items and the query with `text-embedding-3-large`.

Few-shot The actual responses of the three personas most similar to the target, measured by cosine similarity over persona embeddings (`text-embedding-3-large`), are prepended as examples before each question. The four methods apply to the Survey format; Storytelling uses base, CoT, and RAG, as its narrative form leaves no room for in-context examples within the context limit. Further retrieval details are given in Appendix D.

5.2 Results

Table 2 reports performance across the two axes of persona format and prompting method. We read the results through four lenses: question difficulty, model, format and prompting method, and the group versus individual level.

	Naive			CoT			RAG			Few-Shot		
Metric	**GPT**	**Lla**	**Qwen**	**GPT**	**Lla**	**Qwen**	**GPT**	**Lla**	**Qwen**	**GPT**	**Lla**	**Qwen**
Survey Format												
Binary	45.21	49.04	42.87	50.53	47.99	47.13	45.02	45.98	45.79	50.29	45.02	50.86
Q2.1	22.05	24.83	28.95	25.01	17.64	23.60	22.69	24.15	28.92	21.19	11.07	17.37
Q3	15.93	22.56	32.81	33.62	23.56	29.55	33.37	32.62	32.95	30.22	10.16	31.88
Overall	4.02	0.96	4.41	2.49	0.57	1.53	3.07	1.92	3.26	1.72	0.77	0.77
MAE_M	2495	2236	2308	2235	2402	2253	2434	2214	2287	2410	2751	3228
MAE_A	6867	7625	7203	6671	7824	7497	6794	8182	7372	8170	8674	9219
MAPE_M	58.0	46.7	40.6	47.4	56.6	45.3	58.0	42.7	40.6	49.9	69.2	60.8
MAPE_A	44.3	60.5	52.9	42.5	61.5	55.9	43.5	71.3	56.3	53.1	76.7	72.6
CSA	0.72	0.86	0.76	0.86	0.83	0.86	0.83	0.69	0.69	0.72	0.86	0.86
SSA	0.93	0.90	0.86	0.86	0.97	0.90	0.90	0.86	0.86	0.86	0.90	0.86
Jaccard	0.67	0.57	0.20	0.20	0.80	0.57	0.50	0.20	0.33	0.20	0.50	0.20
Storytelling Format *(Few-Shot not applied)*												
Binary	47.61	47.51	49.72	49.05	33.91	49.72	47.90	45.21	49.05	–	–	–
Q2.1	21.80	20.40	30.18	25.19	21.16	23.88	23.79	27.12	27.70	–	–	–
Q3	34.14	16.92	36.35	34.14	17.94	29.33	32.60	30.90	27.43	–	–	–
Overall	0.19	0.38	1.34	1.34	0.57	1.34	0.96	1.92	1.72	–	–	–
MAE_M	2463	2226	2309	2315	2199	2200	2500	2307	2307	–	–	–
MAE_A	6962	8161	7026	7011	7240	7696	6809	7815	7509	–	–	–
MAPE_M	59.3	48.6	40.7	46.8	52.8	45.9	57.1	40.8	40.7	–	–	–
MAPE_A	45.0	64.5	52.8	43.0	55.6	55.1	44.4	68.2	56.1	–	–	–
CSA	0.72	0.69	0.76	0.69	0.93	0.62	0.83	0.83	0.86	–	–	–
SSA	0.93	0.93	0.93	0.93	0.93	0.83	0.90	0.90	0.86	–	–	–
Jaccard	0.60	0.67	0.67	0.60	0.67	0.17	0.40	0.50	0.40	–	–	–

Table 2: LLM performance on the RQ2 conditions, averaged over the music and art domains except for MAE and MAPE, which are reported per domain (_M music, _A art) as the two domains have different price ranges. Binary is the accuracy on `Q1` and `Q2`; `Q2.1` is the open-ended amount; `Q3` is the multiple-choice question; Overall requires all items correct. CSA, SSA, and Jaccard measure group-level coefficient similarity to humans. Few-Shot is not applied to the Storytelling format due to context length constraints.

Accuracy stays unreliable once read against chance level. The questions differ in chance level. A binary item, Q1 or Q2, has a random baseline of 50%, the multiple-choice item Q3 of 20%, and the open-ended amount Q2.1 has no fixed guess. Answering both binary items correctly (TF Total) has a chance level of 25%, and the models reach 40–55% on this measure, above chance but far from reliable. On Q3 they stay near or below the 20% baseline. Q2.1 is the lowest throughout, and since Overall Accuracy requires every item to be correct, it stays below 5% in every condition. The binary and multiple-choice items admit no near-miss, so exact match is the only available score for them. For Q2.1 we also report MAE and MAPE, which credit a near-miss and penalize only the size of the deviation. Even under these measures the predictions remain far from the actual amounts, as MAPE stays above 50% in most conditions, meaning the predicted price deviates from the human amount by more than half of that amount on average.

No model predicts individuals well. GPT-4o, LLaMA, and Qwen differ in places. Qwen reaches the highest Q2.1 accuracy in several conditions, while LLaMA is weakest on the open-ended item. No model separates itself at the individual level, as all three remain below 5% Overall Accuracy across every format and prompting method.

The structured format helps, but prompting methods do not. The Survey format records higher Overall Accuracy than the Storytelling format in most conditions, for example 4.02 against 0.19 for GPT-4o and 4.41 against 1.34 for Qwen under the base prompt. The binary items move little between formats, so the difference is driven by the open-ended item. The advanced prompting methods bring no consistent gain over the base prompt. CoT lowers GPT-4o's Overall Accuracy from 4.02 to 2.49 and Few-shot from 4.02 to 1.72, while RAG stays near the base prompt. MAE and MAPE for Q2.1 stay in the same range across methods as well.

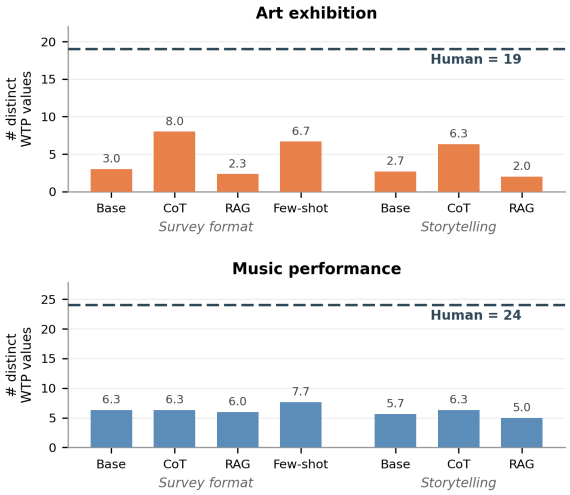

Figure 4: Number of distinct willingness-to-pay values produced over the 522 personas, by condition. The dashed line marks the human level. LLM values are averaged over the three models.

Group-level agreement is high while individual-level accuracy is not. At the individual level the models predict poorly, as the Overall Accuracy and Q2.1 figures show. Group-level agreement is higher. CSA and SSA stay between 0.62 and 0.97 across conditions, indicating broad agreement on the direction and significance of persona coefficients. The Jaccard index is lower, between 0.17 and 0.67, and neither metric moves consistently with format or prompting method.

5.3 Discussion

The failure is not an artifact of how the persona is supplied. RQ2 varied the persona along four points at which it reaches the model, namely the surface form, the amount of information, the reasoning applied to it, and the external examples beside it. Seen through any of the lenses above, the individual-level result does not move: Overall Accuracy stays below 5% and Q2.1 remains the failure point throughout. The structured Survey format is easier for the model to read, but its advantage stays within the same range. Since no variation removes the failure, it cannot be attributed to a single way of supplying the persona, and the limitation lies deeper than persona presentation.

The model does not differentiate one persona from another. One pattern invites a closer look, as the open-ended amounts appear to fall on a narrow set of values. We count the distinct willingness-to-pay values each condition produces over the 522 personas (Figure 4). Human respondents give 19 distinct values for the art scenario and 24 for music, while the LLM conditions give on average 4.4 and 6.2. In several conditions a single amount accounts for most responses, and all 409 art responses of Storytelling-RAG Qwen fall on one value. This count comes from the responses themselves and does not depend on whether any amount is correct, so it is unaffected by exact-match scoring. A model that returns the same few amounts regardless of the persona is not reproducing individuals, and the individual-level failure is one of not differentiating one person from another.

On the group-level metrics. The high CSA and SSA show that the model and human regressions agree on the direction and significance of most coefficients. The low Jaccard index shows that the set of variables the model treats as significant overlaps only partly with the set humans rely on. The model matches humans on which way each attribute pushes the decision, yet draws on a narrower set of attributes when it comes to which ones actually matter, and changing the format or prompting method does not move this overlap.

Across every lens, the individual-level failure holds. The model reproduces aggregate tendencies but assigns the same few amounts to distinct personas, drawing on a narrow part of the persona to do so. Prompt-level intervention does not reach this, and reproducing individuals would require the model to use the persona in a way that current models do not.

6 Conclusion

This study evaluates how well LLMs simulate human economic decision-making, using real personas of 522 participants in PWYW experiments. Across two interaction conditions, two persona formats, and four prompting methods, LLMs reproduce group-level tendencies reasonably well but predict individual decisions poorly. The failure does not depend on how the persona is supplied. Neither a corrected context, a structured persona format, nor chain-of-thought, retrieval, or few-shot prompting moves accuracy out of the same range. The response distributions show that LLMs assign a small set of willingness-to-pay values across distinct personas, so the failure reflects a limited ability to differentiate one person from another. As it persists across every variation we tested, this failure is unlikely to be resolved by prompt engineering.

Limitation

These results provide crucial empirical evidence for the current boundaries of LLM-based human behavior simulation and highlight the need for fundamental advances in model architecture and training paradigms rather than incremental improvements in prompt engineering. However, this study has three limitations. First, we did not conduct detailed analysis of which specific persona components among the 65 collected items most significantly influence prediction performance. Second, we did not explore advanced technical modeling approaches such as ensemble methods or fine-tuning that could potentially improve accuracy. Third, our experiment only involves Korean participants. As our focus was investigating LLM's general ability of simulating human behavior, we decided put more emphasis on the entire situation rather than focusing a specific experimental setting. Despite our valuable findings, further research on component-wise analysis is required.

References

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Bisbee, Joshua D Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M Larson. 2024. Synthetic replacements for human survey data? the perils of large language models. *Political Analysis*, 32(4):401–416.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Myra Cheng, Tiziano Piccardi, and Diyi Yang. 2023. Compost: Characterizing and evaluating caricature in llm simulations. *arXiv preprint arXiv:2310.11501*.

Junhyuk Choi, Yeseon Hong, Minju Kim, and Bugeun Kim. 2025. Examining identity drift in conversations of llm agents. *Preprint*, arXiv:2412.00804.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Torsten J Gerpott. 2017. Pay-what-you-want pricing: An integrative review of the empirical research literature. *Management Science Letters*, 7(1):35–62.

Ali Goli and Amandeep Singh. 2024. Frontiers: Can large language models capture human preferences? *Marketing Science*, 43(4):709–722.

James J Heckman. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of economic and social measurement, volume 5, number 4*, pages 475–492. NBER.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. 2024. Predicting results of social science experiments using large language models. *Preprint*.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Jen-tse Huang, Wenxuan Wang, Eric John Li, Man Ho Lam, Shujie Ren, Youliang Yuan, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. On the humanity of conversational ai: Evaluating the psychological portrayal of llms. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Paul Jaccard. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Sci. Nat.*, 44:223–270.

Jiseon Kim, Jea Kwon, Luiz Felipe Vecchietti, Alice Oh, and Meeyoung Cha. 2025. Exploring persona-dependent llm alignment for the moral machine experiment. *Preprint*, arXiv:2504.10886.

Ju-Young Kim, Martin Natter, and Martin Spann. 2009. Pay what you want: A new participative pricing mechanism. *Journal of marketing*, 73(1):44–58.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Mi-Jin Lee and Hyun Joung Jin. 2022. A study on factors influencing the viewing intention and amount of willingness-to-pay toward the classical music performance under pay-what-you-want pricing system. *Journal of Consumer Studies*, 33(5):49–72.

Yan Leng. 2024a. Can llms mimic human-like mental accounting and behavioral biases? In *Proceedings of the 25th ACM Conference on Economics and Computation*, EC '24, page 581, New York, NY, USA. Association for Computing Machinery.

Yan Leng. 2024b. Can llms mimic human-like mental accounting and behavioral biases? *Available at SSRN 4705130*.

Yan Leng and Yuan Yuan. 2023. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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*, pages 1–22.

Joon Sung Park, Carolyn Q Zou, Aaron Shaw, Benjamin Mako Hill, Carrie Cai, Meredith Ringel Morris, Robb Willer, Percy Liang, and Michael S Bernstein. 2024. Generative agent simulations of 1,000 people. *arXiv preprint arXiv:2411.10109*.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *Preprint*, arXiv:2502.08691.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11(1):1–13.

Jillian Ross, Yoon Kim, and Andrew W Lo. 2024. Llm economicus? mapping the behavioral biases of llms via utility theory. *arXiv preprint arXiv:2408.02784*.

Guijin Son, Hanwool Lee, Suwan Kim, Huiseo Kim, Jaecheol Lee, Je Won Yeom, Jihyu Jung, Jung Woo Kim, and Songseong Kim. 2023. Hae-rae bench: Evaluation of korean knowledge in language models. *arXiv preprint arXiv:2309.02706*.

Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V Nickerson, and Lydia B Chilton. 2025a. Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1272–1286, New York, NY, USA. Association for Computing Machinery.

Karthik Sreedhar, Alice Cai, Jenny Ma, Jeffrey V Nickerson, and Lydia B Chilton. 2025b. Simulating cooperative prosocial behavior with multi-agent llms: Evidence and mechanisms for ai agents to inform policy decisions. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 1272–1286, New York, NY, USA. Association for Computing Machinery.

Yash Verma, Anubhav Jangra, Raghvendra Kumar, and Sriparna Saha. 2023. Large scale multi-lingual multi-modal summarization dataset. *arXiv preprint arXiv:2302.06560*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zeyu Wei, Shuo Wang, Xiaohui Rong, Xuemin Liu, and He Li. 2025. Shadows in the attention: Contextual perturbation and representation drift in the dynamics of hallucination in llms. *arXiv preprint arXiv:2505.16894*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024a. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, et al. 2024b. Can large language model agents simulate human trust behavior? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring collaboration mechanisms for LLM agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.

Useage of Large Language Models

We drafted the manuscript ourselves and used ChatGPT and Gemini-3 for overall refinement assistants.

A Detailed Persona Information

This section provides the full list of 65 items in the persona questionnaire collected from participants, categorized into seven thematic groups. Each item is presented in English translation first, followed by its original Korean version.

A.1 Basic Demographic Information (11 items)

Q1. What is your gender?
성별은 무엇입니까?
① Male (남자) ② Female (여자)

Q2. What is your age?
연령은 어떻게 되십니까?
Age: ____ (만 ____ 세)

Q3. What is your marital status?
결혼 유무는 어떻게 되십니까?
① Married (기혼)
② Single (미혼)
③ Divorced, Bereaved, etc. (이혼, 사별 등)

Q4. How many people currently live with you (including yourself)?
현재 함께 거주하고 있는 가구원 수는 몇 명입니까? (본인을 포함)
Number of household members: ____ (____ 명)

Q5. Who do you live with? (Select all that apply)
귀하의 가구 구성은 어떻게 되십니까? (*중복 응답)

① Grandparent(s) (조부 또는 조모)
② Parent(s) (부 또는 모)
③ Sibling(s) (형제/자매)
④ Relative(s) (친척)
⑤ Spouse (배우자)
⑥ Son(s)/Daughter(s) (아들/딸)
⑦ Alone (혼자)
⑧ Others (기타)

Q6. What is your education level?
학력은 어떻게 되십니까?
① Middle school or less (중학교 졸업 이하) → Q7
② High school graduate (고등학교 졸업) → Q7
③ Currently in university (대학교 재학) → Q6-1
④ University graduate (대학교 졸업) → Q6-1
⑤ Currently in graduate school (대학원 재학) → Q6-1
⑥ Graduate school graduate (대학원 졸업) → Q6-1

Q6-1. What is your university major?
대학교 전공은 어떻게 되십니까?
① Humanities (인문계열)
② Social sciences (사회계열)
③ Business (경상계열)
④ Education (교육계열)
⑤ Engineering (공학계열)
⑥ Natural sciences (자연계열)
⑦ Arts/Physical Education (예체능계열)
⑧ Other (기타)

Q7. What is your occupation?
직업은 어떻게 되십니까?
① Office worker (사무직)
② Civil servant (공무원)
③ Professional (전문직)
④ Sales/service (판매/서비스직)
⑤ Executive/Assembly member (고위임원/의회의원)
⑥ Agriculture/Fishery worker (농업/어업 종사자)
⑦ Technician (기술공)
⑧ Manual labor (단순노무)
⑨ Self-employed (자영업)
⑩ Housewife (전업주부)
⑪ Other (기타)

Q8. What is your father's education level?
아버지의 학력은 어떻게 되십니까?
① Middle school or less (중학교 졸업 이하)
② High school graduate (고등학교 졸업)
③ University graduate (대학교 졸업)
④ Graduate school graduate (대학원 졸업)

Q9. What is your mother's education level?

어머니의 학력은 어떻게 되십니까?
① Middle school or less (중학교 졸업 이하)
② High school graduate (고등학교 졸업)
③ University graduate (대학교 졸업)
④ Graduate school graduate (대학원 졸업)

Q10. What is your average monthly personal income (before taxes and insurance)?
지난 1년간 세금 및 보험을 제하기 전 월 평균 개인소득은 어떻게 되십니까? (가구 전체 합산 아님)
① Not applicable (해당 없음)
② Less than 1 million won (100만원 미만)
③ 1–2 million won (100–200만원)
④ 2–3 million won (200–300만원)
⑤ 3–4 million won (300–400만원)
⑥ 4–5 million won (400–500만원)
⑦ 5–6 million won (500–600만원)
⑧ 6–7 million won (600–700만원)
⑨ Over 7 million won (700만원 이상)

Q11. What is your average monthly personal expenditure?
월 평균 개인 지출은 어떻게 되십니까?
※ Excludes fixed expenses such as rent, taxes, vehicle maintenance, and transportation. Includes food, shopping, cultural and leisure expenses.
※ 월세, 세금, 차량유지비, 교통비 등 고정비 제외. 식비, 쇼핑, 문화 및 여가비 포함.
① Less than 500,000 won (50만원 미만)
② 500,000–1 million won (50–100만원 미만)
③ 1–1.5 million won (100–150만원 미만)
④ 1.5–2 million won (150–200만원 미만)
⑤ 2–2.5 million won (200–250만원 미만)
⑥ 2.5–3 million won (250–300만원 미만)
⑦ 3–3.5 million won (300–350만원 미만)
⑧ 3.5–4 million won (350–400만원 미만)
⑨ 4–4.5 million won (400–450만원 미만)
⑩ 4.5–5 million won (450–500만원 미만)
⑪ Over 5 million won (500만원 이상)

A.2 Cultural Participation (6 items)

In the past year (from September 1, 2020 to August 31, 2021), have you participated in the following activities? If yes, how many times?
귀하는 지난 1년간 (2020년 9월 1일–2021년 8월 31일) 다음의 활동을 하신 적이 있습니까? 하셨다면 얼마나 하셨습니까?

Q1. Attended an art exhibition
미술 전시회 관람
① No (없다) ② Yes (있다): (____) times (회)

Q2. Attended a classical music concert or opera
클래식 음악회, 오페라 공연 관람
① No (없다) ② Yes (있다): (____) times (회)

Q3. Read literary works
문학작품 읽기
① No (없다) ② Yes (있다): (____) times (회)

Q4. Attended a play
연극 관람
① No (없다) ② Yes (있다): (____) times (회)

Q5. Attended a musical
뮤지컬 관람
① No (없다) ② Yes (있다): (____) times (회)

Q6. Attended a dance performance (ballet, modern dance, Korean traditional dance)
무용 공연 관람 – 발레, 현대무용, 한국무용
① No (없다) ② Yes (있다): (____) times (회)

A.3 Cultural Arts Attitudes (19 items)

Please respond based on your usual thoughts.
평소 귀하께서 생각하시는 것을 토대로 응답해주십시오.

Response scale: ① Strongly Disagree (전혀 그렇지 않다), ② Disagree (그렇지 않다), ③ Neutral (보통이다), ④ Agree (그렇다), ⑤ Strongly Agree (매우 그렇다)

Q7. I enjoy beautiful things.
나는 아름다운 것들을 즐긴다.

Q8. I believe having good manners is important.
나는 좋은 매너를 갖는 것이 필요하다고 생각한다.

Q9. I believe culture is more important than material wealth.
나는 물질적 부보다 문화가 더 중요하다고 생각한다.

Q10. I enjoy things related to cultural arts.
나는 문화예술과 관련된 것들을 즐긴다.

Q11. I consider myself cultured in terms of cultural arts.
나는 문화예술 측면에서 교양있는 사람이라고 생각한다.

Q12. People think I have a taste for consuming cultural arts.
사람들은 나를 문화예술을 소비하는 취향을 가진 사람으로 생각한다.

Q13. I try to participate in activities related to cultural arts.
나는 문화예술과 관련된 활동에 참여하려고 노력한다.

Q14. I try to develop tastes related to cultural arts.
나는 문화예술과 관련된 취향을 가지려고 노력한다.

Q15. I try to refine myself by consuming cultural arts.
나는 문화예술을 소비함으로써 고상해지려고 노력한다.

Q16. I prefer classical music performances.
나는 클래식 공연 관람을 선호한다.

Q17. I prefer opera performances.
나는 오페라 공연 관람을 선호한다.

Q18. I prefer dance performances.
나는 무용 공연 관람을 선호한다.

Q19. I prefer art exhibitions.
나는 미술 전시회 관람을 선호한다.

Q20. I prefer traditional music performances.
나는 전통 음악 공연 관람을 선호한다.

Q21. I have a positive view of classical musicians' professions.
나는 클래식 음악 연주자의 직업을 긍정적으로 생각한다.

Q22. I have a positive view of vocal artists such as opera singers.
나는 오페라와 같은 성악가의 직업을 긍정적으로 생각한다.

Q23. I have a positive view of dancers' professions.
나는 무용가의 직업을 긍정적으로 생각한다.

Q24. I have a positive view of artists' professions.
나는 미술가의 직업을 긍정적으로 생각한다.

Q25. I have a positive view of traditional musicians' professions.
나는 전통 음악가의 직업을 긍정적으로 생각한다.

A.4 Cultural Arts Knowledge Level (4 items)

Please answer based on your current knowledge. These questions are meant to assess your level of cultural arts knowledge.
문화예술 지식을 측정하기 위한 질문이니 정확한 측정을 위해 평소 아시는 것을 토대로 응답해주십시오.

Q26. Who is the composer of the classical piece, Symphony "Fate"?
클래식 곡, 교향곡 '운명'의 작곡가는 누구입니까?
① Mozart (모차르트)
② Bach (바흐)
③ Beethoven (베토벤)
④ Chopin (쇼팽)
⑤ I don't know (모름)

Q27. Who is the composer of the opera "Rigoletto"?

클래식 곡, 오페라 '리골레토'의 작곡가는 누구입니까?
① Mozart (모차르트)
② Verdi (베르디)
③ Puccini (푸치니)
④ Wagner (바그너)
⑤ I don't know (모름)

Q28. Who is the painter of the artwork "Mona Lisa"?
미술작품 '모나리자'의 작가는 누구입니까?
① Van Gogh (반 고흐)
② Claude Monet (클로드 모네)
③ Pablo Picasso (파블로 피카소)
④ Leonardo da Vinci (레오나르도 다빈치)
⑤ I don't know (모름)

Q29. Who is the painter of the artwork "The Scream"?
미술작품 '절규'의 작가는 누구입니까?
① Edvard Munch (에드바르 뭉크)
② Paul Gauguin (폴 고갱)
③ Edgar Degas (에드가 드가)
④ Paul Cézanne (폴 세잔)
⑤ I don't know (모름)

A.5 Cultural Education Experience (9 items)

Please answer the following questions regarding your experience with cultural and arts education.
다음의 설문에 응답해 주십시오.

Q1. Before age 18, did you receive education related to classical music, opera, or musical theatre at school (elementary, middle, high school)?
만 18세 이전 학교내 (초, 중, 고등학교)에서 클래식음악, 오페라, 뮤지컬과 관련된 교육을 받은 경험이 있으십니까?
① No (없다) ② Yes (있다)

Q2. Before age 18, did you receive education related to art or exhibitions at school?
만 18세 이전 학교내에서 미술, 전시와 관련된 교육을 받은 경험이 있으십니까?
① No (없다) ② Yes (있다)

Q3. Before age 18, did you receive education related to classical music, opera, or musicals outside of school (e.g., private institutes, cultural centers, private lessons, clubs)?
만 18세 이전 학교외에서 클래식음악, 오페라, 뮤지컬과 관련된 교육을 받은 경험이 있으십니까? (학교외는 사설 학원, 문화센터, 개인레슨, 동호인 모임 등을 의미합니다.)
① No (없다) ② Yes (있다)

Q4. Before age 18, did you receive art or exhibition-related education outside of school?
만 18세 이전 학교외에서 미술, 전시와 관련된 교육을 받은 경험이 있으십니까? (학교외는 사설 학원, 문화센터, 개인레슨, 동호인 모임 등을 의미합니다.)
① No (없다) ② Yes (있다)

Q5. After age 19 and before September 1, 2020, did you receive classical music, opera, or musical education outside of school?
만 19세 이후부터 과거 1년 전까지 (2020년 9월 1일) 학교외에서 클래식음악, 오페라, 뮤지컬과 관련된 교육을 받은 경험이 있으십니까? (학교외는 사설 학원, 문화센터, 개인레슨, 동호인 모임 등을 의미합니다.)
① No (없다) ② Yes (있다)

Q6. After age 19 and before September 1, 2020, did you receive art or exhibition-related education outside of school?
만 19세 이후부터 과거 1년 전까지 (2020년 9월 1일) 학교외에서 미술, 전시와 관련된 교육을 받은 경험이 있으십니까? (학교외는 사설 학원, 온라인 교육, 과외, 방송 교육 등을 의미합니다.)
① No (없다) ② Yes (있다)

Q7. Between September 1, 2020 and August 31, 2021, did you receive classical music, opera, or musical education outside of school?
지난 1년 전부터 현재까지 (2020년 9월 1일–2021년 8월 31일) 학교외에서 클래식음악, 오페라, 뮤지컬 연관된 교육을 받은 경험이 있으십니까? (학교외는 사설 학원, 온라인 교육, 과외, 방송 교육 등을 의미합니다.)
① No (아니오) ② Yes (예)

Q8. Between September 1, 2020 and August 31, 2021, did you receive art or exhibition-related education outside of school?
지난 1년 전부터 현재까지 (2020년 9월 1일–2021년 8월 31일) 학교외에서 미술, 전시와 관련된 교육을 받은 경험이 있으십니까? (학교외는 사설 학원, 온라인 교육, 과외, 방송 교육 등을 의미합니다.)
① No (아니오) ② Yes (예)

Q9. Did you graduate from an arts high school majoring in art or music, or did you major in an art- or music-related field at university?
예술계 고등학교에서 미술 또는 음악 분야를 전공으로 졸업하거나, 대학교에서 미술 또는 음악 관련 학과를 전공하셨습니까?
Art-related majors: Painting, Photography, Sculpture, Craft, Cartoon, Design, etc.
- **미술 분야 전공:** 회화, 사진, 조각, 공예, 만화, 디자인 등
Music-related majors: Classical music, Korean traditional music, popular music, etc
- **음악 분야 전공:** 클래식음악, 국악, 대중음악 등
① No (아니오) ② Yes (예)

A.6 Institution Satisfaction (4 items)

Q30. Have you ever visited the National Museum of Korea?
귀하께서는 국립중앙박물관에 방문하신 경험이 있습니까?
① No (아니오) → Skip to Q31
② Yes (예) → Go to Q30-1

Q30-1. How satisfied were you overall with your visit to the National Museum of Korea?
귀하께서 국립중앙박물관에 방문 후 느꼈던 전반적인 만족도는 어떠십니까?
① Very dissatisfied (매우 불만족스럽다)
② Dissatisfied (불만족스럽다)
③ Neutral (보통이다)
④ Satisfied (만족스럽다)
⑤ Very satisfied (매우 만족스럽다)

Q31. Have you ever visited the Seoul Arts Center (예술의전당)?
귀하께서는 예술의전당에 방문하신 경험이 있습니까?
① No (아니오) → Skip to next section
② Yes (예) → Go to Q31-1

Q31-1. How satisfied were you overall with your visit to the Seoul Arts Center?
귀하께서 예술의전당에 방문 후 느꼈던 전반적인 만족도는 어떠십니까?
① Very dissatisfied (매우 불만족스럽다)
② Dissatisfied (불만족스럽다)
③ Neutral (보통이다)
④ Satisfied (만족스럽다)
⑤ Very satisfied (매우 만족스럽다)

A.7 Social Norms and Attitudes (10 items)

Please respond based on your usual thoughts.
평소 귀하께서 생각하시는 것을 토대로 응답해주십시오.

Response scale: ① Strongly Disagree (전혀 그렇지 않다), ② Disagree (그렇지 않다), ③ Neutral (보통이다), ④ Agree (그렇다), ⑤ Strongly Agree (매우 그렇다)

Q1. I feel good when I act according to my conscience.
나는 양심에 따라 행동을 할 때 기분이 좋다.

Q2. I feel guilty when I pay less than an appropriate amount.
나는 적정금액보다 낮은 금액을 지불했을 때 양심에 걸린다.

Q3. I feel a moral obligation to buy cultural products (e.g., performances, exhibitions) even if they are expensive, if it is for charity.
나는 자선단체 기부를 위해서라면 문화예술상품(공연, 전시 등)이 비싸더라도 구매해야 한다는 도덕적 의무감을 느낀다.

Q4. I feel a moral obligation to attend cultural events (e.g., performances, exhibitions) for charity even if I am busy.
나는 자선단체 기부를 위해서라면 문화예술상품(공연, 전시 등)에 가려고 노력해야 한다는 도덕적 의무감을 느낀다.

Q5. If the quality of cultural products (e.g., performances, exhibitions) is good, I don't mind paying more.

나는 문화예술상품(공연, 전시 등)의 질이 좋다면 더 많은 돈을 지불할지라도 개의치 않는다.

Q6. People around me (family, friends, neighbors) will understand if I pay more to attend cultural events for charity.
나의 주변 사람들(가족, 친구, 이웃)은 자선단체 기부를 위해서라면 문화예술상품(공연, 전시 등) 관람에 더 많은 돈을 지불하는 나의 행동에 공감할 것이다.

Q7. Other people's attitudes and actions influence me.
다른 사람들의 태도나 행동이 나에게 영향을 미친다.

Q8. I sometimes act against my will because I am conscious of others' criticism.
다른 사람의 비난을 의식하여 원치 않는 행동을 하는 경우가 있다.

Q9. I am willing to pay the same amount that others paid, even if I would normally pay less.
나는 사람들의 눈을 의식하여 다른 사람이 지불한 금액을 기꺼이 따라 지불할 용의가 있다.

Q10. I am willing to pay more if others around me also pay a higher amount.
다른 사람들의 눈을 의식하여 다른 사람이 높은 비용을 지불하더라도 기꺼이 따라할 용의가 있다.

B Detailed Question Form

This section provides the full structure of the survey instrument used in the PWYW experiment. Figures 15 to 19 show the original Korean question forms presented to participants, including survey introductions and question items for both the art and music domains, each followed by an English translation. These materials clarify the experimental context and the input provided to both human participants and the LLMs.

C Detailed Human Experiment Analysis

This section details the human experiment with 522 Korean adults and reports the estimated Heckman two-step models that serve as the human reference for the group-level metrics (CSA, SSA, Jaccard) in the main text.

C.1 Art Exhibition Analysis

We employed the Heckman two-step model to analyze participants' Willingness-to-Pay (WTP) for art exhibitions, addressing potential sample selection bias inherent in Pay-What-You-Want (PWYW) scenarios.

Step 1: Selection Equation (Probit Model) In this step, we compute the inverse Mills ratio (λ_i), which corrects selection bias in the subsequent step. Step 1 estimates the probability of participation using a probit regression:

$$\Pr(\texttt{A_Q2} = 1) = \Phi(\gamma' Z), \qquad (1)$$

where the dependent variable `A_Q2` is a binary indicator for willingness to participate in PWYW for the art exhibition (1 = willing, 0 = not willing), and $\Phi(\cdot)$ is the CDF of the standard normal distribution.

The vector Z includes the following independent variables:

- `edu_f`: Father's education level (1=middle school or below, 2=high school, 3=university, 4=graduate school) from A.1, Q8
- `edu_m`: Mother's education level (same coding) from A.1, Q9
- `edu_s`: Respondent's education level (same coding) from A.1, Q6
- `art_act_dum`: Art activity participation dummy from A.2, Q1 (1 if attended an art exhibition ≥ 1 time in the past year, 0 otherwise)
- `art_att_ave`: Average art attitude score, mean of A.3, Q7, Q10–Q15, Q19, and Q24
- `art_qz_tot_sco`: Art knowledge total score, sum of correct answers from A.4, Q28–Q29 (maximum 2)
- `art_ca_edu_b18`: Art cultural education before age 18, from A.5, Q2+Q4
- `art_ca_edu_a19`: Art cultural education after age 19, from A.5, Q6+Q8
- `inc_mo`: Monthly personal income from A.1, Q10

- marital_dum: Marital status dummy (1=married, 0=single/other) from A.1, Q3
- gen_dum: Gender dummy (0=male, 1=female) from A.1, Q1

The inverse Mills ratio for each observation Z_i is obtained as

$$\lambda_i = \frac{\mathcal{N}(\gamma' Z_i \mid 0, 1^2)}{\Phi(\gamma' Z_i)}, \quad (2)$$

where $\mathcal{N}(\cdot \mid 0, 1^2)$ is the PDF of the standard normal distribution.

Step 2: Outcome Equation (Linear Regression with Selection Correction) Step 2 estimates the WTP amount, corrected for selection bias:

$$\texttt{art_pwyw_log_wtp} = \beta' X + \delta\lambda + \varepsilon, \quad (3)$$

where art_pwyw_log_wtp is the natural logarithm of the WTP amount for the art exhibition in KRW, and X includes all Z variables plus social_n (social norm score from A.7, Q7–Q10) and moral_n (moral norm score from A.7, Q2–Q4).

C.2 Music Performance Analysis

The music analysis follows the same methodology with domain-specific variables.

Step 1: Selection Equation (Probit Model)

$$\Pr(\texttt{M_Q2} = 1) = \Phi(\gamma' Z), \quad (4)$$

where M_Q2 is a binary indicator for willingness to participate in PWYW for the music performance. The vector Z includes edu_f, edu_m, edu_s (identical to the art model); mus_act_dum (music activity dummy from A.2, Q2); mus_att_ave (average music attitude score, mean of A.3, Q7, Q10–Q18, Q20–Q23, Q25); mus_qz_tot_sco (music knowledge score from A.4, Q26–Q27); mus_ca_edu_b18 and mus_ca_edu_a19 (music cultural education from A.5); and the sociodemographic controls inc_mo, marital_dum, gen_dum.

Step 2: Outcome Equation

$$\texttt{mus_pwyw_log_wtp} = \beta' X + \delta\lambda + \varepsilon, \quad (5)$$

where X includes all Z variables plus social_n_ave and moral_n_ave.

C.3 Data Processing and Quality Control

Sample Selection Criteria.

- Minimum 10 observations per group to ensure model convergence
- Minimum 2 categories in the selection variable to satisfy variance requirements
- Minimum 5 observations with positive WTP among those willing to participate

Variable Transformations.

- **WTP Amount**: natural logarithm transformation to address skewness and heteroskedasticity.
- **Cultural Education**: binary coding into formal (school) and informal (outside school) experiences.
- **Attitude Scores**: standardized averaging across the relevant Likert items.
- **Knowledge Scores**: summation of correctly answered items.

D Persona Input and Implementation

This section describes how the persona was supplied to the LLMs and the implementation of each condition. All RQ2 conditions run on the Human-guided setting of RQ1: the assistant turns for earlier questions are filled with the actual human responses, and the scenario branches to Q2.1 or Q3 according to the Q2 answer. The conditions differ only in how the persona is supplied.

D.1 Persona Format

Survey Format. In the Survey format, persona information is presented as a structured question–answer list in Korean, mirroring the original questionnaire. Each persona is a compilation of responses across all 65 survey items from the seven categories in Appendix A:

```
What is your gender? : [Male/Female]
What is your age?    : [Age in years]
...
```

Storytelling Format. In the Storytelling format, the same 65 items are rewritten into a biographical narrative with GPT-4o. Each narrative is checked by keyword matching against the 65 items and regenerated until all items are present, so no item is dropped. The generation prompt is shown below.

Original Korean Prompt:

아래 페르소나 정보를 기반으로, 한 인물이 자신의 삶에서 예술을 어떻게 받아들이고 이해해 왔는지를 고백적으로 풀어내는 허구 수필을 작성하세요.

- 텍스트에는 어떤 소제목, 부제, 섹션 제목도 포함되어서는 안 됩니다.
- 자연스러운 단락 구분은 허용되지만, 시각적 강조나 구획 나눔은 삼가세요.
- 전체 흐름은 시간 순이 아니어도 되며, 자유롭게 떠오르는 기억의 방식으로 구성해도 좋습니다.
- 지나치게 극적인 연출은 피하며 사실적인 정서에 기반해 구성하세요.
- 분량은 약 5000 토큰으로, 읽는 이로 하여금 한 사람의 내면을 들여다보는 느낌을 받을 수 있어야 합니다.

PERSONA INFORMATION

English Translation:

Based on the persona information below, write a confessional fictional essay that tells how a person has embraced and understood art throughout their life.

- *The text should not include any subtitles or section titles.*
- *Natural paragraph breaks are allowed, but visual emphasis or partitioning should be avoided.*
- *The flow need not be chronological and may follow freely emerging memories.*
- *Avoid overly dramatic framing; compose based on realistic emotion.*
- *The length should be about 5000 tokens, so that the reader feels they are looking into a person's inner life.*

[PERSONA INFORMATION]

Narratives were generated with the OpenAI API (GPT-4o, temperature 0, max tokens 6,000).

D.2 Prompting Method

CoT. CoT keeps the same input as the base prompt and appends a Korean instruction after each question. The instruction directs the model to reason step by step from the persona and the scenario, to analyze the options, and to state its reasoning before giving a final answer. It also requires the model to commit to exactly one option, which prevents non-committal answers. The verbatim instruction is shown in Figure 6.

RAG. For each persona, a separate FAISS index is built over its own 65 persona items, so retrieval is confined to the target individual. For each question, the question text is the query, and the top k=10 items are retrieved by cosine similarity and placed in the system prompt in place of the full persona. The index is queried anew for every question, so the supplied items change as the scenario proceeds.

Few-shot. A single index is built over all 522 personas, and for each target persona the three most similar other personas are selected as in-context examples, with their actual survey responses prepended before the question. For Q1, the three personas are the nearest to the target by persona-text cosine similarity. From Q2 onward, the candidate pool is first restricted to personas that gave the same answer as the target on the preceding question, and the three nearest within that pool are selected, so the examples match the target both in persona and in response trajectory; Q1 has no preceding answer and omits this filter. The examples are re-selected for every question.

Embedding and Generation. All retrieval uses `text-embedding-3-large` for both personas and queries. The temperature is fixed at 0 for all conditions. The full prompt templates for the four prompting methods are shown in Figures 5–8.

E Detailed Result for RQ1

This section presents the regression outcomes of the RQ1 experimental conditions using the Heckman two-step model. Each table pair (Tables 3 to 6) summarizes how the LLMs (GPT-4o, LLaMA, Qwen) reflect persona attributes when predicting willingness-to-pay decisions, and is provided for reference to illustrate model-level behavioral patterns under the two interaction conditions.

Sequential Condition. See Tables 3 and 4.

Human-Guided Condition. See Tables 5 and 6.

F Detailed Result for RQ2

This section presents the regression outcomes of the RQ2 conditions using the Heckman two-step model. Each table pair (Tables 7 to 18) summarizes how the LLMs reflect persona attributes under different combinations of persona format (Survey vs. Storytelling) and prompting method (Base, CoT, RAG, Few-shot). For some Few-shot conditions the estimates were unstable, with ρ at the boundary and large σ; these are noted in the main text and the corresponding tables should be read with caution.

Survey Format – CoT. See Tables 7 and 8.

Survey Format – RAG. See Tables 9 and 10.

Survey Format – Few-shot. See Tables 11 and 12.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	0.01	0.01	0.21	1.26	0.04	21694.32	1.00	0.00	0.01	0.02	0.51	0.66
edu_m	-0.01	0.01	0.12	-1.57	-0.03	25017.10	1.00	0.00	0.00	0.02	0.90	0.13
mus_act_dum	0.02	0.01	p<0.05	2.00	-0.02	303822.33	1.00	0.00	0.06	0.03	p<0.05	2.38
mus_att_ave	0.02	0.02	0.24	1.18	0.19	22875.96	1.00	0.00	0.26	0.03	p<0.0001	0.02
mus_ca_edu_a19	0.00	0.01	0.52	0.64	-0.03	24289.04	1.00	0.00	0.03	0.02	0.19	10.00
mus_ca_edu_b18	0.00	0.01	0.64	0.47	0.06	18689.85	1.00	0.00	0.00	0.02	0.86	1.31
mus_qz_tot_sco	0.01	0.01	0.42	0.81	0.12	25466.45	1.00	0.00	0.05	0.02	p<0.05	0.18
moral_n_ave	0.00	0.01	0.90	0.12	0.04	21456.43	1.00	0.00	0.04	0.02	p<0.05	2.12
social_n_ave	0.00	0.01	0.65	-0.46	0.04	19449.28	1.00	0.00	0.00	0.02	0.94	0.08
inc_mo	0.00	0.00	0.87	-0.16	0.03	7365.53	1.00	0.00	0.00	0.01	0.65	0.45
matarial_dum	0.00	0.01	0.71	0.37	0.08	27206.99	1.00	0.00	0.04	0.02	0.66	1.88
constant	8.98	0.08	p<0.001	106.26	7.33	103897.22	1.00	0.00	7.28	0.12	p<0.0001	61.51
Model Statistics												
Wald χ^2	9.59				3.1×10^{-10}				183.19			
ρ	0.69				1				0.73			
σ	0.07				299218.13				0.25			
Log Likelihood	-201.81				-				-83.74			
McFadden R^2	0.44				-				0.59			
$N_{\text{nonselected}}$	268				1				69			
N_{selected}	254				521				453			

Table 3: Regression outcomes of Sequential Condition on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.01	0.02	0.64	-0.47	0.04	21694.32	1.00	0.00	0.01	0.02	0.51	0.66
edu_m	0.00	0.03	0.10	0.01	-0.03	25017.10	1.00	0.00	0.00	0.02	0.90	0.13
mus_act_dum	0.05	0.04	0.19	1.30	-0.02	30822.33	1.00	0.00	0.06	0.03	p<0.05	2.38
mus_att_ave	0.19	0.06	p<0.01	3.17	0.19	22875.96	1.00	0.00	0.26	0.03	p<0.0001	10.00
mus_ca_edu_a19	0.00	0.02	0.92	-0.11	-0.03	24289.04	1.00	0.00	0.03	0.02	0.19	1.31
mus_ca_edu_b18	0.03	0.02	0.16	-1.39	0.06	18689.85	1.00	0.00	0.00	0.02	0.86	0.18
mus_qz_tot_sco	0.00	0.02	0.85	-0.19	0.12	25466.48	1.00	0.00	0.05	0.02	p<0.05	2.15
moral_n_ave	0.04	0.02	0.09	1.67	0.04	21456.43	1.00	0.00	0.04	0.02	p<0.05	2.12
social_n_ave	0.01	0.02	0.08	0.51	0.04	19449.28	1.00	0.00	0.00	0.02	0.94	0.08
inc_mo	0.01	0.01	0.71	1.74	0.03	7365.53	1.00	0.00	0.00	0.01	0.65	0.45
matarial_dum	0.01	0.03	p<0.05	0.36	0.08	27206.99	1.00	0.00	0.04	0.02	0.06	1.88
constant	8.87	0.24	p<0.0001	36.91	7.33	103897.22	1.00	0.00	7.28	0.12	p<0.0001	61.51
Model Statistics												
Wald χ^2	9.59				3.1×10^{-10}				183.19			
ρ	0.69				1				0.73			
σ	0.07				299218.13				0.25			
Log Likelihood	-201.81				-				-83.74			
McFadden R^2	0.44				-				0.59			
$N_{\text{nonselected}}$	268				1				69			
N_{selected}	254				521				453			

Table 4: Regression outcomes of Sequential Condition on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.02	0.05	0.73	-0.34	0.03	0.08	0.71	0.37	-0.09	0.17	0.58	-0.55
edu_m	0.04	0.06	0.47	0.72	-0.02	0.09	0.82	-0.22	0.12	0.20	0.53	0.62
mus_act_dum	-0.04	0.12	0.73	-0.34	0.14	0.19	0.47	-0.72	-0.25	0.41	0.54	-0.62
mus_att_ave	0.07	0.11	0.48	0.70	0.11	0.17	0.55	0.60	-0.25	0.37	0.50	-0.67
mus_ca_edu_a19	0.01	0.04	0.75	0.32	0.04	0.05	0.50	0.68	-0.06	0.13	0.63	-0.48
mus_ca_edu_b18	-0.04	0.05	0.40	-0.84	-0.03	0.08	0.76	-0.31	-0.12	0.18	0.50	-0.68
mus_qz_tot_sco	-0.02	0.05	0.67	-0.43	0.03	0.08	0.72	0.35	-0.08	0.17	0.62	-0.50
moral_n_ave	0.04	0.03	0.19	1.30	0.10	0.05	p<0.05	2.07	-0.02	0.10	0.87	-0.17
social_n_ave	0.00	0.03	0.93	0.08	0.03	0.05	0.49	0.69	0.07	0.09	0.46	0.74
inc_mo	0.01	0.01	0.28	1.09	0.03	0.07	0.05	1.93	-0.01	0.04	0.76	-0.31
matarial_dum	0.01	0.05	0.90	0.13	0.00	0.92	0.94	0.07	-0.03	0.16	0.83	-0.22
constant	8.50	0.56	p<0.0001	15.28	7.68	1.65	p<0.0001	8.32	9.91	1.95	p<0.0001	5.08
Model Statistics												
Wald χ^2	20.11				17.88				2.04			
ρ	-1				0.02				-1			
σ	0.46				0.62				1.60			
Log Likelihood	-250.78				-250.78				-250.78			
McFadden R^2	0.08				0.08				0.08			
$N_{\text{nonselected}}$	113				113				113			
N_{selected}	409				409				409			

Table 5: Regression outcomes of Human-Guided Condition on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.09	0.05	0.05	-1.95	-0.05	0.06	0.41	-0.82	-0.04	0.03	0.20	-1.29
edu_m	-0.03	0.05	0.54	0.61	-0.02	0.05	0.65	-0.46	-0.02	0.03	0.63	-0.48
mus_act_dum	-0.09	0.11	0.39	-0.85	-0.02	0.12	0.89	-0.13	-0.03	0.07	0.70	-0.38
mus_att_ave	-0.08	0.14	0.57	-0.57	-0.10	0.17	0.57	-0.57	0.00	0.10	0.98	-0.02
mus_ca_edu_a19	0.02	0.04	0.69	0.40	0.04	0.05	0.38	0.88	0.03	0.03	0.38	0.87
mus_ca_edu_b18	0.00	0.04	0.91	0.11	-0.02	0.05	0.61	-0.51	-0.02	0.03	0.51	-0.66
mus_qz_tot_sco	0.05	0.05	0.30	1.04	0.11	0.06	p<0.05	2.03	0.06	0.03	0.07	1.84
moral_n_ave	0.13	0.03	p<0.0001	3.95	0.13	0.04	p<0.0001	2.99	0.02	0.02	0.37	-0.90
social_n_ave	-0.01	0.03	0.73	-0.34	-0.02	0.04	0.70	-0.39	0.02	0.02	0.38	0.88
inc_mo	0.03	0.01	p<0.05	2.27	-0.03	0.02	0.09	1.70	0.01	0.01	0.44	0.78
matarial_dum	-0.04	0.05	0.46	-0.73	-0.04	0.06	0.46	-0.75	0.01	0.04	0.82	0.23
constant	9.61	0.59	p<0.0001	16.24	9.75	0.72	p<0.0001	13.52	9.77	0.42	p<0.0001	23.22
Model Statistics												
Wald χ^2	39.04				31.44				20.32			
ρ	-0.98				-0.54				-0.87			
σ	0.49				0.53				0.34			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 6: Regression outcomes of Human-Guided Condition on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	0.00	0.05	0.99	-0.01	-0.01	0.02	0.60	-0.52	0.03	0.03	0.31	1.01
edu_m	0.01	0.06	0.92	0.10	0.04	0.03	0.15	1.44	-0.01	0.03	0.84	-0.20
mus_act_dum	0.08	0.13	0.51	0.66	0.00	0.05	0.95	0.06	0.08	0.06	0.19	1.32
mus_att_ave	0.21	0.11	0.07	1.81	0.05	0.05	0.29	1.05	0.14	0.06	p<0.05	2.47
mus_ca_edu_a19	0.00	0.04	0.97	-0.04	0.00	0.02	0.80	0.25	0.01	0.02	0.75	0.32
mus_ca_edu_b18	0.03	0.05	0.56	0.58	-0.01	0.02	0.75	-0.32	-0.01	0.03	0.74	-0.33
mus_qz_tot_sco	0.04	0.05	0.48	0.71	-0.02	0.02	0.32	-0.99	0.01	0.03	0.72	0.36
moral_n_ave	0.04	0.03	0.16	1.42	0.03	0.01	0.07	1.79	0.01	0.02	0.43	0.80
social_n_ave	-0.01	0.03	0.79	-0.27	0.00	0.01	0.80	0.25	0.00	0.01	0.78	-0.28
inc_mo	0.01	0.01	0.48	0.70	0.00	0.01	0.38	0.88	0.00	0.01	0.79	-0.27
matarial_dum	0.04	0.05	0.39	0.86	0.01	0.02	0.58	0.55	0.01	0.02	0.53	0.62
constant	7.71	0.60	p<0.0001	12.85	8.62	0.27	p<0.0001	32.51	8.11	0.29	p<0.0001	27.55
Model Statistics												
Wald χ^2	16.42				37.77				26.87			
ρ	1.00				-0.43				0.86			
σ	0.49				0.18				0.23			
Log Likelihood	-250.78				-250.78				-250.78			
McFadden R^2	0.08				0.08				0.08			
$N_{\text{nonselected}}$	113				113				113			
N_{selected}	409				409				409			

Table 7: Regression outcomes of Survey_CoT on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.10	0.07	0.13	-1.50	-0.03	0.06	0.53	-0.63	-0.06	0.05	0.28	-1.08
edu_m	0.03	0.07	0.62	0.49	-0.02	0.05	0.64	-0.46	0.00	0.05	0.93	-0.08
mus_act_dum	-0.16	0.15	0.29	-1.07	-0.02	0.12	0.89	-0.14	-0.03	0.12	0.81	-0.24
mus_att_ave	-0.07	0.20	0.71	-0.37	-0.12	0.17	0.50	-0.68	-0.16	0.16	0.34	-0.96
mus_ca_edu_a19	0.03	0.06	0.60	0.52	0.03	0.05	0.50	0.68	0.03	0.05	0.55	0.61
mus_ca_edu_b18	-0.04	0.06	0.53	-0.62	-0.02	0.05	0.63	-0.48	-0.03	0.05	0.52	-0.64
mus_qz_tot_sco	0.04	0.07	0.53	0.63	0.11	0.06	0.05	1.95	0.11	0.06	p<0.05	2.05
moral_n_ave	0.07	0.05	0.10	1.63	0.13	0.04	p<0.0001	3.12	0.09	0.04	p<0.05	2.20
social_n_ave	0.02	0.04	0.59	0.54	-0.02	0.04	0.71	-0.38	0.04	0.04	0.32	0.98
inc_mo	0.02	0.02	0.31	1.01	0.02	0.02	0.14	1.48	0.02	0.02	0.29	1.06
matarial_dum	0.01	0.07	0.92	0.10	-0.07	0.06	0.27	-1.11	-0.03	0.06	0.58	-0.56
constant	9.69	0.83	p<0.0001	11.68	9.80	0.72	p<0.0001	13.65	10.00	0.69	p<0.0001	14.43
Model Statistics												
Wald χ^2	14.54				29.98				23.01			
ρ	-1.00				-0.58				-0.90			
σ	0.70				0.53				0.56			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 8: Regression outcomes of Survey_CoT on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	0.01	0.04	0.75	-0.32	0.02	0.02	0.46	0.73	0.00	0.01	0.97	-0.04
edu_m	-0.01	0.04	0.84	-0.21	0.02	0.02	0.51	0.65	0.01	0.01	0.39	0.85
mus_act_dum	0.08	0.09	0.36	-1.59	-0.03	0.05	0.56	-0.58	0.00	0.02	0.88	0.15
mus_att_ave	0.13	0.08	0.11	-0.37	0.06	0.05	0.21	1.26	0.00	0.02	0.81	0.25
mus_ca_edu_a19	-0.01	0.03	0.71	-0.57	0.00	0.02	0.98	0.03	-0.01	0.01	0.08	-1.73
mus_ca_edu_b18	0.02	0.04	0.57	-0.17	-0.02	0.02	0.32	-0.99	0.00	0.01	0.92	-0.10
mus_qz_tot_sco	-0.01	0.04	0.86	0.14	0.00	0.02	0.83	-0.21	-0.01	0.01	0.05	-1.96
moral_n_ave	0.00	0.02	0.89	-0.03	-0.01	0.01	0.45	-0.76	0.00	0.00	0.32	-0.99
social_n_ave	0.00	0.02	0.97	-0.06	0.00	0.01	0.81	-0.24	0.01	0.00	0.17	1.36
inc_mo	0.00	0.01	0.95	0.36	0.00	0.00	0.70	-0.39	0.00	0.00	0.38	0.87
matarial_dum	0.01	0.04	0.72	0.36	0.01	0.02	0.70	-0.38	0.00	0.01	0.83	-0.21
constant	8.35	0.44	p<0.0001	19.19	8.41	0.26	p<0.0001	-32.65	8.51	0.09	p<0.0001	-95.45
Model Statistics												
Wald χ^2	10.45				28.81				15.87			
ρ	1.00				-0.25				-0.33			
σ	0.36				0.18				0.06			
Log Likelihood	-250.78				-250.78				-250.78			
McFadden R^2	0.08				0.08				0.08			
$N_{\text{nonselected}}$	113				113				113			
N_{selected}	409				409				409			

Table 9: Regression outcomes of Survey_RAG on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.05	0.03	0.14	-1.49	-0.06	0.03	0.06	-1.85	-0.08	0.06	0.14	-1.48
edu_m	0.00	0.03	0.97	-0.03	0.02	0.03	0.56	0.58	-0.01	0.05	0.86	-0.18
mus_act_dum	-0.03	0.07	0.70	-0.38	-0.06	0.07	0.39	-0.85	-0.24	0.12	p<0.05	-2.02
mus_att_ave	0.02	0.10	0.82	0.23	-0.05	0.10	0.61	-0.51	0.11	0.17	0.53	-0.63
mus_ca_edu_a19	0.03	0.03	0.23	1.21	0.02	0.03	0.58	0.55	-0.05	0.05	0.27	-1.11
mus_ca_edu_b18	-0.01	0.03	0.71	-0.38	-0.04	0.03	0.14	-1.49	0.00	0.05	0.95	-0.06
mus_qz_tot_sco	0.06	0.03	p<0.05	2.05	0.03	0.03	0.30	1.04	0.00	0.06	0.94	-0.07
moral_n_ave	0.04	0.02	0.13	1.51	0.03	0.02	0.26	1.12	-0.07	0.05	0.13	-1.51
social_n_ave	0.03	0.02	0.22	1.22	0.02	0.02	0.31	1.02	0.07	0.04	0.09	-1.68
inc_mo	0.01	0.01	0.14	1.49	0.00	0.01	0.65	0.46	0.01	0.02	0.38	-0.88
matarial_dum	0.01	0.03	0.82	0.23	0.01	0.04	0.88	0.15	-0.06	0.06	0.35	-0.93
constant	9.32	0.41	p<0.0001	22.70	10.08	0.41	p<0.0001	24.34	9.81	0.73	p<0.0001	-13.47
Model Statistics												
Wald χ^2	30.46				14.85				22.56			
ρ	-0.45				-1.00				-0.11			
σ	0.30				0.35				0.51			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 10: Regression outcomes of Survey_RAG on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.22	0.34	0.53	-0.63	-0.04	0.40	0.92	-0.10	0.03	0.03	0.46	0.74
edu_m	0.26	0.40	0.51	-0.66	0.05	0.46	0.91	0.11	-0.03	0.04	0.51	-0.65
mus_act_dum	-0.48	0.82	0.56	-0.59	0.32	0.98	0.75	0.32	0.04	0.08	0.62	-0.50
mus_att_ave	-0.10	0.75	0.89	-0.14	-0.32	0.89	0.72	-0.35	0.09	0.08	0.24	1.17
mus_ca_edu_a19	-0.02	0.27	0.95	-0.06	-0.29	0.30	0.33	-0.97	0.00	0.03	0.97	-0.04
mus_ca_edu_b18	-0.21	0.36	0.55	-0.60	-0.12	0.42	0.78	-0.28	0.02	0.04	0.53	0.63
mus_qz_tot_sco	0.12	0.34	0.73	0.34	0.16	0.40	0.68	0.41	0.02	0.03	0.55	0.59
moral_n_ave	0.21	0.20	0.31	1.02	-0.25	0.25	0.32	-0.99	0.00	0.02	0.93	-0.08
social_n_ave	-0.05	0.19	0.79	-0.26	0.23	0.23	0.32	0.99	0.02	0.02	0.36	0.92
inc_mo	0.03	0.08	0.70	0.38	-0.09	0.09	0.35	-0.94	0.00	0.01	0.57	0.56
matarial_dum	-0.03	0.32	0.92	-0.09	-0.75	0.36	p<0.05	-2.07	0.00	0.03	0.95	0.06
constant	9.01	3.93	p<0.05	2.30	-8.45	4.71	0.07	1.79	8.44	0.40	p<0.0001	21.17
Model Statistics												
Wald χ^2	3.82				21.44				5.55			
rho	-1.00				-0.76				1.00			
σ	3.23				3.53				0.33			
Log Likelihood	-250.78				-250.78				-250.78			
McFadden R^2	0.08				0.08				0.08			
$N_{\text{nonselected}}$	113				113				113			
N_{selected}	409				409				409			

Table 11: Regression outcomes of Survey_few-shot on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.14	0.12	0.24	-1.17	-0.04	0.21	0.84	-0.20	-0.03	0.02	0.14	-1.47
edu_m	0.05	0.12	0.68	-0.41	-0.02	0.21	0.94	-0.07	0.01	0.02	0.78	0.27
mus_act_dum	-0.21	0.28	0.45	-0.75	-0.36	0.47	0.45	-0.75	-0.01	0.05	0.90	-0.13
mus_att_ave	-0.27	0.36	0.46	-0.74	-0.54	0.62	0.39	-0.87	0.06	0.07	0.44	0.77
mus_ca_edu_a19	0.01	0.12	0.93	0.09	0.01	0.20	0.98	0.03	0.04	0.02	0.08	1.74
mus_ca_edu_b18	-0.05	0.11	0.67	-0.43	-0.10	0.18	0.59	-0.54	-0.02	0.02	0.42	-0.80
mus_qz_tot_sco	-0.01	0.12	0.96	-0.05	0.11	0.21	0.60	0.52	0.04	0.02	0.07	1.78
moral_n_ave	0.11	0.08	0.20	1.29	0.18	0.14	0.21	1.25	0.03	0.02	0.06	1.85
social_n_ave	0.01	0.08	0.91	0.11	0.08	0.14	0.57	0.57	0.01	0.02	0.45	0.76
inc_mo	0.03	0.04	0.39	0.86	0.02	0.06	0.73	-0.34	0.03	0.01	p<0.0001	4.12
matarial_dum	0.05	0.13	0.71	0.38	-0.03	0.23	0.90	-0.13	-0.02	0.03	0.41	-0.83
constant	10.52	1.54	p<0.0001	6.85	11.30	2.63	p<0.0001	4.29	9.52	0.30	p<0.0001	-31.25
Model Statistics												
Wald χ^2	5.72				3.92				45.74			
rho	-1.00				-1.00				-0.55			
σ	1.29				2.21				0.22			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 12: Regression outcomes of Survey_few-shot on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

Base

[System] 귀하의 성별: 여성 / 연령: 38 / 클래식공연 선호: 매우그렇다 / · · ·

[User] 설문 소개문 / 기관 설명 / 가상 상황 설명

[User] Q1. 특별전시를 9,000원을 지불하고 관람할 의도가 있으십니까? ① 아니다 ② 그렇다 둘 중 하나를 선택하세요.

[Assistant] *{Model Response}*

[User] Q2. PWYW 방식이라면 방문하실 의향이 있으십니까? ① 아니다 (→ Q3) ② 그렇다 (→ Q2-1)

[Assistant] *{Model Response}*

[User] Q2-1 or **Q3** (branched by Q2 response)

Figure 5: Base prompt template.

Storytelling Format – Base. See Tables 13 and 14.

Storytelling Format – CoT. See Tables 15 and 16.

Storytelling Format – RAG. See Tables 17 and 18.

G RQ2 Response Distribution Heatmaps

This section provides the full response-distribution heatmaps underlying the diversity analysis in the main text (Figure 4). Figures 9 to 14 show how LLM-generated willingness-to-pay responses are distributed across price points for each combination of persona format and prompting method. Each heatmap makes visible the concentration that the distinct-value counts in the main text summarize: human responses spread across the full range of price points, while LLM responses fall on a small number of them.

Survey Format – CoT. See Figure 9.

Survey Format – RAG. See Figure 10.

Survey Format – Few-shot. See Figure 11.

Storytelling Format – Base. See Figure 12.

Storytelling Format – CoT. See Figure 13.

Storytelling Format – RAG. See Figure 14.

CoT

[System] *(identical to Base)*

[User] Q1. 특별전시를 9,000원을 지불하고 관람할 의도가 있으십니까? ① 아니다 ② 그렇다
자신의 페르소나와 주어진 상황을 바탕으로 단계별로 생각해 보세요. 정확히 하나의 옵션을 선택하고 최종 답변을 명확하게 명시해야 합니다. 아니라면, "상황에 따라 다릅니다"라고 말하거나 여러 답변을 제공하세요. 먼저 자신의 추론을 제시한 다음 최종 답변을 제시하세요.
*Think step-by-step based on your persona and the given situation. You must choose exactly **one** option and clearly state the final answer. Do **not** say "it depends" or give multiple answers. First give your reasoning, then state your final answer.*

[Assistant] *{Reasoning · · · Final Answer: ① / ②}*

▷ The italic instruction is appended to every question (Q1, Q2, Q2-1, Q3).

Figure 6: CoT prompt template. A zero-shot chain-of-thought instruction is appended to every question.

RAG

[System] *Retrieved doc 1:* 성별: 여성 / 클래식 선호: 그렇다 · · ·
Retrieved doc 2: 성별: 남성 / 미술전시 선호: 없다 · · ·
⋮ (top-k=10, query = question text)

[User] Q1. 특별전시를 9,000원을 지불하고 관람할 의도가 있으십니까? ① 아니다 ② 그렇다

[Assistant] *{Model Response}*

▷ System prompt is overwritten with retrieved docs before every question.

Figure 7: RAG prompt template. The system prompt is replaced with top-k retrieved documents before each question.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	0.02	0.07	0.76	0.30	-0.07	0.08	0.38	-0.87	0.00	0.00	0.73	-0.35
edu_m	-0.01	0.08	0.91	-0.11	0.09	0.09	0.30	1.03	0.00	0.00	0.20	1.27
mus_act_dum	0.12	0.17	0.48	0.71	-0.13	0.19	0.50	-0.68	0.00	0.00	0.42	-0.80
mus_att_ave	0.19	0.15	0.21	1.25	0.04	0.17	0.82	0.22	0.00	0.00	0.36	-0.92
mus_ca_edu_a19	-0.03	0.05	0.53	-0.63	0.01	0.06	0.82	0.22	0.00	0.00	0.13	1.53
mus_ca_edu_b18	0.05	0.07	0.50	0.68	-0.03	0.08	0.74	-0.33	0.00	0.00	0.75	-0.32
mus_qz_tot_sco	0.01	0.07	0.90	0.12	-0.06	0.08	0.46	-0.74	0.00	0.00	0.90	-0.12
moral_n_ave	0.00	0.04	0.94	-0.08	-0.02	0.05	0.72	-0.36	0.00	0.00	0.07	1.79
social_n_ave	0.00	0.04	0.91	-0.11	-0.01	0.04	0.83	-0.21	0.00	0.00	0.63	-0.48
inc_mo	-0.01	0.02	0.63	-0.48	0.01	0.02	0.70	0.39	0.00	0.00	0.19	-1.31
matarial_dum	0.01	0.06	0.91	0.12	-0.02	0.07	0.83	-0.21	0.00	0.00	0.62	-0.49
constant	8.14	0.79	p<0.0001	10.30	8.84	0.89	p<0.0001	9.90	8.52	0.02	p<0.0001	505.52
Model Statistics												
Wald χ^2	3.85				6.29				14.58			
ρ	1.00				-1.00				-1.00			
σ	0.65				0.74				0.01			
Log Likelihood	-250.78				-250.78				-250.78			
McFadden R^2	0.08				0.08				0.08			
$N_{\text{nonselected}}$	113				113				113			
N_{selected}	409				409				409			

Table 13: Regression outcomes of Storytelling_Base on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.03	0.03	0.32	-0.99	-0.03	0.05	0.53	-0.63	-0.07	0.03	p<0.05	-2.22
edu_m	-0.01	0.03	0.66	-0.44	-0.02	0.05	0.72	-0.36	-0.01	0.03	0.85	-0.19
mus_act_dum	0.00	0.06	0.97	-0.03	0.02	0.12	0.87	-0.17	-0.06	0.07	0.43	-0.79
mus_att_ave	0.11	0.09	0.20	1.27	-0.11	0.16	0.52	-0.64	-0.04	0.10	0.66	-0.44
mus_ca_edu_a19	0.04	0.03	0.08	1.74	0.04	0.05	0.45	0.75	0.01	0.03	0.66	0.44
mus_ca_edu_b18	0.00	0.02	0.95	0.06	-0.03	0.05	0.49	-0.69	-0.02	0.03	0.48	-0.71
mus_qz_tot_sco	0.05	0.03	0.07	1.81	0.12	0.05	p<0.05	2.32	0.04	0.03	0.22	1.22
moral_n_ave	0.02	0.02	0.30	1.05	0.13	0.04	p<0.0001	3.01	0.05	0.02	0.05	2.00
social_n_ave	0.02	0.02	0.44	0.78	-0.01	0.04	0.74	-0.33	0.00	0.02	0.91	0.11
inc_mo	0.01	0.01	0.26	1.13	0.02	0.02	0.23	1.21	0.01	0.01	0.34	0.96
matarial_dum	0.01	0.03	0.77	0.30	-0.06	0.06	0.32	-1.00	-0.03	0.03	0.34	-0.95
constant	9.07	0.37	p<0.0001	24.47	9.77	0.70	p<0.0001	14.05	9.91	0.40	p<0.0001	-24.67
Model Statistics												
Wald χ^2	31.27				31.22				26.14			
ρ	0.41				-0.43				-0.87			
σ	0.27				0.50				0.32			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 14: Regression outcomes of Storytelling_Base on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.01	0.06	0.84	-0.21	0.03	0.06	0.67	0.42	0.06	0.10	0.56	0.59
edu_m	0.02	0.06	0.71	0.37	0.00	0.07	0.98	-0.03	-0.04	0.12	0.70	-0.38
mus_act_dum	-0.07	0.13	0.58	-0.55	0.15	0.15	0.31	1.01	0.20	0.24	0.41	0.82
mus_att_ave	0.02	0.12	0.86	0.18	0.14	0.14	0.29	1.05	0.21	0.22	0.34	0.95
mus_ca_edu_a19	0.00	0.04	0.96	0.05	-0.01	0.05	0.82	-0.23	0.03	0.08	0.71	0.37
mus_ca_edu_b18	-0.03	0.06	0.57	-0.57	0.05	0.07	0.45	0.75	0.09	0.10	0.37	0.90
mus_qz_tot_sco	-0.01	0.06	0.79	-0.27	0.01	0.06	0.88	0.15	0.01	0.10	0.90	0.13
moral_n_ave	0.03	0.03	0.44	0.77	0.02	0.04	0.58	0.55	0.00	0.06	0.97	-0.03
social_n_ave	0.00	0.03	0.97	-0.03	-0.01	0.03	0.71	-0.37	0.00	0.05	0.96	-0.05
inc_mo	0.00	0.01	0.78	0.28	0.00	0.02	0.99	0.02	0.00	0.02	0.84	0.20
matarial_dum	-0.02	0.05	0.75	-0.32	0.02	0.06	0.78	0.28	0.00	0.09	0.99	-0.02
constant	8.71	0.64	p<0.0001	13.64	8.07	0.72	p<0.0001	11.21	7.71	1.15	p<0.0001	6.72
Model Statistics												
Wald χ^2	3.89				4.15				2.02			
ρ	-1.00				1.00				1.00			
σ	0.53				0.59				0.94			
Log Likelihood	-250.78				-250.78				-250.78			
McFadden R^2	0.08				0.08				0.08			
$N_{\text{nonselected}}$	113				113				113			
N_{selected}	409				409				409			

Table 15: Regression outcomes of Storytelling_CoT on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.05	0.04	0.21	-1.26	-0.06	0.06	0.30	-1.03	0.31	0.16	0.06	1.91
edu_m	0.03	0.04	0.46	0.74	-0.03	0.05	0.61	-0.50	-0.46	0.16	p<0.0001	-2.86
mus_act_dum	-0.01	0.08	0.90	-0.12	-0.03	0.12	0.78	-0.27	0.15	0.36	0.67	-0.42
mus_att_ave	0.04	0.11	0.71	0.37	-0.07	0.17	0.68	-0.42	0.69	0.50	0.17	-1.37
mus_ca_edu_a19	0.04	0.03	0.27	1.09	0.01	0.05	0.77	0.29	0.22	0.15	0.14	-1.47
mus_ca_edu_b18	0.00	0.03	0.98	-0.03	-0.02	0.05	0.61	-0.51	0.06	0.14	0.68	-0.41
mus_qz_tot_sco	0.07	0.04	0.05	1.93	0.12	0.06	p<0.05	2.05	-0.22	0.17	0.18	-1.34
moral_n_ave	0.09	0.03	p<0.0001	2.97	0.14	0.04	p<0.0001	3.14	0.11	0.13	0.40	0.83
social_n_ave	-0.02	0.03	0.46	-0.73	0.01	0.04	0.83	0.21	0.16	0.12	0.18	1.33
inc_mo	0.01	0.01	0.21	1.25	0.03	0.02	0.07	1.78	-0.02	0.05	0.75	-0.32
matarial_dum	-0.03	0.04	0.50	-0.67	0.06	0.06	0.32	-1.00	-0.15	0.18	0.40	-0.84
constant	9.19	0.48	p<0.0001	-19.15	9.50	0.73	p<0.0001	13.00	6.44	2.12	p<0.0001	3.03
Model Statistics												
Wald χ^2	27.60				34.69				17.90			
ρ	-0.36				-0.51				0.67			
σ	0.34				0.53				1.60			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 16: Regression outcomes of Storytelling_CoT on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	0.01	0.04	0.85	0.85	0.00	0.01	0.48	0.71	-	-	-	-
edu_m	-0.01	0.04	0.85	0.85	-0.01	0.01	0.29	-1.06	-	-	-	-
mus_act_dum	0.05	0.09	0.59	0.53	0.01	0.01	0.57	0.05	-	-	-	-
mus_att_ave	0.12	0.08	0.15	1.44	0.01	0.01	0.36	0.92	-	-	-	-
mus_ca_edu_a19	-0.03	0.03	0.29	-1.06	0.00	0.00	0.89	0.14	-	-	-	-
mus_ca_edu_b18	-0.01	0.04	0.89	-0.14	0.00	0.01	0.66	0.45	-	-	-	-
mus_qz_tot_sco	0.01	0.04	0.72	0.36	0.01	0.01	0.39	0.87	-	-	-	-
moral_n_ave	0.04	0.02	0.11	1.61	0.00	0.00	0.30	1.05	-	-	-	-
social_n_ave	-0.03	0.02	0.09	-1.69	0.00	0.00	0.67	-0.43	-	-	-	-
inc_mo	0.01	0.01	0.38	0.88	0.00	0.00	0.86	0.18	-	-	-	-
matarial_dum	-0.01	0.03	0.72	-0.36	0.00	0.01	0.81	0.24	-	-	-	-
constant	8.31	0.42	p<0.0001	19.63	8.44	0.07	p<0.0001	118.37	-	-	-	-
Model Statistics												
Wald χ^2	21.11				5.10				-			
ρ	0.72				0.99				-			
σ	0.31				0.06				-			
Log Likelihood	-250.78				-250.78				-			
McFadden R^2	0.08				0.08				-			
$N_{\text{nonselected}}$	113				113				-			
N_{selected}	409				409				-			

Table 17: Regression outcomes of Storytelling_RAG on the art PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions. In Qwen, as all values were measured identitically, statistical analysis becomes impossible.

	GPT-4o				LLaMA				Qwen			
	coeff	std_error	p value	z value	coeff	std_error	p value	z value	coeff	std_error	p value	z value
Outcome												
edu_f	-0.04	0.03	0.15	-1.44	-0.06	0.03	p<0.05	-2.36	-0.03	0.04	0.43	-0.79
edu_m	-0.01	0.03	0.83	-0.21	0.02	0.02	0.52	0.65	-0.04	0.04	0.26	-1.12
mus_act_dum	-0.02	0.06	0.76	-0.31	-0.05	0.06	0.41	-0.83	-0.02	0.09	0.78	-0.27
mus_att_ave	0.02	0.09	0.81	0.23	-0.03	0.08	0.67	-0.42	0.04	0.12	0.72	0.36
mus_ca_edu_a19	0.03	0.03	0.21	1.25	0.01	0.02	0.71	0.37	0.05	0.04	0.18	1.34
mus_ca_edu_b18	-0.02	0.02	0.52	-0.64	-0.01	0.02	0.72	-0.36	-0.03	0.03	0.31	-1.02
mus_qz_tot_sco	0.05	0.03	0.10	1.64	0.05	0.03	0.06	1.90	0.04	0.04	0.35	0.94
moral_n_ave	0.03	0.02	0.13	1.44	0.04	0.02	p<0.05	2.39	0.03	0.03	0.31	1.02
social_n_ave	0.02	0.01	0.22	0.77	0.01	0.02	0.65	0.45	0.02	0.03	0.41	0.98
inc_mo	0.01	0.01	0.42	0.80	0.00	0.01	0.56	0.59	0.00	0.01	0.99	0.34
matarial_dum	0.00	0.03	0.92	0.10	0.00	0.03	0.87	0.16	0.00	0.04	p<0.05	0.01
constant	9.46	0.37	p<0.0001	25.32	9.89	0.33	p<0.0001	30.24	9.57	0.52	p<0.0001	18.48
Model Statistics												
Wald χ^2	23.42				26.57				22.24			
ρ	-0.39				-0.82				-0.29			
σ	0.27				0.26				0.37			
Log Likelihood	-290.01				-290.01				-290.01			
McFadden R^2	0.16				0.16				0.16			
$N_{\text{nonselected}}$	195				195				195			
N_{selected}	327				327				327			

Table 18: Regression outcomes of Storytelling_RAG on the music PWYW decision-making task. The table reports coefficients, standard errors, z-values, and p-values from a Heckman two-step model. Bottom panel presents model-level statistics including selection bias correction and sample distributions.

Few-shot

[User] 설문 소개문 / 기관 설명 / 가상 상황 설명

[User]
Persona 1: 성별: 여성 / 연령: 35 / 클래식 선호: 그렇다 $\cdots$ **Q1.** 특별전시를 9,000원을 지불하고 관람할 의도가 있으십니까?
Reply: ② 그렇다
Persona 2: 성별: 남성 / 연령: 32 $\cdots$ **Reply:** ① 아니다
Persona 3: $\cdots$ **Reply:** ② 그렇다
****Your Persona**** *{target persona text}*
Q1. 특별전시를 9,000원을 지불하고 관람할 의도가 있으십니까? **Your Answer:**

[Assistant] *{Model Response $\hat{y}_1$}*

▷ After response: few-shot block removed; assistant turn replaced with ground-truth y_1.

▷ Q1: top-3 by persona similarity. Q2+: filtered by matching previous answer.

Figure 8: Few-shot prompt template. Three retrieved personas with ground-truth answers are prepended; after the model responds the block is removed and overwritten with the ground-truth answer.

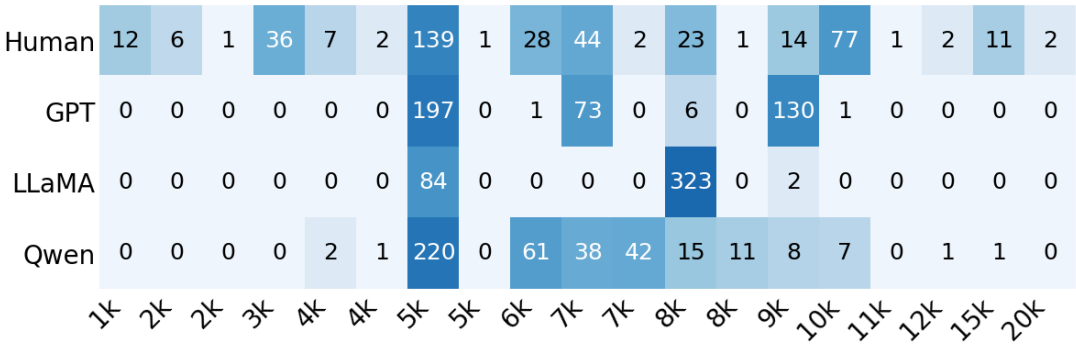

Figure 9: Response distribution in the Survey format × CoT condition.

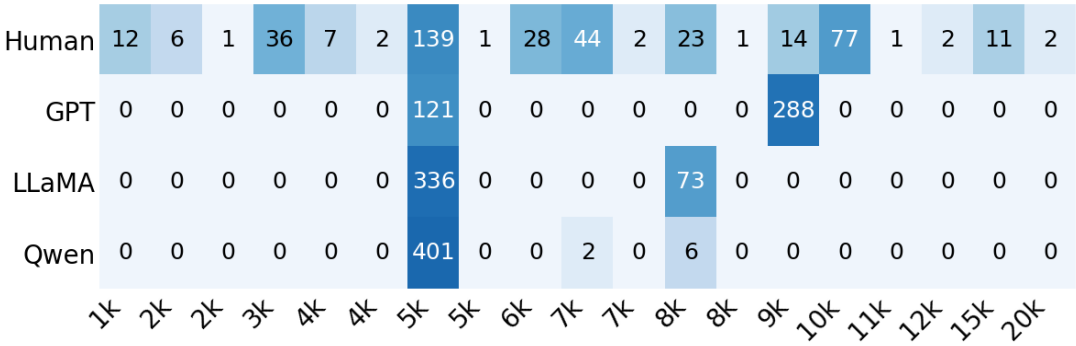

Figure 10: Response distribution in the Survey format × RAG condition.

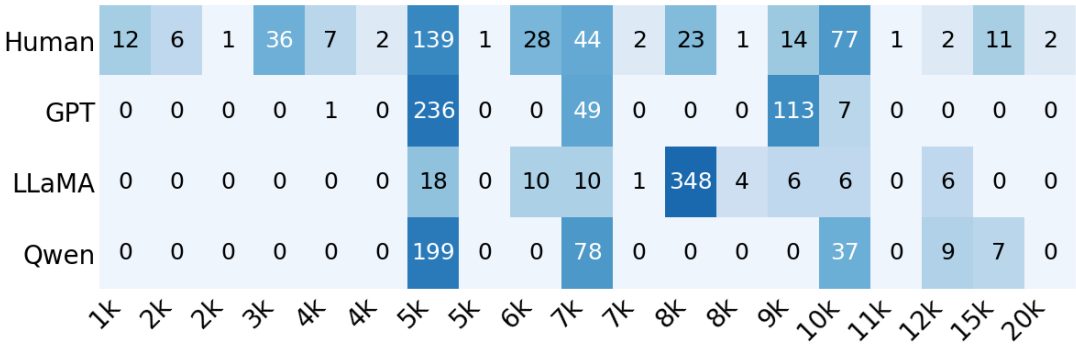

Figure 11: Response distribution in the Survey format × Few-shot condition.

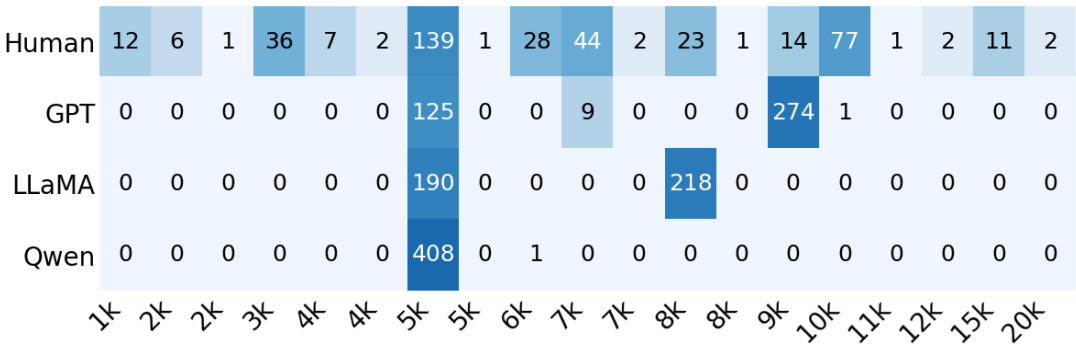

Figure 12: Response distribution in the Storytelling format × Base condition.

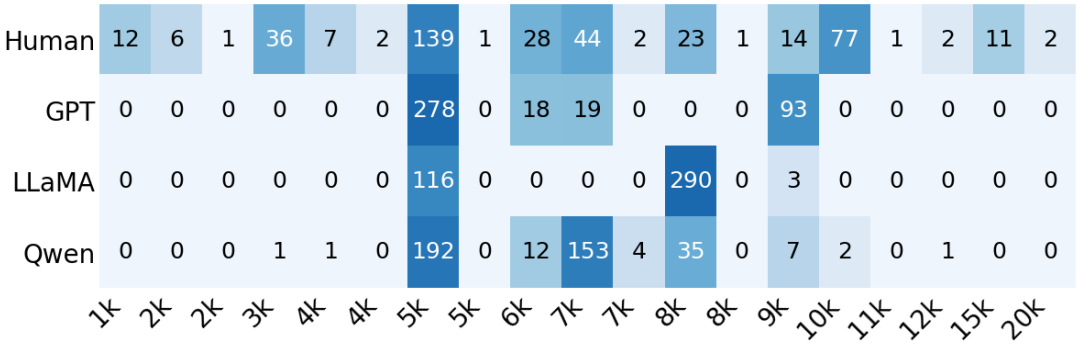

Figure 13: Response distribution in the Storytelling format × CoT condition.

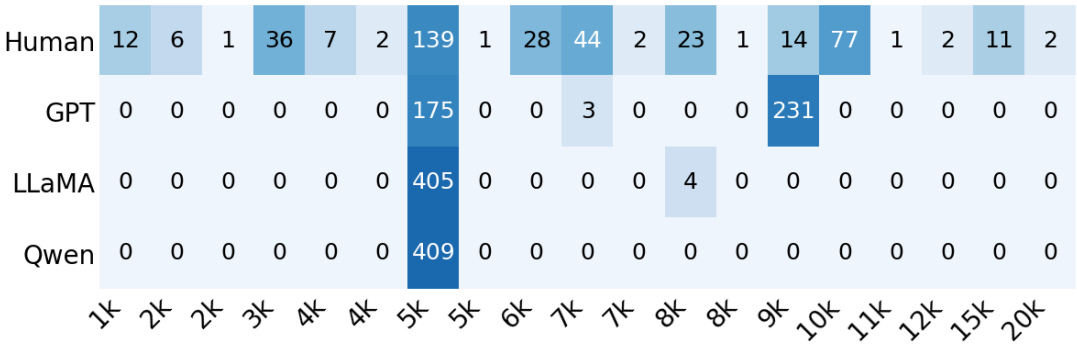

Figure 14: Response distribution in the Storytelling format × RAG condition.

국·공립 박물관 및 공연장 관람료에 대한 소비자 설문조사
Consumer Survey on Admission Fees for National/Public Museums and Performance Halls

[대상 안내 / Participant Criteria]
본 설문은 **30대 (만 30세에서 만 39세)**를 대상으로 진행됩니다. **국립중앙박물관 특별전「시대의 얼굴, 셰익스피어에서 에드 시런까지 2021」을 관람하지 않은 분, 손열음 피아노 독주회를 관람하지 않은 분**만 응답 가능합니다. 위 전시 또는 공연을 관람하신 분은 설문을 종료해 주세요.
This survey targets individuals in their 30s (ages 30–39). It is intended only for those who have **not visited the special exhibition "The Face of the Era"** at the National Museum of Korea and have **not attended the Sunwook Kim piano recital**. If you have seen either, please discontinue this survey.

[조사 목적 / Survey Purpose]
본 설문은 국·공립 박물관 및 공연장 관람료 지불의사에 대한 연구 목적을 가지고 있으며, 응답 내용은 설문 목적 외의 용도로 절대 사용되지 않습니다. 귀하의 응답은 문화예술산업의 발전을 위한 소중한 자료로 활용될 예정입니다. 바쁘시더라도 성실한 응답을 부탁드립니다.
This survey investigates willingness to pay for admissions to national/public museums and performance halls. All answers will be used solely for research purposes and will never be used beyond the scope of the survey. Your responses are valuable for advancing the cultural and arts industry. We kindly ask for your honest participation.

[응답자 보호 / Confidentiality]
수집된 정보는 *통계법* 제33조(비밀의 보호), 제34조(조사자의 의무)에 따라 보호되며, 귀하의 정보는 익명으로 처리됩니다. 어떠한 경우에도 신상 정보가 노출되지 않습니다.
In accordance with Articles 33 and 34 of the Statistics Act, all collected information will be protected and kept strictly confidential. Your answers will remain anonymous, and no identifying information will be disclosed under any circumstances.

※ 각 문항에는 옳고 그른 답이 없으며, 솔직하게 평소 생각을 바탕으로 응답해 주시기 바랍니다.
※ There are no right or wrong answers. Please respond honestly based on your usual thoughts.

다시 한 번, 설문에 참여해 주셔서 감사합니다.
Once again, thank you very much for your participation.

Figure 15: Survey Introduction part of the PWYW questionnaire.

CASE 1

I. 다음은 국립중앙박물관에 대한 설명입니다.
I. The following is an introduction to the National Museum of Korea.

국립중앙박물관은 **우리나라를 대표하는 국립박물관**으로 1915년 12월 1일 경복궁에서 개관한 조선총독부박물관을 인수하여 1945년 12월 3일 *국립박물관*이라는 명칭으로 개관하였습니다. 이후 1972년 7월 *국립중앙박물관*으로 명칭이 변경되었고, 현재 위치인 용산으로 2005년 이전하였습니다. 국립중앙박물관은 과거로부터 이어온 세계 유산을 수집 · 보존 · 전시하고 있으며, 특히 우리 문화의 보급을 통해 세계와 소통하려 노력하고 있습니다.

The National Museum of Korea, the **representative national museum of South Korea**, was originally opened as the Government-General Museum of Korea in Gyeongbokgung Palace on December 1, 1915. It was reestablished under the name "National Museum" on December 3, 1945. The name was changed to "National Museum of Korea" in July 1972, and the museum was relocated to its current site in Yongsan in 2005. The museum is dedicated to collecting, preserving, and exhibiting global heritage from the past, and especially strives to communicate Korean culture to the world.

국립중앙박물관은 **무료**로 운영되는 상설전시관, 어린이박물관 등의 시설과 **유료**로 운영되는 특별기획전시관을 보유하고 있습니다. 또한, **행사, 공연, 교육 등 문화예술 체험 프로그램**을 통해 성인뿐 아니라 어린이, 외국인 등 다양한 관람객이 참여할 수 있도록 운영하고 있습니다.

The museum operates permanent exhibition halls and a children's museum that are **free of charge**, along with **special exhibitions that require paid admission**. It also offers **cultural and artistic programs**, such as events, performances, and educational experiences, designed to encourage participation from a wide audience, including adults, children, and foreign visitors.

[그림 1] 국립중앙박물관 전경 [Figure 1] View of the National Museum of Korea

[그림 2] 국립중앙박물관 외관 [Figure 2] Exterior of the National Museum of Korea

Figure 16: Case 1 (Art) Introduction part of the PWYW questionnaire.

II. 다음은 전시 방문 가상상황에 대한 설명입니다.
II. The following is a hypothetical scenario about visiting an exhibition.

여러분은 현재 서울 용산구 국립중앙박물관 근처에 있고 3–4시간 정도 여유시간이 있습니다. 바로 옆 국립중앙박물관에서는 성인기준 **9,000원**으로 《*시대의 얼굴, 셰익스피어에서 에드 시런까지 2021*》가 특별전시하고 있습니다.

You are currently near the National Museum of Korea in Yongsan, Seoul, and have about 3–4 hours of free time. Next door, the National Museum is holding a special exhibition titled *"The Face of the Era: From Shakespeare to Ed Sheeran 2021"* with an admission fee of **₩9,000** for adults.

[그림] 전시 포스터 / [Figure] Exhibition Poster

《**전시소개**》**Exhibition Description**
이번 전시는 '76명의 역사적 인물'이 그려진 **'초상화'**라는 예술 장르를 중심으로 한 기획전입니다. 영국 국립초상화박물관의 소장품이 한국에서 처음 공개되며, 세계 역사와 문화를 빛낸 인물들을 만날 수 있습니다. 셰익스피어, 엘리자베스 1세, 뉴턴, 비틀즈, 에드 시런 등이 등장합니다.

This exhibition focuses on the genre of **portraiture**, showcasing depictions of 76 historical figures. For the first time, the National Portrait Gallery (UK) brings its collection to Korea. Visitors will encounter world-renowned figures such as Shakespeare, Queen Elizabeth I, Newton, The Beatles, and Ed Sheeran.

1. 귀하께서는 **국립중앙박물관의 특별전시 '시대의 얼굴'을 9,000원을 지불하고 관람**할 의도가 있으십니까?
   Would you be willing to pay ₩9,000 to see the special exhibition "The Face of the Era" at the National Museum of Korea? ① 아니오 (No) ② 그렇다 (Yes)

2. 만약 **위 전시에 대해 방문객이 원하는 가격을 스스로 정해서 지불**할 수 있다면, 귀하께서는 이 경우 방문할 의향이 있으십니까?
   If you could choose your own admission fee for this exhibition, would you still want to visit it? ① 아니다 (No) → Q3로 이동 (Go to Q3)
   ② 그렇다 (Yes) → Q2-1로 이동 (Go to Q2-1)

Q2-1. **방문하실 경우 얼마 정도 관람료로 낼 의향이** 있으십니까?
If you choose to visit, how much would you be willing to pay as admission? _____ 원 (KRW)

Q3. 귀하께서 **위 전시에 대해 관람 의사가 없다고 답변하신 이유는 무엇입니까?** (복수선택 가능)
If you are not willing to attend the exhibition, what are the reasons? (Check all that apply)

- ① 해당 기관에서 제공하는 전시 및 프로그램에 관심이 없다
  I am not interested in the exhibitions or programs offered at the institution.
- ② 위 전시에 대한 프로그램이 만족스럽지 않다
  The content of the exhibition is not satisfactory.
- ③ 위 전시의 관람 여부를 판단하기 위해 필요한 정보가 충분히 주어지지 않았다
  There is insufficient information to judge the value of the exhibition.
- ④ 나는 위의 전시와 관련된 주제에 관심이 없다
  I am not interested in the theme of the exhibition.
- ⑤ 나는 상대적으로 다른 여가활동을 더 선호한다
  I prefer other leisure activities over this exhibition.
- ⑥ 기타 (직접 서술): _____
  Other: _____

Figure 17: Case 1 (Art) Question part of the PWYW questionnaire.

CASE 2

III. 다음은 예술의전당에 대한 설명입니다.
III. The following is an introduction to the Seoul Arts Center.

예술의전당은 **우리나라를 대표하는 공공기관 공연장**으로 문화예술의 창달과 진흥, 문화예술향유기회 확대를 목표로 클래식 음악, 오페라 등의 공연을 비롯하여 미술작품의 수집 · 전시를 목적으로 1988년 설립되었습니다.

The Seoul Arts Center is a **leading public cultural venue in Korea** established in 1988 with the goal of promoting and advancing the arts. It aims to expand public access to cultural experiences, including classical music, opera performances, and exhibitions of visual artworks.

예술의전당은 2016년 누적 관람객이 5천만 명을 넘었고 음악당, 오페라하우스, 서울서예박물관, 한가람미술관, 한가람디자인미술관 등의 시설을 갖추고 있습니다. 예술의전당은 관람객에게 **공연, 전시, 교육 등의 문화예술경험**을 제공하고 있으며 **서울시 서초구**에 위치한 국가대표 문화예술기관입니다.

As of 2016, the center had welcomed over 50 million visitors. It houses multiple facilities such as a concert hall, opera house, calligraphy museum, Hangaram Art Museum, and the Design Museum. Located in **Seocho-gu, Seoul**, it offers diverse **cultural and artistic experiences**, including performances, exhibitions, and educational programs.

예술의전당 음악당에서는 ① 예술가나 기획자가 직접 대관하여 진행하는 **대관공연**과 ② 예술의전당에서 **특별 기획한 프로그램들**을 소개하고 있으며, ③ 한화생명, KT, 신세계 등과 같은 기업의 **후원으로 매달 한 번 오전 11시에 클래식 음악 콘서트**를 제공하고 있습니다.

At the Seoul Arts Center's Music Hall, ① **private performances** organized by artists or planners and ② **special programs curated by the center** are presented. ③ With sponsorship from companies such as Hanwha Life, KT, and Shinsegae, the center also offers a **monthly classical music concert at 11 a.m.**

[그림 1] 예술의전당 전경
[Figure 1] Full View of Seoul Arts Center

[그림 2] 예술의전당 음악당 외관
[Figure 2] Exterior of SAC Music Hall

Figure 18: Case 2 (Music) Introduction part of the PWYW questionnaire.

IV. 다음은 가상상황에 대한 설명입니다.
IV. The following is a hypothetical scenario.
여러분은 현재 서울 서초구 예술의전당 근처에 있고 3–4시간 정도 여유시간이 있습니다. 바로 옆 예술의전당에서는 성인기준 **B석 30,000원**으로 <*손열음 피아노 독주회*>가 곧 시작할 예정입니다.
You are currently near the Seoul Arts Center in Seocho-gu, Seoul, and have about 3–4 hours of free time. At the Arts Center nearby, the *"Son Yeol-eum Piano Recital"* will begin soon, with **B-seat tickets priced at 30,000 KRW**.

[그림] 공연 포스터 / [Figure] Performance Poster

<연주자 및 공연소개> About the Artist and Performance
피아니스트 손열음은 2005년 루빈스타인 국제 피아노 콩쿠르에서 3위를 수상하고 2009년 벨클라이번 국제 피아노 콩쿠르에서도 준우승한 대한민국 대표 피아니스트입니다.
In 2005, pianist Son Yeol-eum placed 3rd in the Rubinstein International Piano Competition, and in 2009 won the silver medal at the Van Cliburn International Piano Competition. She is one of Korea's most celebrated classical pianists.

이번 독주회는 그녀가 가장 아끼는 작곡가 슈만과 모차르트를 중심으로 구성된 **"All Schumann Program"**으로, 손열음의 음악적 해석이 기대되는 무대입니다.
The recital focuses on her favorite composers — Schumann and Mozart — under the **"All Schumann Program"**. Audiences can look forward to her deep musical interpretation.

1. 귀하께서는 **손열음 피아노 독주회를 B석 관람료 가격 30,000원을 지불하고 관람**할 의도가 있으십니까?
   Would you be willing to pay 30,000 KRW (B-seat) to attend the recital?
   ① 아니요 (No) ② 그렇다 (Yes)

2. 만일 **위 공연에 대해 방문객이 원하는 가격을 스스로 정해서** 지불할 수 있다면, 관람 의향이 있으십니까?
   If you could choose your own admission fee, would you still be willing to attend the recital?
   ① 아니다 (No) → Q3로 이동 (Go to Q3)
   ② 그렇다 (Yes) → Q2-1로 이동 (Go to Q2-1)

Q2-1. **방문하실 경우 얼마 정도 관람료로 낼 의향이** 있으십니까?
If you choose to attend, how much would you be willing to pay for admission?
1회 입장 시 _____ 원 (KRW)

Q3. 귀하께서 **예술의전당 공연 관람 의사가 없다고 답변하신 이유는 무엇입니까?**
If you are not willing to attend the performance, what are the reasons? (Check all that apply)

- ① 해당 기관에서 제공하는 공연 및 프로그램에 관심이 없다
  I am not interested in the institution's programs or performances.
- ② 위 공연에 대한 프로그램이 만족스럽지 않다
  The program content is not satisfactory.
- ③ 위 공연의 관람 여부를 판단하기 위한 정보가 충분히 주어지지 않았다
  There is insufficient information to make a decision.
- ④ 나는 위의 공연과 관련된 주제에 관심이 없다
  I am not interested in the theme of this performance.
- ⑤ 나는 상대적으로 다른 여가활동을 더 선호한다
  I prefer other leisure activities.
- ⑥ 기타 (직접 서술): _____
  Other: _____

Figure 19: Case 2 (Music) Question part of the PWYW questionnaire.