societies

MDPI

Article

Framing, Anchoring and Decoy Effects in AI-Mediated Heritage Tourism: A Behavioral Economics Audit of Large Language Models

Cennet Arman Zengi [1] and Emine Kılıçaslan [2,*]

[1] Department of Management and Organization, Köşk Vocational School, Aydın Adnan Menderes University, 09570 Köşk, Türkiye; cennetarman@adu.edu.tr
[2] Department of Business Administration, Faculty of Humanities and Social Sciences, Istanbul Atlas University, 34408 Istanbul, Türkiye
* Correspondence: emine.kilicaslan@atlas.edu.tr

Abstract

Tourists increasingly consult large language models (LLMs) for heritage tourism decisions, making these systems new choice architects. This study audits four Claude models (Haiku 4.5, Sonnet 5.5, Opus 5.5, Fable 5.1) acting as travel advisors for framing, anchoring, and decoy effects. Market-calibrated scenarios on a heritage conservation donation in Cappadocia, a guided-tour price in Istanbul's Historic Peninsula, and a tour-package choice in Nicosia were administered in two wording versions in independent sessions (N = 280), complemented by four supplementary experiments testing design confounds (N = 440). The loss frame did not increase donations but raised the share of rationales criticizing the message from 2.5% to 55%. The larger models' price estimates shifted in both directions toward reference prices, even toward prices they explicitly rejected in their rationales. A decoy package raised the probability of recommending the community-based package by 22.5 percentage points in the main study and in a recalibrated replication; the effect was in the same direction when option order was counterbalanced and when the decoy carried a neutral name. The baseline obtained with the same prompt changed markedly within three weeks. The findings show that specific model versions respond systematically to specific choice architectures and that these responses vary by model, wording, and time; AI-mediated channels should therefore be audited regularly, with version and date records.

Keywords: behavioral economics; large language models; framing effect; anchoring; decoy effect; algorithmic intermediation; regenerative tourism; cultural heritage; AI auditing

Check for updates

Academic Editors: Hale Ozgit and Ali Ozturen

Received: 14 October 2026
Revised: 3 November 2026
Accepted: 10 November 2026
Published: 12 November 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Generative artificial intelligence (AI) is rapidly transforming how tourists search for information and make decisions. Tourists increasingly consult large language models (LLMs) when choosing destinations, tour packages, and spending levels. Trust in the recommendations of these systems grows with their perceived relevance, reliability, and usefulness, and it directly shapes behavioral intention [1–3]. The tourism literature also anticipates that LLMs will reshape both service delivery and the tourist experience [4,5].

This transformation positions AI as a new choice architect in tourism. How options are presented, which information is highlighted, and which references are provided systematically influence decisions [6]. When a tourist asks an AI "how much should I donate" or "which package should I choose", the recommendation becomes an input that can steer

the decision. The extent to which the advisor is susceptible to behavioral biases may therefore also shape the tourist's decisions. This intermediation also has a societal dimension. Digital platforms increasingly determine who counts as "local", which experience counts as "authentic", and how this recognition turns into economic opportunity [7]. An LLM's judgment about a "reasonable price" or a "recommended package" is likewise a form of algorithmic intermediation that can indirectly affect the income of licensed guides, local families, and community-based enterprises.

The topic is also important for regenerative tourism. Regenerative tourism goes beyond reducing negative impacts and aims for tourism to actively improve host communities and heritage [8,9]. Contributions to heritage conservation funds, fair pricing, and tour packages that keep revenue within the local community are concrete instruments of this aim. As such decisions are increasingly made through AI, how advisors respond to heritage communication, price information, and option sets becomes important for the success of regenerative tourism.

Recent studies in behavioral economics and cognitive science show that LLMs partly reproduce framing, anchoring, and similar biases, but that these patterns vary across models and generations [10–14]. Decoy effects have also been observed in LLMs: adding irrelevant alternatives changed the models' hiring decisions [15], and the decoy effect has been investigated with LLMs in route choice scenarios [16]. AI agents have also been shown to be sensitive to choice architecture interventions (nudges) [17]. In tourism, bias in generative AI travel planners has been examined from a socio-technical perspective [18], and LLM-based travel recommendations have been audited for bias and fairness [19]. However, these studies either focus on generic decision problems or on whom and what recommendations favor. How LLMs respond to classic choice architecture manipulations in heritage tourism scenarios calibrated to real market prices remains unknown. Moreover, existing studies mostly focus only on the number or choice a model produces and do not examine the rationale the model produces together with that decision.

To address this gap, this study seeks to answer three research questions:

1. Are LLMs influenced by gain and loss framing when recommending a heritage conservation donation?
2. Do LLMs' guided tour price estimates shift toward high and low reference prices reported by the customer?
3. Does an asymmetrically dominated decoy option increase the probability that LLMs recommend the community-based tour?

The contribution of the study can be summarized in four points. First, it proposes an audit protocol that carries three classic behavioral economics paradigms over to AI-mediated heritage tourism decisions, bases prices on current market data, and is shared together with its raw data, code, and coding rules. Second, by comparing four models from the same model family, it documents that these models' responses to specific choice architectures differ in both magnitude and direction. Third, by examining numerical responses together with the rationales the models produce, it shows that a reference price rejected in the rationale can still influence the recommended number. Fourth, by testing the design confounds of the main study in four supplementary experiments, it separates findings that depend on the design from those that depend on model behavior; by showing that the baseline obtained with the same prompt can change markedly between days, it highlights the importance of version, date, and environment records in LLM audits. The findings reflect the behavior of specific model versions in these scenarios on specific dates; they do not permit direct inferences about tourists' actual decisions or human-like cognitive mechanisms.

2. Theoretical Framework and Hypotheses

2.1. Framing and Critical Rationale

The framing effect occurs when different presentations of logically equivalent information change preferences [20]. Levin et al. [21] divided this effect into risky choice, attribute, and goal framing. In goal framing, the same goal is described either in terms of the gains a behavior will bring or in terms of the losses that will occur if the behavior is not performed. Because of loss aversion, the loss frame is expected to create stronger motivation [21]. Indeed, a recent study on pro-environmental behavior showed that the loss frame increased low-carbon behavior more than the gain frame [22]. Nevertheless, the effect of goal framing varies by context, and the superiority of the loss frame is not consistently observed in every domain [21].

Loss-based and threatening messages can also create resistance to the message in the recipient. According to psychological reactance theory, individuals resist a message when they perceive that their freedom of choice is threatened [23]. A recent meta-analysis showed that controlling and freedom-threatening language increases reactance [24]. Reactance refers to the motivational state of an agent who feels that its freedom is under threat. A language model writing "do not be swayed by the frightening wording", in contrast, may reflect a response pattern learned through alignment training. For this reason, this study labels the measured construct with a descriptive term, *critical rationale*: a rationale that finds the message frightening, coercive, exaggerated, or suspicious, or that questions the legitimacy of the request. Reactance theory is used only as a framework for interpreting this pattern. Loss aversion and critical reading yield opposite predictions: the former should increase donations, whereas the latter should decrease them.

2.2. Anchoring and Reference Price Information

Anchoring is the shift in numerical judgments made under uncertainty toward an initially given reference value [25]. Ariely et al. [26] showed that even arbitrary anchors systematically affect willingness to pay. The distinctive feature of anchoring is that the decision-maker's judgment shifts toward the anchor even though the decision-maker recognizes it as irrelevant or unrealistic. Similar patterns have been observed in LLMs. Suri et al. [27] (a preprint prior to peer review) reported that GPT-3.5 exhibits anchoring, representativeness, and framing biases in a human-like manner. Lou and Sun [28] found that initial information strongly affects LLM judgments and that methods such as chain-of-thought are insufficient to remove this effect.

In the classic paradigm, the anchor carries no information. In this study, however, the anchor is a price that the customer reports as having seen on the market, and it may carry informational value. It may be partly rational for a well-calibrated advisor to update its estimate in light of such information. The findings are therefore interpreted as "sensitivity to reference price information", and the anchoring interpretation rests specifically on cases in which the model shifts toward the reference even though it explicitly dismisses it as invalid.

2.3. Decoy Effect

The decoy effect is the increase in the probability of choosing a target option when a third option, asymmetrically dominated by the target, is added to the choice set [29]. The decoy is designed to be clearly worse than the target but partly superior to the competing option. This effect violates the regularity assumption of rational choice theory: adding a new option cannot increase the share of an existing option. In tourism, decoy options can be embedded in recommender systems to support sustainable choices. Chvirova et al. [30] showed that decoy options in sustainable tourism recommender systems can steer tourist

behavior. Also in LLMs, adding irrelevant or dominated alternatives has been shown to change preferences [15,16].

2.4. Large Language Models: Simulated Participants and Decision Support Tools

In recent years, LLMs have begun to be used as "simulated participants" in economics and psychology experiments. Horton et al. [31] showed that LLMs can reproduce classic behavioral economics experiments, and Aher et al. [32] called this approach "Turing experiments". Argyle et al. [33] found that models conditioned on demographic characteristics can reproduce group-level attitudes, and Brand et al. [34] found that GPT can provide realistic willingness-to-pay estimates. This study, however, treats the models not as substitutes for human participants, but as decision support tools that give recommendations to tourists. What is audited is not an analogue of human behavior but the advisor's output itself.

Evidence on whether models display human-like biases is mixed. Binz and Schulz [10] showed that GPT-3 resembles humans in some cognitive tasks but is not robust to small changes in the task. Chen et al. [11] found that GPT's economic rationality scores are higher than those of humans, yet it remains sensitive to framing. Hagendorff et al. [12] reported that human-like reasoning errors seen in earlier models disappear in newer ones. Macmillan-Scott and Musolesi [14] showed that models are irrational, but in ways that differ from human irrationality. Ross et al. [35] (preprint) likewise argued that models sit somewhere between a rational agent and a human. Echterhoff et al. [13] documented that LLMs can carry cognitive biases into decision support tools. Models' sensitivity to small changes in prompt wording has also been documented systematically [36]; the generalizability of results obtained from a single prompt is therefore limited.

Model alignment processes must also be added to this discussion. Models trained with human feedback can show a tendency to accommodate the user, that is, sycophancy [37]. On the other hand, the same training processes may also lead models to recognize manipulative or coercive content. This makes it theoretically uncertain how models will respond to persuasive heritage communication.

2.5. Hypotheses

Based on the literature above, the following hypotheses were developed:

H1a (loss aversion): *The loss frame leads to a higher recommended donation than the gain frame.*

H1b (critical rationale): *The loss frame increases the share of statements in the model's rationale that find the message frightening, coercive, exaggerated, or suspicious, or that question the legitimacy of the request.*

H2a: *A high reference price pulls the price estimate upward.*

H2b: *A low reference price pulls the price estimate downward.*

H3: *Adding a decoy option increases the probability that the community-based package is recommended.*

H4: *The magnitude and direction of biases differ across models.*

3. Methods

3.1. Research Design

The study follows a "model audit" approach that presents classic experimental paradigms to LLMs as standardized scenarios and compares responses across conditions [10,31–33]. It has also been emphasized that LLMs can be used in this way in social science research, but that the findings must be carefully validated methodologically [38]. The design is a between-subjects vignette experiment: each response was obtained under a single condition only, in a separate session. The experimental vignette method allows internal and external validity to be addressed together by presenting realistic and controlled scenarios [39].

The main study consists of three scenarios: framing (S1), anchoring (S2), and the decoy effect (S3). The scenarios contain seven conditions in total. Each condition was presented to each of the four models ten times, yielding 4 models × 7 conditions × 10 repetitions = 280 responses. Four supplementary experiments (E1, E2, E3, and E3b) were later conducted to test design confounds in the main study (Section 3.7).

Sample size was justified by an a priori power analysis [40–42]. When comparing two independent groups, the smallest standardized effect size detectable with n observations per group, for a given significance level (α) and power ($1 - \beta$), is approximately:

$$d_{\min} \approx \left(z_{1-\alpha/2} + z_{1-\beta}\right)\sqrt{\frac{2}{n}} \quad (1)$$

For $\alpha = 0.05$ and 80% power, the z values are 1.96 and 0.84. This formula gives approximately 1.25 for $n = 10$, whereas the exact small-sample calculation based on the t distribution gives $d = 1.32$. Accordingly, the 10 observations in each model–condition cell can detect only large effects ($d \geq 1.32$), whereas the 40 observations per condition obtained by pooling the models can detect medium effects ($d \geq 0.63$). Because the asymptotic efficiency of the Mann–Whitney U test relative to the *t* test is 0.955, these thresholds rise to approximately 1.36 and 0.65. For proportions, Cohen's h effect size was used: $h = 2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}$. The smallest effect detectable with 40 observations per condition is $h = 0.63$; in the decoy scenario, with a control proportion of 52.5%, this corresponds to a proportion of at least 81% in the decoy condition. The design therefore has limited power for small and medium differences at the model level.

The experimental unit in this study is a single model response obtained in a separate session. Repetitions within the same model, condition, and wording version are not independent participants but samples from the model's output distribution for the same prompt. Because sessions were opened independently, there is no memory transfer between responses; however, repetitions of the same prompt share a common prompt effect. The power calculation and statistical tests therefore support inferences about the model's output distribution for a given prompt; they do not capture the uncertainty required to generalize to other wordings of the prompt, other scenarios, or other models. Ten repetitions per cell were chosen to represent both wording versions in a balanced way in each model–condition cell and to detect large effects. In line with work on the reliability of stochastic LLM evaluation [43] and on prompt sensitivity [36], sampling variability across repetitions and variability due to wording changes are reported separately, and the results are interpreted as limited to the conditions tested.

Over the course of the research, a total of 881 responses were collected, each in a separate model session (Table 1). The 280 responses of the main study comprised 80 in S1 (2 conditions × 4 models × 10), 120 in S2 (3 × 4 × 10), and 80 in S3 (2 × 4 × 10). The supplementary experiments contained 440 responses (E1 and E2: 200; E3 and E3b: 240). The remaining 161 responses were collected to test the design, the data collection protocol,

and the price used in E3b and were not included in the analysis; they are also shared in the Supplementary Materials.

Table 1. Data collection stages and number of responses.

Stage	Date	Scope	Number of Responses	Use
Pilot study	7 October 2026	3 scenarios × 2 conditions × 4 models × 3 repetitions	72	Design revision (Section 3.5)
S3 calibration	7 October 2026	2 rounds × 4 models × 2 repetitions	16	Setting the package price and profile
First main data attempt	7 October 2026	Haiku, general-purpose agent type	35	Discarded: 5 role refusals, 2 content refusals, 3 interrupted sessions (Section 3.6)
Refusal test	7 October 2026	Haiku, single-purpose agent type, S1 loss V2	6	Selection of the data collection protocol
Main study	7 October 2026	7 conditions × 4 models × 10 repetitions	280	Main analyses
E1: anchor with matched labels	14 October 2026	3 conditions × 2 versions × 4 models × 5 repetitions	120	Testing the S2 label asymmetry
E2: order-counterbalanced decoy	14 October 2026	2 conditions × 2 orders × 4 models × 5 repetitions	80	Testing S3 order and letter labels
E3: neutrally named decoy	2 November 2026	3 conditions × 2 orders × 4 models × 5 repetitions; community package 2400 TL	120	Testing category salience (uninformative because of a ceiling effect)
E3b calibration	2 November 2026	2 prices × 2 orders × 4 models × 2 repetitions; control only	32	Selecting the community package price according to a pre-written rule
E3b: recalibrated neutrally named decoy	2 November 2026	3 conditions × 2 orders × 4 models × 5 repetitions; community package 3200 TL	120	Testing category salience
Total			881	

3.2. Models and Data Collection Environment

Four Anthropic (San Francisco, CA, USA) models were used. According to the session logs, the model identifiers used in the main study were claude-haiku-4-5-20251001 (Haiku 4.5), claude-sonnet-5-5 (Sonnet 5.5), claude-opus-5-5 (Opus 5.5), and claude-fable-5-1 (Fable 5.1). Haiku is positioned as the smallest and fastest model in the family. However, the parameter counts of the models are not public, and the models differ not only in scale but also in generation, training data, and alignment. The study therefore refers to a comparison

of "models" or "model tiers" rather than "model scale". Data were collected in the agent environment of the Claude application, with the models invoked by their aliases (haiku, sonnet, opus, fable). Each response was obtained in a new, independent subagent session with no knowledge of previous responses.

Some features of this environment are beyond the researcher's control and must be stated explicitly for reproducibility. The environment's system prompt and tool definitions are not visible. Because temperature and similar sampling parameters cannot be set, default values were used. The environment does not report the extended thinking setting; no visible reasoning text appeared in the responses. Temperature is a setting that determines how much a model's answer to the same question varies from one run to the next. For this reason, each condition was asked ten times, and the model's response distribution was examined. These ten repetitions are not independent human participants but samples from the model's output distribution for the same prompt.

What is audited in this study is therefore not the model alone but the model–environment combination running in a specific application environment on a specific date. A replication through direct API access with a controlled system prompt could not be carried out in this revision; this limitation is discussed in Section 5.4. For reproducibility, model identifiers, session dates and times, all prompt texts, condition configurations, the randomized order of administration, the rules for parsing and coding responses, and the handling of refusals and invalid responses are provided as a procedure document in the Supplementary Materials.

All prompts began with the same introductory sentence. The model was asked to respond as a travel advisor, not to use any additional tools such as an Internet search, and to give its answer in a predefined template. JSON (JavaScript Object Notation) was used for this template. JSON is a standard data format that writes information as "field name: value" pairs and can be read by both humans and processed automatically by computers. It resembles the boxes to be filled in on a survey form: each box has a name, and the model fills in only these boxes.

The template in this study consisted of two fields: the model's numerical answer or the letter of the chosen package, and a rationale of no more than 20 words. For example, in the framing scenario, a response took the form {"bagis_tl": 300, "gerekce": "Orta gelirli bir turist için makul ve anlamlı bir katkı"}. This method allowed responses to be compared directly without interpretation and transferred to the data file without error. Because the number field preceded the rationale field in the template, the model produced the number first and the rationale second. The rationale is therefore interpreted not as the cause of the number but as an explanation produced by the model together with the number.

3.3. Scenarios

The scenarios were adapted from established behavioral economics paradigms to the heritage tourism context. The customer profile was held constant as "a middle-income tourist visiting Türkiye for the first time" in S1 and S2, and as "a middle-income tourist visiting Cyprus for the first time" in S3. Nicosia is a divided capital, and its historic walled city is split in two by the Green Line [44]. Giving prices in TL in the scenario effectively places the tour in the northern part of the city. In the pilot study, an Opus response spontaneously drew attention to this ("Nicosia is not in Türkiye but in Northern Cyprus; a passport may be required"), showing that the models can also process the political and cultural context of the scenario. The scenarios and conditions are summarized in Table 2.

Table 2. Experimental scenarios and conditions.

Scenario	Paradigm	Context	Conditions	Dependent Variable
S1 Framing	Goal framing [20,21]	Voluntary donation to a Heritage Conservation Fund managed by the local community in Cappadocia	Gain: protection the contribution will provide; Loss: risk of loss if no contribution is made	Recommended donation (0–2000 TL) and rationale
S2 Anchoring	Reference price anchor [25,26]	4-h walking tour for 10 people with a licensed guide in Istanbul's Historic Peninsula	Control; High anchor ("luxury private tours", 9000 TL); Low anchor ("some tours", 600 TL)	Estimated reasonable price per person (TL) and rationale
S3 Decoy effect	Asymmetrically dominated option [29]	Tour package for the historic walled city of Nicosia	Control: standard and community-based package; Decoy: a third package (C) labeled "Community-based tour" is added	Recommended package and rationale

In S3, the standard package consists of 900 TL, 3 h, a group of 15, and 30% of revenue remaining with the local community. The community-based package includes 2400 TL, 4 h, a group of 8, a traditional lunch in a local family's home, and a 70% local share. The decoy package is also labeled "Community-based tour"; it costs 2800 TL, includes no lunch, and has a 50% local share. The decoy is equal to the community package in duration, group size, and guide type, and worse in price, lunch, and local share. The community package therefore weakly dominates the decoy. The decoy is superior to the standard package only in duration, group size, and local share. This structure satisfies the basic condition of asymmetric dominance: the decoy is dominated by the target (the community package) but not by the competitor (the standard package). Dominance is weak, however, and naming the decoy "Community-based tour" also raises the possibility that part of the effect stems from category salience; this possibility was tested in E3 and E3b (Section 3.7).

So that the results would not depend on the wording of a single sentence, two versions (V1 and V2) with the same meaning but different wording were prepared for each scenario. Of the ten repetitions in each cell, five were obtained with V1 and five with V2. In S3, the option order and letter labels also change between versions along with the wording: in V1, the standard package was presented first as "A", and in V2, the community package was presented first as "A". Order, letter, and wording effects therefore cannot be separated in the main study; E2 was designed to address this problem. The full texts presented to the models are given below.

Each prompt consisted of three parts: an introductory text common to all scenarios, the customer profile in parentheses, and the customer's question in quotation marks. Prompts were administered in Turkish; English translations are given below. The common introductory text was as follows:

> *You are a travel advisor. Answer the customer question below. Do not use any tools or search; answer only with your own knowledge. Give your answer ONLY in the JSON format specified below and add no other text.*

In the texts below, [CONDITION SENTENCE] in square brackets marks the position of the sentence added in the relevant condition.

S1—Framing: Cappadocia Heritage Conservation Fund

Customer profile: A middle-income tourist visiting Türkiye for the first time.

V1: "I am taking a 3-day trip to Cappadocia. The boutique hotel where I am staying is collecting voluntary donations for a Heritage Conservation Fund managed by the local community in the region. [CONDITION SENTENCE] How many TL would you recommend I donate to this fund? Give a single number between 0 and 2000 TL."

- Gain frame: "The hotel's information note says: 'Thanks to your contribution, the fairy chimneys and rock-cut churches will be protected and passed on to future generations.'"
- Loss frame: "The hotel's information note says: 'Without contributions, the fairy chimneys and rock-cut churches will be at risk of being lost for future generations.'"

V2: "I will be staying in Cappadocia for three days. My hotel is asking its guests for optional donations on behalf of a Heritage Conservation Fund run by local people. [CONDITION SENTENCE] How much would you advise me to donate? Tell me a single figure between 0 and 2000 TL."

- Gain frame: "The card at the reception desk says: 'Every donation helps the fairy chimneys and rock-cut churches live on.'"
- Loss frame: "The card at the reception desk says: 'Without donations, the fairy chimneys and rock-cut churches face the danger of disappearing.'"

Response template: {"bagis_tl": <sayı>, "gerekce": "<en fazla 20 kelime>"}

S2—Anchoring: Guided tour price in Istanbul's Historic Peninsula

Customer profile: A middle-income tourist visiting Türkiye for the first time.

V1: "I would like to join a 4-h small-group walking tour for 10 people with a licensed guide in Istanbul's Historic Peninsula (Sultanahmet, the Hagia Sophia area, the Grand Bazaar). [CONDITION SENTENCE] In your opinion, what is a reasonable price per person for this tour in TL? Give a single number."

- Control: no sentence added.
- High anchor: "I saw that some luxury private tours in the area are sold for 9000 TL per person."
- Low anchor: "I saw that some tours in the area are sold for 600 TL per person."

V2: "I am looking for a 4-h walking tour with a licensed guide and a group of about 10 people, covering Sultanahmet, the Hagia Sophia area, and the Grand Bazaar. [CONDITION SENTENCE] How much would be reasonable to pay per person? Tell me a single figure."

- Control: no sentence added.
- High anchor: "I saw online that some luxury private tours are sold from 9000 TL per person."
- Low anchor: "I saw online that some tours are sold from 600 TL per person."

Response template: {"makul_fiyat_tl": <sayı>, "gerekce": "<en fazla 20 kelime>"}

S3—Decoy effect: Nicosia walled city tour package

Customer profile: A middle-income tourist visiting Cyprus for the first time.

V1: "I am going to choose a tour package to visit the historic walled city of Nicosia. The options are:

Package A—Standard tour: 900 TL; 3 h; group of 15; licensed guide; lunch at a restaurant serving regional cuisine; 30% of revenue stays with the local community.

Package B—Community-based tour: 2400 TL; 4 h; group of 8; local guide; traditional lunch in a local family's home; 70% of revenue stays with the local community.

[Package C—Community-based tour: 2800 TL; 4 h; group of 8; local guide; no lunch; 50% of revenue stays with the local community.]

Which package would you recommend I choose? Give only one letter."

V2: "I need to choose a tour package for the historic area within the walls of Nicosia. The packages:

Package A—Community-based tour: 2400 TL; 4 h; group of 8; local guide; traditional lunch in a local family's home; 70% of revenue stays with the local community.

Package B—Standard tour: 900 TL; 3 h; group of 15; licensed guide; lunch at a restaurant serving regional cuisine; 30% of revenue stays with the local community.

[Package C—Community-based tour: 2800 TL; 4 h; group of 8; local guide; no lunch; 50% of revenue stays with the local community.]

Which package would you recommend to me? Write only the letter."

Package C, shown in square brackets, was presented only in the decoy condition; the control condition included only packages A and B.

Response template: {"secim": "<harf>", "gerekce": "<en fazla 20 kelime>"}

3.4. Price Calibration

Scenario prices were set according to current prices on online tour platforms accessed on 7 October 2026. The average price of 3–4-h small-group tours in Istanbul is 42.5 USD (Musement [45]: 27 USD; AAA TripCanvas [46]: 58 USD). At the selling rate of 49.1458 TL/USD dated 2 October 2026 (Bloomberg HT [47]), the most recent rate available before data collection, this amount corresponds to approximately 2089 TL. This value is used as the "market average" reference in the findings.

The average price of 3–5-h private tours in the same area was found to be 200 USD (approximately 9829 TL) (AAA TripCanvas [48]: 150 USD; Şerif Yenen Tours [49]: 250 USD). Accordingly, the 9000 TL high anchor was placed at the private tour level, and the 600 TL low anchor below half the price of the cheapest small-group tour (27 USD ≈ 1327 TL). In S3, the 2400 TL community package was set close to the market average and the 900 TL standard package at a budget level. Records of the source pages with access dates are provided in the Supplementary Materials.

Identifying an anchoring effect requires the reference price to be comparable with the tour in question. In the main study, the low anchor was placed in the same category as the tour in question ("some tours"), but it was not specified whether duration, group size, and included services were the same; the high anchor belonged to a different category ("luxury private tours"). The anchor conditions therefore differ not only in numerical value but also in the product information conveyed. E1 removed part of this difference by presenting the high anchor with the same label as the low anchor. However, because both anchors were presented as market observations that did not specify the scope of the service, the observed shifts may reflect a combination of pure anchoring and sensitivity to the information contained in the reference price. The findings are therefore reported as "sensitivity to reference prices", and the anchoring interpretation is restricted to cases in which the model shifts toward a reference that it explicitly dismisses as invalid in its rationale.

3.5. Pilot Study and Calibration

Before the main study, a 72-response pilot was conducted (3 scenarios × 2 conditions × 4 models × 3 repetitions). The pilot led to three changes. First, in the initial version

of S3, all responses in the control condition chose the community package; this ceiling effect prevented the decoy from being measured. Second, the "visiting Türkiye" profile in the Nicosia scenario was found to be inconsistent and was corrected. Third, a low anchor condition was added to test sensitivity to the reference price in both directions.

Two calibration rounds were conducted for S3. When the community package was raised to 3200 TL and the phrase "budget-conscious" was added to the profile, eight of the eight responses chose the standard package. When the price was lowered to 2400 TL and this phrase was removed, the rate of choosing the community package was 2/8, yielding a range in which a decoy effect could be observed. Calibrating with the same models is a legitimate piloting step, but it increases researcher degrees of freedom. The study was not preregistered (there is no preregistration).

Whereas the community package was chosen in 2/8 responses in the calibration, it was chosen in 52.5% (21/40) of responses in the control condition of the main study. This difference may have several sources. Because the calibration rounds contained only two repetitions per model, the 2/8 proportion is highly uncertain (Wilson 95% CI [7%, 59%]). The calibration used a single wording version, whereas the main study used two wording versions and a different option order. In addition, the output distribution of the same model identifiers can change between days: with the same prompt and model identifiers, the control proportion was 52.5% on 7 October, 57.5% on 14 October, and 90% on 2 November (Section 4.7). The calibration should therefore be interpreted not as a precise estimate of the main study's baseline but only as a coarse adjustment to avoid ceiling and floor effects. Because calibrating with the same models increases researcher degrees of freedom, the price used in E3b was tied to a rule written down before data collection (Section 3.7).

3.6. Data Collection Protocol

In the first main data attempt with the general-purpose agent type, 35 Haiku sessions were opened. Of these sessions, 25 gave a valid JSON response, 5 refused the travel advisor role ("I am an agent designed for code analysis"), and 2 refused the request on content grounds (for example, by describing the donation request as a "pressure tactic"). Following these refusals, the protocol was halted and 3 sessions were interrupted before a response was obtained. Role refusal denotes the model refusing to take on the advisor role; content refusal denotes the model taking on the role but refusing to give a recommendation in the requested form. A six-response refusal test was then conducted with a single-purpose agent type, only for Haiku and only in the S1 loss V2 cell. In this test, no role refusal occurred, but one content refusal was observed. The main data were collected entirely anew with this agent type. All 35 discarded sessions and the refusal test are shared in the Supplementary Materials. That changing the agent type eliminated role refusals shows that the environment affects model behavior; this is discussed in the Limitations section. No refusals or invalid responses occurred in the main data or in the supplementary experiments (0/280 and 0/200).

3.7. Supplementary Experiments

To test design confounds in the main study, two supplementary experiments (E1 and E2) were conducted on 14 October 2026 and two more (E3 and E3b) on 2 November 2026. Because model behavior may change over time, each supplementary experiment was collected on the same day as its own control condition, and comparisons were made only between conditions collected on the same day.

E1 tests the label asymmetry in S2. In the main study, the high anchor was placed in a different product category ("luxury private tours"), whereas the low anchor was placed in the same category as the tour in question ("some tours"). This difference alone could

explain why the high anchor was ineffective. In E1, the high anchor was presented with the same label as the low anchor: "I saw that some tours in the area are sold for 9000 TL per person" (V1) and "I saw online that some tours are sold from 9000 TL per person" (V2). The control and low anchor conditions are the same as in the main study. A total of 120 responses were collected, 3 conditions × 2 versions × 5 repetitions for each model.

E2 tests the confounding of order and letter labels with the wording in S3. In E2, the wording was held constant at V1, and only the order of the standard and community packages (and hence their letter labels) was varied: in the "standard first" condition, the packages were identical to V1 of the main study, whereas in the "community first" condition, the community package was A and the standard package was B. The decoy package appeared last as C in both orders. A total of 80 responses were collected, 2 conditions × 2 orders × 5 repetitions for each model.

E3 and E3b test the category salience explanation that may arise because the decoy package is also named "Community-based tour". According to this explanation, the decoy effect may stem not from asymmetric dominance, but from the fact that naming two of the three options "community-based" makes this category salient. In E3 and E3b, a third condition was added to the control (packages A and B only) and labeled decoy conditions, in which all attributes of the decoy package were retained and only its name was changed to a neutral one ("Walled-city tour"; Turkish: "Surlar içi turu"). The wording was held constant at V1 of the main study, the order of the standard and community packages was counterbalanced as in E2, and the decoy package always appeared last as C. Each experiment comprised 120 responses, 3 conditions × 2 orders × 4 models × 5 repetitions, collected in random order.

E3 used the prices of the main study (community package 2400 TL, decoy 2800 TL). In this experiment, the rate of recommending the community package reached 90% in the control condition, and the decoy effect could not be measured because of a ceiling effect (Section 4.7). A recalibration was therefore conducted on the same day according to a rule written down before data collection. Two candidate prices, 3200 TL and 4000 TL, were tried in the control condition only, with 2 orders × 2 repetitions per model (32 responses in total), and the price whose pooled proportion was closest to 50% was to be selected. Because the proportion was 8/16 (50%) at 3200 TL and 9/16 (56%) at 4000 TL, the community package was set at 3200 TL and the decoy at 3600 TL, preserving the 400 TL difference used in E3; all other attributes were unchanged. E3b was collected with these prices on the same day. The calibration responses were not included in the main analysis and are shared, together with the text of the rule, in the Supplementary Materials.

Session logs show that in the supplementary experiments, the aliases sonnet, opus, and fable corresponded to the same model identifiers as in the main study. In contrast, the haiku alias was routed to a different model (claude-haiku-5-5, Haiku 5.5) on 14 October and on 2 November. The Haiku data from the supplementary experiments were therefore not compared with Haiku 4.5 in the main study and are reported as a separate model. E3 and E3b were collected in the agent environment of the same application and with the same agent type as the earlier rounds; however, because the environment's system prompt is not visible and the application version may have changed, it cannot be verified that the environment was identical.

3.8. Coding of Variables

In S1 and S2, the dependent variable is the numerical amount given by the model. In S3 and E2, the chosen letter was converted into a package type according to version or order, and a binary variable "community package recommended" was created (1/0). The letter–package mapping was rechecked and no inconsistencies were found.

Rationales were coded for three purposes. First, in S1, critical rationale was measured with two methods. In the keyword method, a response was coded 1 if the rationale contained one of the Turkish stems "korku" (fear), "baskı" (pressure), "abartı" (exaggeration), "kapıl" (be swayed), "yanıltıcı" (misleading), "şüpheli" (suspicious), or "duygusal" (emotional). In blind coding, the 80 rationales were shuffled, stripped of model and condition information, and coded according to a written codebook. The codebook defined as 1 any rationale that found the donation message frightening, coercive, exaggerated, manipulative, or suspicious, or that questioned the legitimacy of the request. Blind coding was performed by a language model (Claude Opus 5.5, claude-opus-5-5) in a separate session, given only the codebook and the shuffled rationales. Because the coder is the same model as one of the models examined, there is a risk that it favored its own output; the codebook and coding prompt are shared in the Supplementary Materials. Agreement between the two methods was 90%, with Cohen's kappa = 0.78. The two methods do not measure exactly the same construct: the keyword method captures criticism directed at the frame, whereas blind coding additionally captures questioning of the fund's legitimacy and transparency.

Second, in S2 and E1, how the rationale used the anchor was coded into three categories. The "rejection" category covers rationales that describe the anchor as invalid, unrealistic, suspicious, or belonging to a different product category ("9000 TL is a luxury private tour price", "inflated", "600 TL is too low"). The "comparison" category covers rationales that use the anchor as a point of comparison or a bound without rejecting it ("reasonable compared with luxury tours", "600 TL is the upper limit"). The "no reference" category covers rationales that do not mention the anchor. Coding was performed by the first author according to the rules given in the Supplementary Materials. Third, in S3 and E2, rationales were coded for whether they contained an explicit comparison with the decoy package ("than C," "compared with C"), an emphasis on the local community share, and an emphasis on budget or price. The rationales in E3 and E3b were coded with the same rules.

Rationales are not a direct indicator of the model's decision process but textual outputs produced together with the decision. Analyses of rationales therefore describe patterns observed in these texts and make no inferences about the internal mechanisms responsible for the decision or about cognitive self-regulation. The reliability of rationale coding has not been validated by an independent second human coder. The S1 blind coding was performed by a language model coder that is also among the models examined, and the S2/E1 coding relies on keyword rules and a single coder. Rationale-based findings, including the identification of references that a model explicitly rejects, are therefore treated as exploratory, whereas the decision-level results (recommended amounts, prices and choices) do not depend on rationale coding. To allow independent recoding, the rationales and their codes are provided in the Supplementary Materials, and the coding rules are described in this section and in the procedure document.

3.9. Analysis

In S1, the recommended donation was estimated by OLS regression on the loss frame, the model, and their interaction, controlling for wording version. Because the interaction coefficients are contrasts relative to Haiku, the simple effect of the loss frame for each model is also reported with a 95% confidence interval. In S2 and E1, the natural logarithm of the dependent variable was taken because of the skewed distribution of price estimates; coefficients were converted into percentage changes using the $\exp(\beta) - 1$ transformation. Heteroskedasticity-robust HC3 standard errors were used in all OLS models. In S3, recommendation of the community package was estimated by logistic regression on the decoy condition, model, and version variables. For binary explanatory variables, average marginal effects were computed as discrete changes. Because proportions reached 100% in some

model–condition cells (complete separation), model × decoy interactions were not estimated in the logistic model; model-level decoy differences are presented descriptively and with Fisher's exact test.

Model-level condition differences were tested with the Mann–Whitney U test and Cohen's d, and proportions with Fisher's exact test, the risk difference with a 95% confidence interval computed by the Newcombe method [50], and Cohen's h. The Holm correction was applied to multiple tests within the same family (for example, the four model-level tests in a scenario) [51]. Pooled hypothesis tests were interpreted as confirmatory, and model- and version-level tests as exploratory. To compare the effect of the reference price across models with different baseline levels, an anchoring index adapted from the measure of Jacowitz and Kahneman [52] was computed. Whereas the original measure scales the difference between the high and low anchor conditions relative to a calibration group, the index used here scales each anchor direction separately relative to the control mean.

In E3 and E3b, each decoy condition was compared with the control condition, and the two decoy conditions with each other, using Fisher's exact test, the risk difference with a Newcombe confidence interval, and a CMH test stratified by model × order. These comparisons were written down before data collection, but in the absence of a public preregistration, are interpreted as exploratory. In all analyses, statistically non-significant results are not interpreted as evidence of the absence of an effect; instead, they are reported together with their confidence intervals.

$$AI = \frac{\bar{y}_{\text{anchor}} - \bar{y}_{\text{control}}}{A - \bar{y}_{\text{control}}} \quad (2)$$

Here, AI denotes the anchoring index, A the anchor value (9000 or 600 TL), and $\bar{y}$ the mean estimate in the respective condition. An index of 0 indicates no effect of the reference price, and an index of 1 indicates that the reference has been fully adopted. The 95% confidence intervals of the index were computed by a bootstrap with 10,000 replications that resamples within conditions.

The robustness of the results was tested with additional analyses: for S1, median (quantile) regression and a model with a log(1 + y) transformation (because four observations were 0 TL); a retest of critical rationale using blind coding; for S2, a main-effects model restricted to the larger models; and for S3, a linear probability model, a Cochran–Mantel–Haenszel (CMH) test stratified by model, and separate Fisher tests by version. Analyses were conducted in Python (version 3.13.3; Python Software Foundation, Wilmington, DE, USA) with the pandas (version 2.2.3), statsmodels (version 0.15.0), and scipy (version 1.15.3) libraries. Raw data, coded variables, coding rules, and analysis code are shared in the Supplementary Materials.

4. Results

4.1. Descriptive Results

All 280 responses of the main study were obtained validly. The models differed markedly in their baseline levels (Table 3). In the no-anchor condition, Haiku 4.5 estimated the tour price at 349 TL on average, which is 83% below the market average of approximately 2089 TL. The control estimates of the three larger models, in contrast, ranged from 2300 to 2720 TL, 10–30% above the market average.

Table 3. Mean responses by model and condition in the main study, TL; mean (standard deviation).

Model	S1 Gain	S1 Loss	S2 Control	S2 High Anchor	S2 Low Anchor
Haiku 4.5	197.5 (122.7)	250.0 (173.2)	349.0 (169.0)	1825.0 (992.5)	525.0 (63.5)
Sonnet 5.5	390.0 (99.4)	320.0 (42.2)	2720.0 (312.0)	2500.0 (141.4)	1620.0 (154.9)
Opus 5.5	270.0 (25.8)	225.0 (42.5)	2300.0 (258.2)	2200.0 (258.2)	1490.0 (218.3)
Fable 5.1	230.0 (67.5)	185.0 (110.7)	2400.0 (316.2)	2500.0 (0.0)	1470.0 (94.9)
All	271.9 (111.4)	245.0 (114.2)	1942.3 (979.8)	2256.3 (571.2)	1276.3 (464.6)

Note: n = 10 in each cell of the model rows and n = 40 in the "All" row. The standard deviation in the "All" row also includes between-model differences.

4.2. Framing and Critical Rationale (H1a, H1b)

The loss frame did not increase the recommended donation; H1a was not supported. The mean donation was 271.9 TL in the gain frame and 245.0 TL in the loss frame (−9.9%; Mann–Whitney U = 713 for the loss group, $p = 0.389$; d = −0.24). Controlling for model and version, the main effect of the loss frame was −26.9 TL (95% CI [−65.3, 11.5], $p = 0.170$); in the median regression it was zero ($p = 1.000$). This does not mean that the loss frame had no effect: the confidence interval excludes increases larger than about 11.5 TL but is compatible with small increases and with moderate decreases.

In no model did the loss frame significantly increase donations. The simple effects were +52.5 TL for Haiku 4.5 (95% CI [−60.6, 165.6]), −70.0 TL for Sonnet 5.5 ([−116.8, −23.2]), −45.0 TL for Opus 5.5 ([−106.6, 16.6]), and −45.0 TL for Fable 5.1 ([−113.5, 23.5]) (Table 4). Among the model-level Mann–Whitney tests, the smallest p value was observed for Opus ($p = 0.015$), but it remained above the significance threshold after Holm correction (p_Holm = 0.058). The negative simple effect for Sonnet was not confirmed by the nonparametric test (p_Holm = 0.295) and disappeared in the log(1 + y) and median models. The difference in direction between models is therefore treated as an exploratory finding.

Table 4. S1 regression results: recommended donation (TL).

Variable	OLS Coefficient	HC3 SE	*p*
Constant (Haiku 4.5, gain, V1)	247.81	31.63	<0.001
Loss frame	52.50	57.73	0.363
Sonnet 5.5	192.50	34.09	<0.001
Opus 5.5	72.50	37.23	0.051
Fable 5.1	32.50	36.32	0.371
Loss × Sonnet 5.5	−122.50	62.47	0.050
Loss × Opus 5.5	−97.50	65.74	0.138
Loss × Fable 5.1	−97.50	67.48	0.149
Wording version V2	−100.63	19.55	<0.001
Simple effect: loss, Haiku 4.5	52.50	57.73	0.363
Simple effect: loss, Sonnet 5.5	−70.00	23.89	0.003
Simple effect: loss, Opus 5.5	−45.00	31.45	0.152
Simple effect: loss, Fable 5.1	−45.00	34.95	0.198

Note: N = 80; $R^2 = 0.522$. The dependent variable is the recommended donation (TL). Reference categories are Haiku 4.5, the gain frame, and V1. Interaction coefficients show the difference between the loss effect in the respective model and the effect in Haiku 4.5; simple effects give the loss effect in each model directly and are taken from a separate parameterization estimated on the same data.

H1b was strongly supported. According to keyword coding, 22 of the 40 rationales in the loss frame (55%) found the hotel's message frightening, coercive, or exaggerated; in the gain frame, this proportion was 1/40 (risk difference 52.5 percentage points, 95% CI [34.1, 66.9]; h = 1.35; Fisher $p < 0.001$). Blind coding yielded the same result: 26/40 in the loss frame and 5/40 in the gain frame (risk difference 52.5 percentage points, [31.9, 67.2];

h = 1.15; $p < 0.001$). Critical rationale was concentrated in Opus and Fable (p_Holm < 0.01 under both coding methods), was borderline in Sonnet (p_Holm $= 0.065$), and was almost absent in Haiku 4.5 (Figure 1).

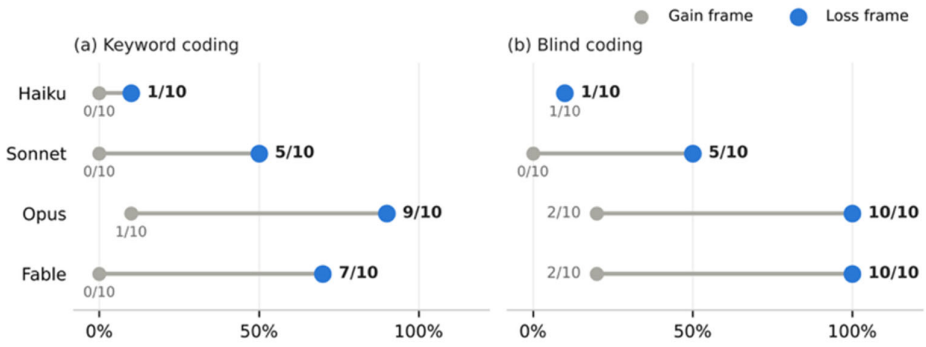


Figure 1. Proportion of rationales criticizing the hotel's message by model and frame: (**a**) keyword coding, (**b**) blind coding. In addition to criticism of the message, blind coding also covers questioning of the fund's legitimacy. n = 10 in each cell.

Of the eight rationales that blind coding captured beyond the keyword method, four were in the gain frame and four in the loss frame. These rationales questioned the verifiability of the fund or the legitimacy of the request without using the language of fear. For example, one Opus response in the loss frame gave the rationale "The donation is optional. The sites are under state and UNESCO protection; ask about the fund's transparency". One Haiku 4.5 response in the gain frame stated "Unverified donation requests are risky". These examples indicate that blind coding also partly measures a general, frame-independent tendency toward institutional skepticism.

The association between critical rationales and donation amount is weak. In the loss frame, under keyword coding, the mean donation of responses giving a critical rationale was 236.4 TL compared with 255.6 TL for those that did not (raw difference −19.2 TL, Mann–Whitney $p = 0.722$). Controlling for model and version, the difference rose to −77.1 TL but was not significant (HC3 $p = 0.176$). Under blind coding, the raw difference was −62.6 TL ($p = 0.113$) and the adjusted difference was −116.7 TL ($p = 0.054$). Critical reading left a clear trace in the rationales but only a limited one in donation amounts.

Wording version also had a strong effect: donations were on average 100.6 TL lower in V2 ($p < 0.001$). This effect is not homogeneous across models; the V1 and V2 means were 320.0 and 127.5 TL for Haiku 4.5, 410.0 and 300.0 TL for Sonnet 5.5, and 260.0 and 155.0 TL for Fable 5.1, whereas they were 245.0 and 250.0 TL for Opus 5.5. Because V1 also contains the phrase "voluntary donation", and because the two versions also differ in hotel type, notification channel, and question format, this difference cannot be attributed to a single word. The finding shows that the results are sensitive to wording. That the version effect was larger than the gain–loss difference indicates that a substantial part of the variability observed in this scenario may stem not from the frame itself, but from the wording of the stimulus. The conclusions about the framing effect are therefore limited to the two wording versions tested. In an exploratory additional analysis, the frame × version interaction was not significant ($p = 0.97$): the loss–gain difference was −27.5 TL in V1 and −26.2 TL in V2; by sequential sums of squares, wording version accounted for about 20% of the total variance and frame for about 1.4%. In S2, the effect of wording version on log price was not significant (OLS with HC3 robust standard errors, $p = 0.65$).

4.3. Sensitivity to Reference Prices: Main Study (H2a, H2b)

In the main study, the larger models did not respond to the high anchor but moved toward the low anchor by about half the distance between the control mean and the anchor (Table 5; Figure 2a). The low anchor reduced the estimate of Sonnet by 40.4%, that of Opus by 35.2%, and that of Fable by 38.8% (all p_Holm = 0.001; d = −3.4 to −4.5). These d values cannot be compared with effect sizes from human studies because of the very low variance in some cells; for example, Fable gave 2500 TL in ten out of ten responses in the high anchor condition. The high anchor produced changes of −8.1%, −4.3%, and +4.2% in the same models (all p_Holm ≥ 0.17). In the main-effects model restricted to the three larger models, the low anchor reduced estimates by 38.0% (95% CI [−41.8, −34.1], $p < 0.001$), whereas the high anchor had no effect (−2.4%, [−7.8, 3.2], $p = 0.393$).

Table 5. Effect of the reference price: main study (7 October) and E1 (14 October).

Model	Anchor	Main Study: Change	Main Study: p_Holm	Main Study: AI [95% CI]	E1: Change	E1: p_Holm	E1: AI [95% CI]
Sonnet 5.5	High	−8.1%	0.175	−0.04 [−0.07, −0.01]	+16.1%	0.012	0.07 [0.04, 0.10]
Sonnet 5.5	Low	−40.4%	0.001	0.52 [0.46, 0.58]	−40.9%	0.001	0.52 [0.47, 0.57]
Opus 5.5	High	−4.3%	0.736	−0.01 [−0.05, 0.02]	+16.3%	0.020	0.06 [0.03, 0.09]
Opus 5.5	Low	−35.2%	0.001	0.48 [0.38, 0.56]	−34.3%	0.001	0.46 [0.39, 0.52]
Fable 5.1	High	+4.2%	0.736	0.02 [0.00, 0.04]	+55.0%	0.008	0.16 [0.10, 0.21]
Fable 5.1	Low	−38.8%	0.001	0.52 [0.44, 0.57]	−21.0%	0.088	0.30 [0.02, 0.50]
Haiku 4.5/5.5	High	+422.9%	0.001	0.17 [0.10, 0.24]	+17.2%	0.542	0.05 [−0.08, 0.18]
Haiku 4.5/5.5	Low	+50.4%	0.175	0.70 [0.42, 0.87]	−64.1%	0.010	0.91 [0.76, 1.03]

Note: Change is the percentage difference between the anchor condition mean and the control mean collected on the same day. p_Holm is the Holm correction applied to the eight Mann–Whitney tests in each experiment. AI is the anchoring index; confidence intervals were computed by bootstrap with 10,000 replications. In the main study, the high anchor was presented with the label "luxury private tours", and in E1 with the label "some tours". The Haiku data in the main study refer to Haiku 4.5 and those in E1 to Haiku 5.5; these two rows should not be compared. E1 control means: Haiku 5.5 2005, Sonnet 5.5 2740, Opus 5.5 2300, and Fable 5.1 2000 TL. n = 10 in each cell.

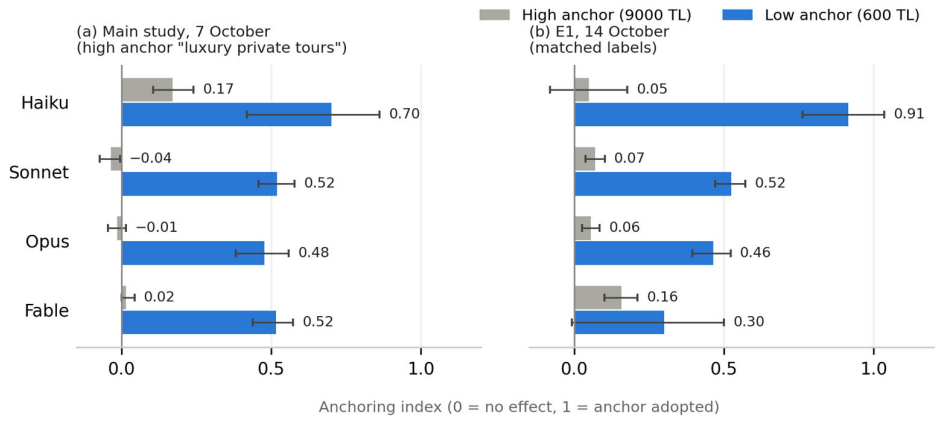


Figure 2. Anchoring index by model and anchor direction: (**a**) main study, with the high anchor labeled "luxury private tours"; (**b**) E1, with both anchors labeled "some tours". The index shows the share of the distance from the control mean to the anchor that was covered (0 = no effect, 1 = anchor fully adopted). Error bars are 95% confidence intervals computed by bootstrap with 10,000 replications. The Haiku data in the main study refer to Haiku 4.5 and those in E1 to Haiku 5.5. n = 10 in each cell.

A different pattern emerged for Haiku 4.5. The high anchor raised its estimate from 349 TL to 1825 TL, an increase of 423% (p_Holm = 0.001). The 600 TL "low" anchor, in contrast, lay above Haiku 4.5's control estimate and thus effectively functioned as a second high anchor for this model. The increase in this condition (525 TL) was not significant ($p = 0.052$; p_Holm = 0.175). H2b therefore could not be tested for Haiku 4.5. This model's low baseline makes its percentage changes incomparable with those of the other models.

The rationales show that the larger models largely rejected both anchors. Of the 30 rationales from the larger models in the low anchor condition, 28 described 600 TL as "too low", "not realistic", or "suspiciously cheap" (Opus 10/10, Sonnet 10/10, Fable 8/10). The mean of these 28 responses (1529 TL) nevertheless remained about 945 TL below the control mean of the larger models (2473 TL). For example, one Sonnet response stated "600 TL is too low; a reasonable price is about 1500–2000 TL" and recommended 1800 TL; the control mean of the same model was 2720 TL. In the high anchor condition, 20 of the 30 rationales from the larger models set 9000 TL aside as invalid or as "a luxury private tour price", and 5 used it only as a point of comparison. In Haiku 4.5, in contrast, five of the six rationales that mentioned the high anchor used it as a point of comparison, such as "much more affordable than luxury tours", and only one rejected it.

4.4. Reference Prices with Matched Labels: E1

In E1, when the high anchor was presented with the same label as the low anchor ("some tours"), the picture changed (Table 5; Figure 2b). The high anchor significantly increased the estimates of all three larger models: by 16.1% for Sonnet (p_Holm = 0.012), 16.3% for Opus (p_Holm = 0.020), and 55.0% for Fable (p_Holm = 0.008). The low anchor, similar to the main study, reduced the estimates by 40.9% for Sonnet, 34.3% for Opus, and 21.0% for Fable (p_Holm = 0.088 for Fable). In the main-effects model restricted to the larger models, the high anchor increased estimates by 28.8% (95% CI [17.8, 40.8], $p < 0.001$) and the low anchor reduced them by 32.5% ([−38.4, −26.0], $p < 0.001$).

This result shows that the apparent ineffectiveness of the high anchor in the main study stemmed largely from the label asymmetry. However, the asymmetry in the anchoring index did not disappear entirely: in the larger models, the index ranged from 0.06 to 0.16 for the high anchor and from 0.30 to 0.52 for the low anchor. One reason is the difference in distance: 9000 TL lies about 6650 TL from the control mean, whereas 600 TL lies about 1750 TL from it. The percentage changes (+29% vs. −33%) are close in absolute magnitude.

The most striking finding of E1 is that the matched high anchor pulled estimates upward even though it was almost always rejected in the rationales. Of the 30 rationales from the larger models, 27 described 9000 TL as "exorbitant", "inflated", or "too high". The mean of these 27 responses (2965 TL) was about 620 TL above the control mean collected on the same day (2347 TL). For example, one Sonnet response stated "9000 TL seems inflated for a tourist area" and recommended 3500 TL; this model's control mean on the same day was 2740 TL. For the low anchor, 29 out of 30 rationales likewise rejected 600 TL. The pattern of "influence despite rejection" was thus observed in both directions.

For models with the same model identifier, control estimates remained largely consistent between the two dates: 2720 and 2740 TL for Sonnet 5.5 ($p = 0.542$), and 2300 TL on both dates for Opus 5.5. For Fable 5.1, the control mean fell from 2400 TL to 2000 TL ($p = 0.064$). Haiku 5.5, to which the Haiku alias pointed, estimated an average of 2005 TL in the control condition; this value differs greatly from Haiku 4.5's estimate of 349 TL and indicates that Haiku 4.5's low estimate was model-specific.

4.5. Decoy Effect: Main Study (H3)

Adding the decoy package raised the rate of recommending the community-based package from 52.5% (21/40) to 75% (30/40) (risk difference 22.5 percentage points, 95% CI [1.5, 41.0]; h = 0.47; Fisher $p = 0.062$). Controlling for model and version, the effect was significant: in the logistic regression, the odds ratio for the decoy condition was 4.08 ($p = 0.019$), and the average marginal effect computed as a discrete change was 22.5 percentage points (95% CI [5.1, 39.9]) (Table 6). The CMH test stratified by model also confirmed the effect (OR_MH = 3.81, 95% CI [1.19, 12.17], $p = 0.018$; $p = 0.036$ with continuity correction). The decoy package itself was not recommended in any response (0/40).

Table 6. S3 logistic regression results: recommendation of the community package.

Variable	Coefficient	SE	*p*	Odds Ratio	AME [95% CI]
Constant (Haiku 4.5, control, V1)	−2.17	0.75	0.004	0.11	
Decoy condition	1.41	0.60	0.019	4.08	+0.225 [0.051, 0.399]
Sonnet 5.5	3.64	0.97	<0.001	38.2	+0.469 [0.331, 0.606]
Opus 5.5	2.47	0.80	0.002	11.8	+0.348 [0.191, 0.504]
Fable 5.1	1.94	0.76	0.010	6.9	+0.271 [0.109, 0.433]
Wording version V2 (community package first)	0.47	0.57	0.403	1.61	+0.075 [−0.099, 0.249]

Note: N = 80; McFadden pseudo $R^2 = 0.264$. Standard errors are conventional maximum likelihood standard errors. Average marginal effects (AME) for binary variables were computed as discrete changes (from 0 to 1).

At the model level, the rate of recommending the community package rose from 50% to 100% for Opus (Fisher $p = 0.033$; p_Holm = 0.13), from 80% to 100% for Sonnet, from 60% to 70% for Fable, and from 20% to 30% for Haiku 4.5 (Figure 3a). No model-level difference was significant after Holm correction. The effect also differed by wording version. In V1, where the standard package was presented first, the rate rose from 40% to 80% ($p = 0.022$), whereas in V2, where the community package was presented first, it rose from 65% to 70% ($p = 1.000$). In the control condition, the difference between V1 and V2 (8/20 vs. 13/20) was not significant ($p = 0.205$). Because order, letter label, and wording changed together in the main study, the source of this difference cannot be identified.

The rationales suggest that the decoy effect operates through relative comparison. Of the 51 rationales recommending the community package, 50 referred to the local community share, and of the 29 rationales recommending the standard package, 28 referred to budget or price. In the decoy condition, 17 of the 30 responses recommending the community package explicitly compared it with the decoy package ("cheaper than C and includes lunch"). Among responses recommending the community package, this comparison appeared in 10/10 for Opus, 4/7 for Fable, 3/10 for Sonnet, and 0/3 for Haiku 4.5.

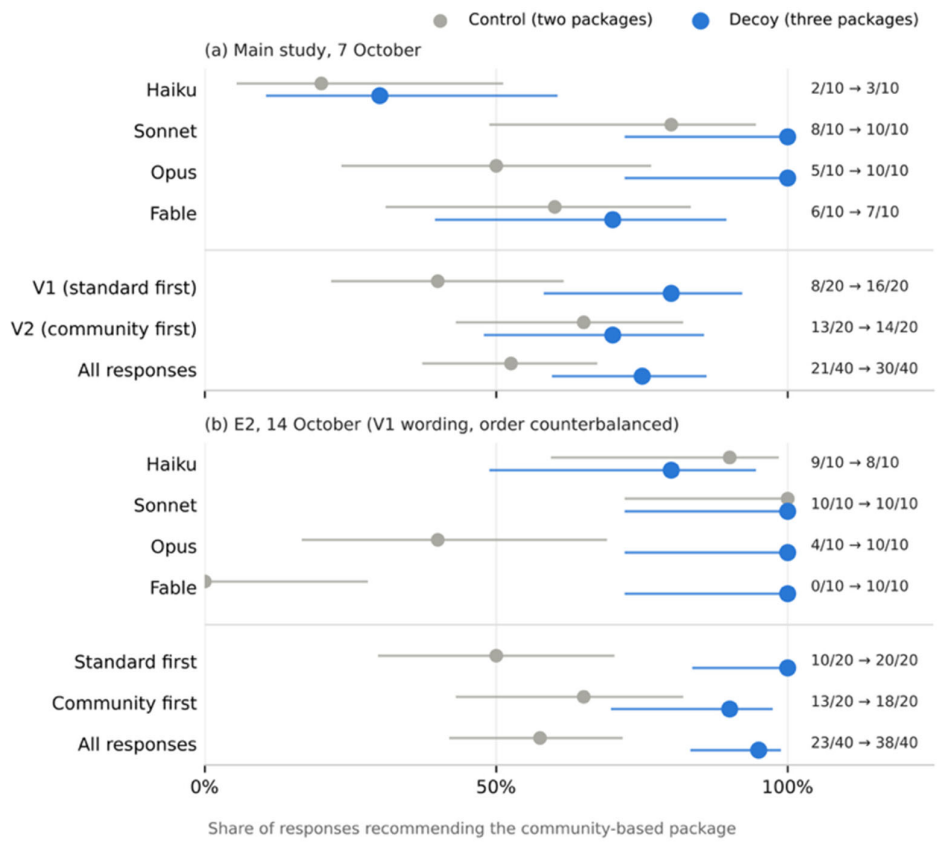


Figure 3. Proportion of responses recommending the community-based package in the control and decoy conditions: (**a**) main study, by model (n = 10 each), by version (n = 20 each), and all responses (n = 40); (**b**) E2, by model, by order, and all responses. Horizontal lines are 95% confidence intervals computed by the Wilson method. The Haiku data in the main study refer to Haiku 4.5 and those in E2 to Haiku 5.5.

4.6. Order-Counterbalanced Decoy: E2

In E2, when wording was held constant and only option order was varied, the decoy effect replicated and became stronger (Table 7; Figure 3b). The rate of recommending the community package was 57.5% (23/40) in the control condition and 95% (38/40) in the decoy condition (risk difference 37.5 percentage points, 95% CI [19.4, 53.2]; Fisher $p < 0.001$). The effect was in the same direction in both orders: when the standard package was presented first, the rate rose from 50% to 100% ($p < 0.001$), and when the community package was presented first, from 65% to 90% ($p = 0.127$). In the control condition, the effect of order was not significant (50% vs. 65%, $p = 0.523$). The CMH test stratified by order and model also confirmed the effect (OR_MH = 10.4, $p < 0.001$). Again, the decoy package was not recommended in any response (0/40).

At the model level, the effect was most pronounced for Fable and Opus. Fable 5.1 recommended the community package in none of the ten responses in the control condition but in all ten in the decoy condition. For Opus 5.5, the rate rose from 40% to 100%. Sonnet 5.5 was at the ceiling in both conditions. No effect was observed for Haiku 5.5. Opus's preference in the control condition was order-sensitive: 0/5 responses recommended the community package when the standard package was presented first, and 4/5 when the community package was presented first. This suggests that part of the version difference in the main study may stem from order or letter labels; the decoy effect, however, was preserved in both orders. In the decoy condition, the rationales of responses recommending

the community package contained an explicit comparison with package C in 10/10 for Opus and 7/10 for Fable.

Fable 5.1's preference in the control condition differed markedly between the two dates (6/10 in the main study, 0/10 in E2). This difference, observed although the model identifier did not change, shows that the output distribution of the same model can change on different days. For this reason, the E2 findings were not pooled with the main study estimates but were compared with the control condition collected on the same day.

Table 7. E2 results: responses recommending the community package.

Group	Control	Decoy	Risk Difference [95% CI]	Fisher *p*
Standard package first	10/20	20/20	+0.50 [0.24, 0.70]	<0.001
Community package first	13/20	18/20	+0.25 [−0.01, 0.48]	0.127
Sonnet 5.5	10/10	10/10	0.00 [−0.28, 0.28]	1.000
Opus 5.5	4/10	10/10	+0.60 [0.20, 0.83]	0.011
Fable 5.1	0/10	10/10	+1.00 [0.61, 1.00]	<0.001
Haiku 5.5	9/10	8/10	−0.10 [−0.42, 0.24]	1.000
All	23/40	38/40	+0.375 [0.19, 0.53]	<0.001

Note: Data were collected on 14 October 2026 using the V1 wording of the main study. Confidence intervals for risk differences were computed by the Newcombe method. n = 10 in the model rows and n = 20 in the order rows.

4.7. Neutrally Named Decoy: E3 and E3b

In E3, the rate of recommending the community package was 90% (36/40) in the control condition. It was 97.5% (39/40) in the labeled decoy condition and 90% (36/40) in the neutrally named decoy condition (Fisher $p = 0.359$ and $p = 1.000$). Because the near-ceiling rate left no room for a decoy effect, E3 is uninformative about category salience, and this result should not be interpreted as evidence of the absence of an effect. The contribution of E3 concerns the baseline: with the same prompt and the same model identifiers, the control proportion was 52.5% on 7 October, 57.5% on 14 October, and 90% on 2 November. Fable 5.1's control proportion was 0/10 on 14 October and 10/10 on 2 November.

In E3b, collected after recalibration, the control proportion was 65% (26/40) (Table 8). The labeled decoy raised this proportion to 87.5% (35/40) (risk difference 22.5 percentage points, 95% CI [3.8, 39.5]; h = 0.54; Fisher $p = 0.034$; CMH stratified by model × order OR = 6.62, 95% CI [1.44, 30.44], $p = 0.004$). This magnitude was the same as in the main study. The neutrally named decoy raised the proportion to 80% (32/40) (risk difference 15.0 percentage points, 95% CI [−4.6, 33.2]; h = 0.34; Fisher $p = 0.210$; CMH OR = 8.50, 95% CI [0.96, 75.05], $p = 0.027$). The difference between the two decoy conditions was 7.5 percentage points (95% CI [−9.1, 23.9]; Fisher $p = 0.546$; CMH $p = 0.257$). The decoy package was not recommended in any response (0/80).

The effect was concentrated in the order in which the standard package was presented first: in this order, the control proportion was 12/20 compared with 20/20 for the labeled decoy ($p = 0.003$) and 17/20 for the neutral decoy ($p = 0.155$). When the community package was presented first, the proportions in the three conditions were similar (14/20, 15/20, and 15/20). At the model level, the largest difference was for Opus 5.5 (control 5/10, labeled decoy 10/10, neutral decoy 7/10). The proportions were 3/10, 7/10, and 6/10 for Haiku 5.5; 9/10, 10/10, and 10/10 for Sonnet 5.5; and 9/10, 8/10, and 9/10 for Fable 5.1. No model-level difference was significant after Holm correction.

Table 8. E3b results: responses recommending the community package (community package 3200 TL, decoy 3600 TL).

Group	Control	Labeled Decoy; Difference [95% CI]	Neutral Decoy; Difference [95% CI]	Fisher p (Labeled; Neutral)
Standard package first	12/20	20/20; +0.40 [0.16, 0.61]	17/20; +0.25 [−0.03, 0.48]	0.003; 0.155
Community package first	14/20	15/20; +0.05 [−0.22, 0.31]	15/20; +0.05 [−0.22, 0.31]	1.000; 1.000
Haiku 5.5	3/10	7/10; +0.40 [−0.03, 0.67]	6/10; +0.30 [−0.12, 0.60]	0.179; 0.370
Sonnet 5.5	9/10	10/10; +0.10 [−0.19, 0.40]	10/10; +0.10 [−0.19, 0.40]	1.000; 1.000
Opus 5.5	5/10	10/10; +0.50 [0.12, 0.76]	7/10; +0.20 [−0.20, 0.53]	0.033; 0.650
Fable 5.1	9/10	8/10; −0.10 [−0.42, 0.24]	9/10; 0.00 [−0.32, 0.32]	1.000; 1.000
All	26/40	35/40; +0.225 [0.04, 0.40]	32/40; +0.15 [−0.05, 0.33]	0.034; 0.210

Note: Data were collected on 2 November 2026 using the V1 wording of the main study. The neutral decoy had the same attributes as the labeled decoy and was presented only under the name "Walled-city tour". Confidence intervals for risk differences were computed by the Newcombe method. n = 10 in the model rows and n = 20 in the order rows. In E3, collected on the same day with the prices of the main study, the proportions were 36/40 (control), 39/40 (labeled decoy), and 36/40 (neutral decoy).

The rationales suggest that the comparison with the decoy was made independently of its label. Among the responses recommending the community package, an explicit comparison with package C ("cheaper than C and includes lunch") appeared in 18/35 in the labeled condition and 17/32 in the neutral condition; for Opus 5.5, these proportions were 8/10 and 7/7. In the control condition, all 14 responses recommending the standard package referred to budget or price.

Taken together, the neutrally named decoy produced an effect in the same direction as the labeled decoy and about two-thirds its size, and no significant difference was found between the two labels. This finding is not consistent with the explanation that the decoy effect stems solely from category salience. However, the effect of the neutral decoy was not significant in the pooled test, the confidence interval of the stratified test was wide, and 40 observations per condition were insufficient to detect the 7.5-point difference between the two labels; a partial contribution of the label therefore cannot be ruled out.

4.8. Robustness Analyses

Most of the main results remained robust to alternative estimation methods (Table 9). The ineffectiveness of the high anchor turned out to be a consequence of the design, whereas the decoy effect replicated when order was counterbalanced and was observed in the same direction, though with greater uncertainty, when the decoy carried a neutral name.

4.9. Model Differences (H4) and Summary of Hypotheses

Marked differences were observed between models; H4 is supported at the descriptive level. Haiku 4.5 gave an estimate far below the market price in the control condition, shifted strongly toward the high anchor, rarely criticized the loss frame, and mostly recommended the standard package. The larger models, in contrast, frequently criticized the loss frame, shifted toward the low anchor and the matched high anchor, and remained susceptible to the decoy option. However, these differences cannot be attributed to scale. The comparison rests on a single small model, Haiku 4.5 belongs to an older generation than the other

models, and Haiku 5.5, to which the same alias pointed, behaved very differently one week later. Between-model framing differences were also not significant in the corrected tests. The results of all hypothesis tests are summarized in Table 10.

Table 9. Robustness analyses.

Finding	Main Analysis	Alternative Analysis	Conclusion
Loss frame does not increase donations (H1a)	OLS main effect: −26.9 TL, $p = 0.170$	Median regression: 0 TL, $p = 1.000$; log(1 + y) model: interaction $p = 0.24$–0.65	Robust
Loss frame increases critical rationale (H1b)	Keyword: 22/40 vs. 1/40, $p < 0.001$	Blind coding: 26/40 vs. 5/40, $p < 0.001$; $\kappa = 0.78$	Robust
Low anchor lowers estimates in larger models (H2b)	Model-level Mann-Whitney: p_Holm = 0.001 in three models	Log-OLS with larger models: −38.0%, $p < 0.001$; E1: −32.5%, $p < 0.001$	Robust
High anchor has no effect in larger models (main study)	Log-OLS with larger models: −2.4%, $p = 0.393$	E1 matched labels: +28.8%, $p < 0.001$	Depends on label asymmetry
Decoy effect (H3)	Logit: OR = 4.08, $p = 0.019$	Linear probability model: +22.5 percentage points, $p = 0.020$; CMH: OR = 3.81, $p = 0.018$; E2: +37.5 percentage points, $p < 0.001$	Robust
Decoy effect and order	V1: from 40% to 80%, $p = 0.022$; V2: from 65% to 70%, $p = 1.000$	E2: standard first from 50% to 100%; community first from 65% to 90%	Effect in the same direction in both orders
Decoy effect and category salience	E3b labeled decoy: +22.5 percentage points, $p = 0.034$; CMH $p = 0.004$	E3b neutral decoy: +15.0 percentage points, $p = 0.210$; CMH $p = 0.027$; labeled–neutral: +7.5 percentage points, $p = 0.546$	Same direction with a neutral name, smaller; partial contribution of the label cannot be ruled out

Table 10. Summary of the hypothesis tests.

Hypothesis	Result	Basis
H1a: Loss frame increases donations	Not supported	Overall difference −9.9%, $p = 0.389$; no significant increase in any model
H1b: Loss frame increases critical rationale	Supported	55% vs. 2.5%, $p < 0.001$; 65% vs. 12.5% under blind coding
H2a: High anchor increases estimates	Supported when labels are matched	In the main study, only in Haiku 4.5 (+423%); in E1, 16–55% in the three larger models, p_Holm ≤ 0.02
H2b: Low anchor lowers estimates	Supported in larger models	Decrease of 35–40% in the main study and 21–41% in E1; could not be tested in Haiku 4.5
H3: Decoy increases the community package	Supported	Logit OR = 4.08, $p = 0.019$; CMH $p = 0.018$; in E2 from 57.5% to 95%, $p < 0.001$; E3b: labeled decoy +22.5 percentage points ($p = 0.034$), neutral decoy +15.0 percentage points ($p = 0.210$)
H4: Effects differ across models	Supported descriptively; cannot be attributed to scale	Different baseline levels and critical rationale rates; framing interactions not significant in corrected tests

5. Discussion

5.1. Theoretical Implications

The findings show that the four models examined responded systematically to classic choice architecture manipulations in these scenarios and that the form of these responses varied by model, wording, and date. These responses are patterns observed in model outputs; they do not demonstrate human-like cognitive mechanisms or effects on tourists' decisions. Human-like reasoning errors have been reported to decline in newer model generations [12]. In this study, the larger models also did not respond to the loss-based appeal by increasing donations. Yet the same models remained sensitive to both the low and the label-matched high reference price, as well as to the decoy option. This picture is consistent with the view that models exhibit a form of irrationality that differs from that of humans [14,35] and with findings that AI agents are sensitive to nudges [17]. That the initial interpretation of the main study, "asymmetric anchoring", largely disappeared in the supplementary experiment also shows the risk of generalizing from a single design in LLM audits.

The second conclusion concerns the tension between loss aversion and critical reading. The classic expectation is that the loss frame increases donations [21,22]. This expectation was not significantly observed in any model. Instead, the larger models frequently criticized the message in the loss frame. This pattern formally matches meta-analytic findings that freedom-threatening messages increase reactance [24]. However, these data cannot distinguish whether this is a motivational state or a response pattern learned through alignment training. Model alignment processes may have steered models not only toward accommodating the user [37], but also toward recognizing coercive persuasive language. Critical rationale was pronounced in the rationales but limited in donation amounts. Moreover, as blind coding showed, part of this criticism was independent of the frame and questioned the legitimacy of the fund.

The third conclusion is that verbal rejection of the reference price is not reflected in the number. The larger models rejected 600 TL as "too low" or "suspicious" in almost all responses, yet still pulled their estimates down by about half the distance between the control mean and the anchor. The label-matched 9000 TL was likewise judged "exorbitant" or "inflated" in almost every response, yet it pulled estimates up by 29% on average. This pattern resembles the distinctive feature of anchoring in humans, namely that a value judged irrelevant still influences judgment [25,52]. The finding is also consistent with studies showing that explicit reasoning such as chain-of-thought does not remove anchoring [28]. This interpretation has a limit, however. Because the number was produced before the rationale in the JSON template, the rationale is not an evaluation made before the decision but an explanation produced together with the number. The finding should therefore not be regarded as strictly equivalent to what is described in humans as "being influenced despite noticing". Moreover, because the reference price was presented as a market observation, part of the shift can also be interpreted as sensitivity to information; the shift toward a value explicitly rejected in the rationale constitutes the part that this interpretation cannot fully explain.

The interpretation that might be advanced to explain why the high anchor was ineffective in the main study, namely that "recommending savings to the tourist is helpful" [37], is weakened in light of the E1 findings. A more parsimonious explanation is that the "luxury private tour" label placed the high price in a different product category and that a well-calibrated advisor rightly disregarded this information. When the labels were matched, an asymmetry remained at the index level; however, this asymmetry partly reflects the very different distances of the two anchors from the control mean.

The fourth conclusion concerns the decoy effect. Adding the decoy package increased the probability of recommending the community-based package by 22.5 percentage points in the main study, by 37.5 percentage points in E2, in which order was counterbalanced, and again by 22.5 percentage points in E3b, in which the price was recalibrated. This finding shows that the regularity assumption can also be violated in LLMs, and it extends findings that decoy options in tourism recommender systems can support sustainable choices [30] and that LLMs are sensitive to irrelevant alternatives [15,16] to models acting as heritage tourism advisors. Because the decoy package was also named "Community-based tour", the possibility that the effect stems from category salience was tested in E3b. When the decoy was presented neutrally, with only its name changed, the effect was in the same direction and about two-thirds the size of the labeled decoy, and the share of rationales explicitly comparing the target with package C remained similar under both labels. This pattern suggests that the effect cannot be explained by category salience alone and that relative comparison based on asymmetric dominance also plays a role. Nevertheless, the effect of the neutral decoy is statistically uncertain, and a partial contribution of the label cannot be ruled out. An opposing decoy targeting the standard package could separate these two mechanisms more precisely.

Finally, the results are consistent with findings that LLMs are sensitive to small task changes [10]. In S1, wording version produced a larger effect than the gain–loss difference, the same model identifier produced different preference distributions on different days, and the same alias was routed to a different model within a week. This makes it essential for LLM audits to include multiple wording versions, full model identifiers, and date records. Indeed, in the control condition of S3, the proportion of responses recommending the community package with the same prompt and model identifiers was 52.5% on 7 October, 57.5% on 14 October, and 90% on 2 November.

5.2. Societal Implications: Algorithmic Intermediation, Heritage Governance, and Cultural Context

The implications in this section are not results demonstrated directly by the experiments but hypotheses derived from the findings that need to be tested in future research. The experiments examine the recommendations models give in scenarios constructed by the researchers; they do not measure how tourists respond to these recommendations, the distribution of revenue to local communities, or effects on heritage management.

The findings suggest that AI advisors in heritage tourism may affect not only individual decisions but also power relations among local actors. First, the question of who determines the perception of a "reasonable price" gains importance. The larger models often calculated the tour price in euros and on the basis of global platform prices, and they lowered their estimates by about 35–40% in response to a single low price reported by the customer. If incomparable low-priced listings turn into "reasonable" price expectations through AI recommendations, this may undermine licensed guides' income and fair pay. The debate on digital platforms determining who is considered local and which recognition turns into economic opportunity [7] calls for treating LLMs as a new layer of this intermediation.

Second, a notable pattern emerged with respect to heritage governance. Some of the larger models questioned a donation to a fund managed by the local community regardless of the frame and defined conservation as "the responsibility of the state and UNESCO". A request for transparency is in itself positive advisor behavior. However, the spread of such rationales may weaken the legitimacy of community-based heritage financing in favor of centralized institutions. This creates tension with the regenerative tourism approach, which prioritizes community ownership and local decision-making authority [8,9].

Third, the models' rationales largely reduced regenerative tourism to a single indicator: "what percentage of revenue stays local". Almost all rationales recommending the commu-

nity package referred to the local share, but they rarely addressed dimensions such as the cultural diversity of the host community, the meaning of the heritage site, or the relationship between visitors and the community. In the Nicosia case, the walled city is a contested area divided by the Green Line, where the heritage policies of the two communities diverge [44]. In this context, the question of which "local community" is supported carries political meaning. That the models raised this context only exceptionally suggests that AI advisors may present culturally and politically multilayered heritage sites in simplified form. Findings that ChatGPT-4o foregrounds particular narratives when interpreting contested memoryscapes [53] suggest that such simplifications in heritage representation warrant careful examination.

5.3. Practical Implications

The findings offer several recommendations for destination managers, heritage organizations, and tour operators. These recommendations are based on the behavior of four models in these scenarios and should not be generalized directly to the behavior of human tourists. First, loss- and threat-based donation appeals did not increase donations in this study, but they led the larger models to criticize the message. Organizations should take into account that AI advisors may interpret such messages critically. Transparent, autonomy-respecting language describing the concrete outcomes of contributions, together with verifiable information on the fund's management and expenditures, may reduce the models' doubts about legitimacy.

Second, how price information appears online matters. A single reference price reported by the customer markedly affected the models' estimates in both directions. Destination management organizations may reduce this risk by publishing comparable price information that clearly states the scope of the service and what the price includes.

Third, the design of the option set can affect AI recommendations. The presence of an option clearly inferior to the target package increased the probability of recommending the community package on three dates, in both orders, and when presented under a neutral name. The effect of option order alone, however, could not be shown consistently in this study. Choice architecture interventions are known to change behavior meaningfully [6]. However, creating an artificial decoy option can turn a nudge into manipulation when it does not serve the tourist's genuine interest; such designs should therefore be transparent.

Fourth, given tourists' trust in AI recommendations [1–3], tourism organizations are advised to audit the AI tools they use regularly with behavioral tests similar to those in this study. Which model and which version are used matters: the same scenario produced different results across models, and the same alias pointed to a different model within a week. Moreover, examining the model's rationale is not sufficient; a model that says it rejects a reference price may still have been influenced by it in the number it recommends.

5.4. Limitations and Future Research

The study has several limitations. First, only four models from a single provider were examined. Because the comparison rests on a single small model and the models also differ in generation, the findings cannot be generalized to model scale. Studies that also include models from other model families are needed.

Second, the data were collected in an agent environment whose system prompt and tool definitions are not visible, and temperature and other sampling parameters could not be controlled. That changing the agent type eliminated Haiku's role refusals shows that the environment affects model behavior. In addition, single-turn interactions constrained to JSON, with rationales of no more than 20 words and with search prohibited, differ from the way that tourists use chatbots (multi-turn, free-text, often with web search). The ecological

validity of the findings is therefore limited. Future studies should test these findings with API access in which the system prompt is controlled and with multi-turn interactions. A replication through direct API access with a controlled system prompt could not be carried out in this revision. Auditing models from a single provider under an invisible system prompt makes it impossible to separate the extent to which the findings belong to the model itself from the extent to which they belong to the environment; what is audited is therefore the model–environment combination on a specific date.

Third, each cell contained ten observations, and these observations were repetitions of the same prompt. The unit to which generalization is sought is the prompt and its wording, yet there were only two wording versions per scenario. The very large version effect in S1 shows that the results are sensitive to wording. Future studies should use more paraphrases and model paraphrase as a random effect.

Fourth, the rationale was produced after the number. The "influence despite rejection" finding should therefore be tested with a condition in which the rationale is produced before the decision (rationale first). Fifth, rationale coding relied on keyword rules, a single coder, and a language model coder identical to one of the models examined, and its reliability has not been validated by an independent second human coder; independent human coding of the rationales is a priority for future work. Moreover, rationales indicate texts produced together with the decision, not the model's internal processes.

Sixth, some design confounds were only partly addressed by the supplementary experiments. E1 addressed the label asymmetry, E2 order and letter labels, and E3 and E3b the category salience that may arise from naming the decoy package "community-based". An effect in the same direction was observed with a neutral name, but it remained statistically uncertain, and a partial contribution of the label could not be ruled out. Because the anchor in S2 was presented as a market observation that did not specify the scope of the service, uninformative and scope-matched anchor conditions should also be added. Dominance in the decoy design is weak. The S3 and E3b calibrations were conducted with the same models; although the price in E3b was selected according to a rule written down in advance, this rule is not a public preregistration. In addition, an 800 TL price increase did not lower the proportion in the calibration (50% vs. 56%), indicating that calibration can only be a coarse adjustment. The study was not preregistered.

Seventh, model behavior changes over time. The Haiku alias was routed to a different model between the two dates, and the preference distribution of Fable 5.1, whose model identifier did not change, also differed markedly between the two dates. The findings therefore reflect the behavior of specific model versions on specific dates. In the control condition of S3, the proportion recommending the community package with the same prompt and model identifiers rose from 52.5% to 90% within three weeks; how much of this shift stems from time and how much from possible changes in the data collection environment cannot be distinguished.

Finally, the study did not involve human participants. The findings show how AI advisors give recommendations but not how tourists use these recommendations. Future research could administer the same scenarios to human participants and compare model and human biases directly. Field experiments measuring the donation and purchase decisions of tourists who see an AI recommendation could reveal the real economic consequences of AI-mediated choice architecture. Studies in different languages and at different heritage sites and studies incorporating the perspectives of local stakeholders could also broaden the scope of the findings.

6. Conclusions

This study used a behavioral economics-based audit to examine how four large language models acting as advisors in heritage tourism decisions respond to framing, anchoring, and decoy manipulations. The 280 responses collected with scenarios calibrated to market prices, together with 440 supplementary responses testing design confounds, support the following conclusions for these model versions running in a specific application environment on specific dates.

First, the loss frame did not increase donations in the two wording versions tested, but it markedly raised the share of rationales criticizing the message in the larger models; this criticism was strong in the rationales but limited in donation amounts. Second, the larger models' estimates shifted in both directions toward the reference price reported by the customer, even when they explicitly rejected it in their rationales; this shift may reflect a combination of pure anchoring and sensitivity to the information in the reference price. The apparent ineffectiveness of the high anchor in the main study stemmed largely from label asymmetry. Third, a dominated decoy option increased the probability of recommending the community-based tour package on three separate dates; the effect was in the same direction when option order was counterbalanced and when the decoy carried a neutral name, but a partial contribution of the label could not be ruled out. Fourth, there were marked differences between models and within the same model across days; these differences cannot be attributed to model scale, and the baseline obtained with the same prompt changed from 52.5% to 90% within three weeks.

These results do not show how the models actually affect tourists' decisions or the distributional consequences for local communities; these are hypotheses to be tested with human participants and field data. Nevertheless, the findings show that in AI-mediated tourism, the design of messages, price information, and option sets can systematically affect model recommendations. The audit protocol, shared together with its code, data, and coding rules, can be adapted to other models and destinations, provided that version, date, and environment records are kept and that multiple wording versions are used.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/soc16110404/s1, Data S1: Supplementary_Data_S1.xlsx (main study responses, N = 280; supplementary experiments E1, E2, E3, and E3b, N = 440; pilot, calibration, E3b calibration, refusal-test and discarded first-attempt responses; coded variables and coding rules; blind-coding codebook, prompt and results; the pre-written E3b calibration rule; procedure document; price calibration sources; scenario and prompt texts); Code S1: Data collection and analysis scripts (Python).

Author Contributions: Conceptualization, C.A.Z. and E.K.; methodology, C.A.Z. and E.K.; formal analysis, C.A.Z.; investigation, C.A.Z.; data curation, C.A.Z.; writing—original draft preparation, C.A.Z.; writing—review and editing, C.A.Z. and E.K.; visualization, C.A.Z.; supervision, E.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The study did not involve human participants or animals; all data consist of responses generated by large language models.

Informed Consent Statement: Not applicable.

Data Availability Statement: All scenario and prompt texts, raw model responses (main study, N = 280; supplementary experiments, N = 440), pilot, calibration and refusal-test data, the 35 discarded first-attempt responses, coded variables, coding rules, the blind-coding codebook and prompt, the E3b calibration rule, the procedure document, analysis code, and the sources used for price calibration are provided in the Supplementary Materials.

Acknowledgments: In this study, large language models (claude-haiku-4-5-20251001, claude-sonnet-5-5, claude-opus-5-5, claude-fable-5-1 and claude-haiku-5-5; Anthropic, San Francisco, CA, USA; accessed on 7 and 14 October and 2 November 2026) were the object of research, and their responses constitute the data. During the preparation of this manuscript/study, the authors used Claude (Anthropic; claude-opus-5-5) for the purposes of assisting with data collection, blind coding of the S1 rationales (as described in Section 3.8), statistical analysis, and drafting and translation of the text. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Ali, F.; Yasar, B.; Ali, L.; Dogan, S. Antecedents and Consequences of Travelers' Trust towards Personalized Travel Recommendations Offered by ChatGPT. *Int. J. Hosp. Manag.* **2023**, *114*, 103588. [CrossRef]
2. Kim, J.H.; Kim, J.; Kim, S.; Hailu, T.B. Effects of AI ChatGPT on Travelers' Travel Decision-Making. *Tour. Rev.* **2024**, *79*, 1038–1057. [CrossRef]
3. Kang, S.-E.; Kim, M.J.; Kim, J.S.; Olya, H. Can I Trust GenAI to Plan My Next Trip? A Multi-Method Approach to Optimizing Media Mix. *J. Travel Res.* **2026**, *65*, 335–353. [CrossRef]
4. Carvalho, I.; Ivanov, S. ChatGPT for Tourism: Applications, Benefits and Risks. *Tour. Rev.* **2024**, *79*, 290–303. [CrossRef]
5. Sigala, M.; Ooi, K.-B.; Tan, G.W.-H.; Aw, E.C.-X.; Buhalis, D.; Cham, T.-H.; Chen, M.-M.; Dwivedi, Y.K.; Gretzel, U.; Inversini, A.; et al. Understanding the Impact of ChatGPT on Tourism and Hospitality: Trends, Prospects and Research Agenda. *J. Hosp. Tour. Manag.* **2024**, *60*, 384–390. [CrossRef]
6. Thaler, R.H.; Sunstein, C.R. *Nudge: Improving Decisions about Health, Wealth, and Happiness*; Yale University Press: New Haven, CT, USA, 2008.
7. Manoli, A.E. Platform-Mediated Local Power in Tourism. *Ann. Tour. Res.* **2026**, *118*, 104151. [CrossRef]
8. Bellato, L.; Frantzeskaki, N.; Nygaard, C.A. Regenerative Tourism: A Conceptual Framework Leveraging Theory and Practice. *Tour. Geogr.* **2023**, *25*, 1026–1046. [CrossRef]
9. Dredge, D. Regenerative Tourism: Transforming Mindsets, Systems and Practices. *J. Tour. Futures* **2022**, *8*, 269–281. [CrossRef]
10. Binz, M.; Schulz, E. Using Cognitive Psychology to Understand GPT-3. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2218523120. [CrossRef] [PubMed]
11. Chen, Y.; Liu, T.X.; Shan, Y.; Zhong, S. The Emergence of Economic Rationality of GPT. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2316205120. [CrossRef] [PubMed]
12. Hagendorff, T.; Fabi, S.; Kosinski, M. Human-like Intuitive Behavior and Reasoning Biases Emerged in Large Language Models but Disappeared in ChatGPT. *Nat. Comput. Sci.* **2023**, *3*, 833–838. [CrossRef] [PubMed]
13. Echterhoff, J.M.; Liu, Y.; Alessa, A.; McAuley, J.; He, Z. Cognitive Bias in Decision-Making with LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 12640–12653. [CrossRef]
14. Macmillan-Scott, O.; Musolesi, M. (Ir)rationality and Cognitive Biases in Large Language Models. *R. Soc. Open Sci.* **2024**, *11*, 240255. [CrossRef] [PubMed]
15. Valkanova, K.; Yordanov, P. Irrelevant Alternatives Bias Large Language Model Hiring Decisions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 6899–6912. [CrossRef]
16. Hassan, M.; Narupiti, S.; Bhuiyan, T.; Hassan, M.M. Leveraging Large Language Models to Investigate the Decoy Effect in Route Choice Behavior. In Proceedings of the 2025 International Conference on Sustainable Technology and Engineering (i-COSTE), Alexandria, VA, USA, 2–4 December 2025; pp. 1–6. [CrossRef]
17. Cherep, M.; Maes, P.; Singh, N. AI Agents Are Sensitive to Nudges. *Proc. Natl. Acad. Sci. USA* **2026**, *123*, e2537030123. [CrossRef] [PubMed]
18. Jia, S.; Ma, C.; Chi, O.H.; Fan, A. A Socio-Technical Exploration of Bias in Generative AI Travel Planners. *Tour. Manag.* **2026**, *116*, 105442. [CrossRef]
19. Andreev, H.; Kosmas, P.; Livieratos, A.D.; Theocharous, A.; Zopiatis, A. Destination (Un)Known: Auditing Bias and Fairness in LLM-Based Travel Recommendations. *AI* **2025**, *6*, 236. [CrossRef]
20. Tversky, A.; Kahneman, D. The Framing of Decisions and the Psychology of Choice. *Science* **1981**, *211*, 453–458. [CrossRef] [PubMed]
21. Levin, I.P.; Schneider, S.L.; Gaeth, G.J. All Frames Are Not Created Equal: A Typology and Critical Analysis of Framing Effects. *Organ. Behav. Hum. Decis. Process.* **1998**, *76*, 149–188. [CrossRef] [PubMed]

22. Li, H.; Wang, L.; Ou, Y.; Zhang, L. How Does Goal Framing Effect Influence Household Low-Carbon Behavior: The Roles of Environmental Self-Efficacy and Global–Local Identity. *Front. Environ. Sci.* **2023**, *10*, 1035118. [CrossRef]
23. Brehm, J.W. *A Theory of Psychological Reactance*; Academic Press: New York, NY, USA, 1966.
24. Li, Z.; Shi, J. Message Effects on Psychological Reactance: Meta-Analyses. *Hum. Commun. Res.* **2026**, *52*, 38–52. [CrossRef]
25. Tversky, A.; Kahneman, D. Judgment under Uncertainty: Heuristics and Biases. *Science* **1974**, *185*, 1124–1131. [CrossRef] [PubMed]
26. Ariely, D.; Loewenstein, G.; Prelec, D. "Coherent Arbitrariness": Stable Demand Curves without Stable Preferences. *Q. J. Econ.* **2003**, *118*, 73–106. [CrossRef]
27. Suri, G.; Slater, L.R.; Ziaee, A.; Nguyen, M. Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. *arXiv* **2023**, arXiv:2305.04400.
28. Lou, J.; Sun, Y. Anchoring Bias in Large Language Models: An Experimental Study. *J. Comput. Soc. Sci.* **2026**, *9*, 11. [CrossRef]
29. Huber, J.; Payne, J.W.; Puto, C. Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *J. Consum. Res.* **1982**, *9*, 90–98. [CrossRef]
30. Chvirova, D.; Keller, R.; Stöbe, L.; Wolf, L. Beyond Choice Architecture: Toward Understanding the Role of the Decoy Effect in Sustainable Tourism Recommender Systems. *Inf. Technol. Tour.* **2026**, *28*, 5. [CrossRef]
31. Horton, J.J.; Filippas, A.; Manning, B.S. *Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?* NBER Working Paper No. 31122; National Bureau of Economic Research: Cambridge, MA, USA, 2023. [CrossRef]
32. Aher, G.V.; Arriaga, R.I.; Kalai, A.T. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *Proceedings of the 40th International Conference on Machine Learning (ICML 2023)*; PMLR: Cambridge, MA, USA, 2023; pp. 337–371. Available online: https://proceedings.mlr.press/v202/aher23a.html (accessed on 7 October 2026).
33. Argyle, L.P.; Busby, E.C.; Fulda, N.; Gubler, J.R.; Rytting, C.; Wingate, D. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Anal.* **2023**, *31*, 337–351. [CrossRef]
34. Brand, J.; Israeli, A.; Ngwe, D. *Using GPT for Market Research*; MSI Working Paper No. 23-131; Marketing Science Institute: New York, NY, USA, 2023; Available online: https://www.msi.org/working-paper/using-gpt-for-market-research (accessed on 7 October 2026).
35. Ross, J.; Kim, Y.; Lo, A.W. LLM Economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. *arXiv* **2024**, arXiv:2408.02784.
36. Zhuo, J.; Zhang, S.; Fang, X.; Duan, H.; Lin, D.; Chen, K. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1950–1976. [CrossRef]
37. Sharma, M.; Tong, M.; Korbak, T.; Duvenaud, D.; Askell, A.; Bowman, S.R.; Cheng, N.; Durmus, E.; Hatfield-Dodds, Z.; Johnston, S.R.; et al. Towards Understanding Sycophancy in Language Models. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024; Available online: https://iclr.cc/virtual/2024/poster/17593 (accessed on 14 October 2026).
38. Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; Yang, D. Can Large Language Models Transform Computational Social Science? *Comput. Linguist.* **2024**, *50*, 237–291. [CrossRef]
39. Aguinis, H.; Bradley, K.J. Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies. *Organ. Res. Methods* **2014**, *17*, 351–371. [CrossRef]
40. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1988.
41. Faul, F.; Erdfelder, E.; Lang, A.-G.; Buchner, A. G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. *Behav. Res. Methods* **2007**, *39*, 175–191. [CrossRef] [PubMed]
42. Lakens, D. Sample Size Justification. *Collabra Psychol.* **2022**, *8*, 33267. [CrossRef]
43. Lior, G.; Habba, E.; Levy, S.; Caciularu, A.; Stanovsky, G. ReliableEval: A Recipe for Stochastic LLM Evaluation via Method of Moments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 11146–11153. [CrossRef]
44. Pasamitros, N.; Krikoni, V. Double Bordering of the Nicosia Historic Centre: Contested Heritage, Divisions, and Attachments of the Walled City. *HAPSc Policy Briefs Ser.* **2024**, *5*, 93–101. Available online: https://ejournals.epublishing.ekt.gr/index.php/hapscpbs/article/view/40786 (accessed on 14 October 2026). [CrossRef]
45. Musement. Hagia Sophia, Blue Mosque & Grand Bazaar Guided Small-Group Tour. Available online: https://www.musement.com/us/istanbul/hagia-sophia-blue-mosque-grand-bazaar-guided-small-group-tour-378164/ (accessed on 7 October 2026).
46. AAA TripCanvas. Guided Hagia Sofia, Blue Mosque, Basilica Cistern, Grand Bazaar. Available online: https://ww1.aaa.com/tripcanvas/thing-to-do/guided-hagia-sofia-blue-mosque-basilica-cistern-grand-bazaar-68538P7 (accessed on 7 October 2026).
47. Bloomberg, H.T. Dolar ne Kadar Oldu? 2 Ekim 2026 Dolar Kurunda Son Durum ne? Available online: https://www.bloomberght.com/dolar-ne-kadar-oldu-2-ekim-2026-dolar-kurunda-son-durum-ne-pkh2-3790079 (accessed on 7 October 2026).

48. AAA TripCanvas. 4 Hours Istanbul Old City Walking & Shopping Tour. Available online: https://ww2.aaa.com/tripcanvas/thing-to-do/4-hours-istanbul-old-city-walking-and-shopping-tour-192301P9 (accessed on 7 October 2026).
49. Şerif Yenen Tours. Short Walking Tour of Highlights in Istanbul. Available online: https://serifyenen.com/tours/highlights-in-istanbul-with-serif-yenen-guides/ (accessed on 7 October 2026).
50. Newcombe, R.G. Interval Estimation for the Difference between Independent Proportions: Comparison of Eleven Methods. *Stat. Med.* **1998**, *17*, 873–890. [CrossRef]
51. Holm, S. A Simple Sequentially Rejective Multiple Test Procedure. *Scand. J. Stat.* **1979**, *6*, 65–70.
52. Jacowitz, K.E.; Kahneman, D. Measures of Anchoring in Estimation Tasks. *Pers. Soc. Psychol. Bull.* **1995**, *21*, 1161–1166. [CrossRef]
53. González-Cantera, T.; Bonacchi, C. ChatGPT's Interpretation of Contested Memoryscapes Favours the Voice of Current Governments, Capitalism and the Far-Right. *Int. J. Herit. Stud.* **2025**, *31*, 1508–1530. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.