Artificial Intelligence for Transportation 3–4 (2025) 100034

Contents lists available at ScienceDirect

Artificial Intelligence for Transportation

journal homepage: www.elsevier.com/locate/ait

Can large language models mimic airline passenger preferences?

Augusto Voltes-Dorta [a,*], Pere Suau-Sanchez [b,c]

[a] University of Edinburgh Business School, Edinburgh EH8 9JS, United Kingdom
[b] Faculty of Economics and Business, Digital Transformation and Governance Research Centre, Universitat Oberta de Catalunya, Rambla del Poblenou 156, 08018 Barcelona, Spain
[c] Faculty of Engineering and Applied Sciences, Cranfield University, Martell House, University Way, Cranfield MK43 0TR, United Kingdom

ARTICLE INFO

Keywords:
Airline
Discrete choice
Large language model
Knowledge audit
Mixed logit

ABSTRACT

Large Language Models (LLMs) are being used in the air travel sector to simulate passenger behaviour. While commercial LLMs provide ready-made solutions, concerns over reliability limit widespread use. This paper leverages decades of discrete-choice research to audit 23 LLMs' responses to a flight-ticket choice experiment with zero-shot prompting. The results of our logit regressions indicate that LLMs can simulate highly rational preferences, correctly sign ticket attributes, differentiate between relevant and irrelevant factors, and deliver plausible willingness-to-pay estimates. However, LLMs might fall short in replicating nuanced demographic segmentation and show sensitivity to cultural bias in their training data. Their outputs are best used to inform early-stage modelling, with traditional market research remaining essential for validation.

1. Introduction

With the adoption of dynamic pricing in the airline industry, many firms have integrated Artificial Intelligence (AI) to enhance customer offerings (Dadoun et al., 2021). Amadeus, for example, uses AI explicitly to predict customer preferences (Amadeus, 2025). LATAM Airlines leverages Sabre's AI-powered Air Price IQ solution to optimise airfare pricing based on traveller behaviour and market competition (Sabre, 2024). Singapore Airlines similarly applies AI for tailored product promotions (Singapore Airlines, 2025). Together, these examples highlight the global airline industry's strategic deployment of AI to enhance customer satisfaction and revenue optimisation.

Large Language Models (LLMs) are central to this trend (Eigner & Händler, 2024). Their ability to process vast data and adapt to feedback enables them to replicate complex behaviours (Nguyen et al., 2024). Setting aside epistemological debates, AI now serves as a practical means of imitating human decision-making (Mei et al., 2024) by applying existing knowledge to routine tasks. Beyond language, LLMs demonstrate economic rationality across classic domains—often matching or exceeding human benchmarks—while showing less heterogeneity (Chen et al., 2023) and uneven but notable social-cognitive abilities (Strachan et al., 2024). However, recent stress tests caution that such human-like behaviour is not guaranteed across all tasks, underscoring the need for domain-specific evaluation (Gao et al., 2024) and persona conditioning (Kitadai et al., 2025).

Air travel, with its rich behavioural evidence base, provides an ideal testbed for assessing how closely LLMs align with human preferences. Indeed, a vast body of discrete choice research has established a detailed understanding of air travellers' preferences (Fukushi et al., 2024). Key drivers of flight choice include airfare, frequency, travel time, and service reliability, with business travellers valuing time and punctuality more highly (Hess & Adler, 2011). Comfort and environmental factors, such as seat pitch or CO_2 emissions, also influence demand (Balcombe et al., 2009; Crosby et al., 2024). Most studies rely on stated preference surveys analysed with random utility models—typically the Multinomial Logit (McFadden, 1974) and its extensions, including mixed and latent-class logit, which capture taste heterogeneity (Hess & Polak, 2005). A central outcome is travellers' willingness to pay (WTP) for key attributes. WTP for time savings, for instance, consistently differentiates business from leisure travellers (Warburg et al., 2006; Hess et al., 2011).

These openly available WTP datasets are likely included in the *Common Crawl* corpus used to train major commercial LLMs (Brown et al., 2020). As a result, models such as OpenAI's ChatGPT and Google's Gemini have absorbed domain-specific economic knowledge from the web, providing low-cost tools for organisations lacking resources to train bespoke systems. Yet, these off-the-shelf models raise key concerns. Their "black-box" nature limits transparency and accountability (Castelvecchi, 2016), while factual unreliability remains common (Zhou et al., 2024), partly due to the architectural constraints of transformer models (Nikolenko, 2025).

On the other hand, LLMs are being constantly improved, with intense competition, e.g. entry of DeepSeek in early 2025, driving rapid advances. However, other than the general KPIs that companies use to

* Corresponding author.
E-mail address: avoltes@becarios.ulpgc.es (A. Voltes-Dorta).

https://doi.org/10.1016/j.ait.2025.100034
Received 25 June 2025; Received in revised form 9 November 2025; Accepted 10 November 2025
3050-8606/© 2025 The Author(s). Published by Elsevier Inc. on behalf of Chinese Overseas Transportation Association. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

benchmark their models against their competitors, little is known about their task-specific performance. This raises a central question: can commercial LLMs, without additional tuning, reliably imitate airline passengers' decisions? If not yet perfect, are they at least showing meaningful progress towards that goal?

To explore this, we adopt a "knowledge audit" approach—comparing LLM predictions against published findings. While this builds on earlier work using psychological experiments to probe LLMs' cognitive abilities (Hagendorff et al., 2023), it is conceptually closer to studies such as Brand et al. (2024), which directly compared LLM and human responses in consumer surveys. In similar fashion, a growing body of economic research now treats LLMs as surrogate decision-makers in controlled experiments, testing behavioural hypotheses "in silico" before real-world application (Horton, 2023; Filippas et al., 2024; Ludwig et al., 2024). Overall, findings are mixed: LLMs can replicate human-like economic choices (Horton, 2023) but sometimes behave more rationally than humans (Chen et al., 2023), misrepresent certain demographic groups (Wang et al., 2025b), and exhibit cultural biases (Segerer, 2025).

In transport, LLM research remains limited, with most applications focusing on mode choice (Liu et al., 2024; Mo et al., 2023) or broader functions such as safety, automation, or customer engagement (Wandelt et al., 2024; Nie et al., 2025). Recent work has pushed towards behaviourally explicit agents and system-level realism. Guo et al. (2024) surveyed LLM-based multi-agent systems; Wang et al. (2025a) introduced an LLM-powered agent reproducing human-like route switching in congestion games; and Liu et al. (2025) integrated LLM travellers into a morning-commute simulation, showing realistic dynamics. At the modelling end, Sfeir et al. (2025) found that LLMs can generate behaviourally plausible model specifications—and sometimes estimate them—under structured prompting. Closest to our aims, Yan et al. (2025) audited GPT-4o's implicit WTP for travel time savings within a discrete-choice framework, finding human-like aggregate estimates and demographic sensitivity but muted contextual responsiveness. We extend this approach to air travel, a domain not yet examined despite its complex, highly attribute-driven choice sets.

This paper fills that gap by employing the same type of discrete-choice experiment (DCE) that originally revealed air passenger preferences. We designed a flight-ticket DCE reflecting the demographics of San Francisco International Airport (SFO) travellers. To account for potential cultural biases, we engaged 23 LLMs from OpenAI, Google, and DeepSeek—spanning 2024–2025 releases, small and large sizes, and both general and reasoning models—using zero-shot prompting throughout. Mixed and latent-class logit regressions yielded WTP estimates for ticket attributes and traveller segments. These serve as the basis for auditing LLM performance across five alignment dimensions: rationality, selective attention, discriminatory power, plausibility, and class variability. By simulating the full decision environment of airline travellers, we provide the first systematic evaluation of LLMs in air transport choice modelling. The results offer a practical benchmark for assessing whether commercial LLMs can emulate passenger preferences—particularly valuable for smaller airlines lacking the capacity to train bespoke AI systems.

The remainder of this paper is structured as follows: Section 2 reviews the literature on human–LLM alignment and air-travel DCE research; Section 3 details the modelling framework; Section 4 presents the results and discusses their main implications; and Section 5 concludes with key takeaways and future directions.

2. Literature review

2.1. Human-LLM alignment

Recent research on the alignment between LLMs and human choices has focused on several key dimensions, the first being *rationality*. LLM responses should conform to classical axioms, such as dominance and transitivity, and results so far are promising. Chen et al. (2023) evaluated GPT-4's choices and found near-perfect performance across rationality indices—including the Critical Cost Efficiency Index (CCEI), Houtman–Maks Index (HMI), and Minimum Cost Index (MCI)—with only minimal adjustments needed for full consistency. Similarly, Horton (2023) observed that GPT models display coherent, utility-maximising choices in standard economic games, while Yan et al. (2025) confirmed that LLMs make logically consistent trade-offs in travel-choice experiments. Collectively, these findings suggest LLMs can act as internally rational agents.

However, Hagendorff et al. (2023) caution that such models may exhibit hyper-rationality, behaving more consistently than humans and missing the heuristic, bounded aspects of real decision-making (Simon, 1955; Ke et al., 2025). Following Mukherjee and Chang (2024), bounded rationality in LLMs can be operationalised through instrumental simplification—evaluating how many attributes the model attends to when making decisions. In our framework, this equates to assessing *selective attention* in choice tasks, i.e., whether the LLM focuses on a subset of salient attributes rather than processing all available information.

Cultural biases from training data may further shape this selective attention. Kazemi et al. (2024) showed that LLMs' reflection of national values strongly correlates with the linguistic and digital resources in training, leading to higher error rates in low-resource contexts. As a result, GPT models align more with Western, individualist values, while models trained on Chinese corpora—such as DeepSeek—show more collectivist orientations (Tao et al., 2024; Segerer, 2025; Lu et al., 2025). In air-travel contexts, where regional preferences for time, comfort, or environmental factors differ, these cultural biases may guide which attributes an LLM prioritises when simulating travellers.

Beyond the number of attributes attended, a well-aligned LLM should also prioritise important attributes and down-weight minor ones—displaying *discriminatory power*. Brand et al. (2024) found GPT-3.5 produced realistic WTPs and preserved the relative importance of features, reflecting human-like recognition of attribute relevance.

Another dimension, *plausibility*, concerns whether the model's stated trade-offs quantitatively resemble those of rational human decision-makers. Even without fine-tuning, LLMs have been shown to produce human-like WTP values: Brand et al. (2024) found GPT-3.5 generated realistic, comparable figures to human surveys, and Yan et al. (2025) reported GPT-4's value of travel time (VOT) was statistically indistinguishable from human travellers. Plausibility can thus be tested by verifying that estimated WTP coefficients have expected economic signs and magnitudes within human benchmarks.

Finally, in air-travel, *class variability* is crucial, as preferences vary by income, purpose, and demographics—e.g., business passengers show higher values of time than leisure travellers. Yan et al. (2025) found GPT-4o replicated these patterns, with higher estimated VOTs for higher incomes and business trips, though variability was dampened—showing LLMs can capture contextual trends but understate real heterogeneity.

In sum, the emerging literature across psychology, economics, and transport science supports evaluating LLM–human alignment along five dimensions: rationality, selective attention, discriminatory power, plausibility, and class variability. Each captures a distinct aspect of human-like reasoning, together forming a comprehensive framework as applied in this study. In airline travel, these metrics are reinforced by robust DCE research, providing validated benchmarks for assessing alignment with real passenger preferences.

2.2. DCE research in air travel

In the air travel domain, travellers' preferences are well understood thanks to decades of discrete-choice research (Fukushi et al., 2024). Studies have analysed diverse decision contexts, including ticket or itinerary choice (Adler et al., 2005), airport selection in multi-airport regions (Windle & Dresner, 1995), combined airport-airline choice (Hess et al., 2011), and ground access mode to airports (Bergantino et al., 2020). Across this literature, several attributes consis-

tently drive air travel decisions. Airfare is often dominant, especially for cost-sensitive leisure travellers (Ashford & Benchemam, 1987). Lower fares substantially increase an option's appeal, making price a near-universal variable in airline choice models. Beyond cost, schedule convenience is crucial. Pels et al. (2001) found that greater flight frequency enhances attractiveness by offering flexibility, while travel time strongly influences business travellers, who value time savings and frequent service most (Hess & Adler, 2011). Non-stop flights are also far more appealing than those with layovers, which add delay and risk (Adler et al., 2005).

Service quality and reliability also matter. Travellers value punctuality and schedule reliability, as delays or missed connections impose high inconvenience (Hess et al., 2011). Flights or airlines with stronger on-time records are therefore preferred. Comfort and amenities influence choices too: Balcombe et al. (2009) found significant WTP for wider seats, greater pitch, in-flight meals, and entertainment, showing that many passengers weigh experience as well as price.

Ancillary fees—especially for checked baggage—have become key determinants of choice. Scotti and Dresner (2015) show that baggage fees reduce demand but allow fare adjustments to balance revenue and market share. Cho and Dresner (2018) find baggage fee waivers particularly sway leisure and long-haul travellers with higher baggage needs. Thus, baggage policies affect choices independently of fare and must be modelled explicitly. Frequent-flyer programs (FFPs) similarly shape behaviour: Hartmann and Viard (2008) demonstrate that accumulated miles and elite status can justify higher fares, as travellers rationally trade price for loyalty rewards.

Environmental considerations now appear regularly in passenger surveys. Some travellers and willing to pay premiums for lower-emission flights (Crosby et al., 2024), though WTP for CO_2 reductions varies widely. Wendt et al. (2024) show that framing offsets "per flight" rather than "per tonne of CO_2" increases WTP, while higher income and education correlate with stronger environmental valuations—underscoring preference heterogeneity.

Stated preference (SP) surveys remain the main method for collecting air travel choice data (Fukushi et al., 2024). Travellers select among hypothetical itineraries with varying attributes (fare, time, etc.), allowing inference of trade-offs. Analytically, most studies employ random utility maximisation frameworks. The Multinomial Logit (MNL) model (McFadden, 1974) has long been foundational, with subsequent extensions addressing its limitations. Nested Logit models capture correlated alternatives such as flights from the same airport (Hess et al., 2011), while conditional logit models incorporate alternative-specific attributes. To represent taste heterogeneity, mixed logit and latent class logit models allow parameters to vary across individuals or segments (Hess & Polak, 2005; Crosby et al., 2024).

A key outcome of these models is estimation of WTP for service attributes. WTPs for travel time savings, for example, differ markedly by trip purpose: leisure travellers' mean values typically range from $19.64 to $36.00 per hour, while for business travellers, mean values span $36.00 to $70.20 per hour (Warburg et al., 2006; Garrow et al., 2007; Hess & Adler, 2011; Seelhorst & Liu, 2015; Hess et al., 2011). Decades of such WTP estimates are now available through open-access repositories, enabling their widespread use in policy evaluation and transport planning—as well as in training the very LLMs we aim to audit.

2.3. Research gap and contribution

Despite rapid advances in modelling human-like decision-making, current evidence on LLM-human alignment remains fragmented across psychology, economics, and transport. Studies have shown that LLMs can act as rational agents (Chen et al., 2023), approximate human trade-offs (Yan et al., 2025), and exhibit boundedly rational heuristics (Mukherjee & Chang, 2024). However, systematic frameworks for assessing how they align with human behaviour in applied domains remain limited. At the same time, decades of discrete-choice research have precisely mapped air travellers' preferences, offering an exceptionally rich behavioural benchmark. We extend this emerging alignment literature to air travel—a domain not yet studied despite its complex, highly attribute-driven choice sets. By introducing an integrated five-dimensional framework—rationality, selective attention, discriminatory power, plausibility, and class variability—we evaluate LLM–human alignment empirically. This approach contributes both a replicable audit method and new insights into the behavioural fidelity of modern LLMs in realistic, multidimensional decision environments.

3. Data and methodology

Our method proceeds in three stages. First, we combined a stated-preference DCE with LLM-based synthetic respondents, instructing each model to act as a traveller completing 17 flight-choice tasks of increasing attribute complexity. Demographically grounded personas ensured realistic representation, while a fixed zero-shot prompt and zero-temperature setting maintained response consistency. Second, the simulated data were analysed using mixed and latent class logit models to capture preference heterogeneity. WTP estimates were derived from model coefficients to quantify trade-offs among service attributes. Finally, we audited the LLM-generated choices and WTPs for human alignment across five dimensions—rationality, selective attention, discriminatory power, plausibility, and class variability.

3.1. Modelling approach

Our stated-preference DCE is underpinned by random utility theory (McFadden, 1974). For individual i choosing among J flight ticket alternatives in choice task t, the utility U_{ijt} of alternative j is defined as:

$$U_{ijt} = \beta' x_{ijt} + \varepsilon_{ijt}, \tag{1}$$

where ε denotes the random error, x includes all observable price and non-price attributes, and β are the associated taste parameters.

In our study, the mixed logit serves as the main estimation framework, while a latent class specification is used illustratively to highlight class variability. In the mixed logit framework, the taste parameters are treated as random variables, capturing preference heterogeneity across individuals:

$$\beta_i = \beta + \eta_i, \tag{2}$$

where β is now the population mean, and η_i captures individual-level deviations. The probability of individual i choosing alternative j in task t is then:

$$Prob_{ijt} = \int \frac{\exp(\beta' x_{ijt})}{\sum_{l=1}^{J}(\beta' x_{ilt})} f(\beta_i \mid \theta)\, d\beta_i, \tag{3}$$

where $f(\beta_i|\theta)$ is the density of the random parameters, characterised by distributional parameters θ.

For those parameters exhibiting significant variance in the mixed logit, we also explored a finite mixture (latent class) specification (Kim & Mokhtarian, 2023). In this hybrid approach, only the parameters with significant variance are allowed to take on class-specific values β^k, while the others remain fixed. Conditional on class k, the choice probability is:

$$Prob_{ijt|k} = \frac{\exp\left(\beta' x_{ijt} + \beta^{k'} x^k_{ijt}\right)}{\sum_{l=1}^{J}(\beta' x_{ilt} + \beta^{k'} x^k_{ilt})} \tag{4}$$

The unconditional probability is then given by a weighted average across classes:

$$Prob_{ijt} = \Sigma_{k=1}^{K} \pi_{ik} Prob_{ijt|k}, \tag{5}$$

where π_{ik} is the probability that individual i belongs to class k. These membership probabilities are modelled through a multinomial logit

function of observed socio-demographic covariates ω_i:

$$\pi_{ik} = \frac{\exp\left(\gamma_k' \omega_i\right)}{\sum_{m=1}^{K} \exp\left(\gamma_m' \omega_i\right)}. \quad (6)$$

This structure enables the model to disentangle subgroups in the population with internally homogeneous but mutually heterogeneous preference structures.

As usual, the WTP for a unit change in non-price attribute r can be computed as:

$$WTP_r = -\frac{\beta_r}{\beta_{price}} \quad (7)$$

The parameters will be estimated via maximum likelihood assisted by the expectation-maximisation algorithm (Bhat, 1997; Yoo, 2020).

3.2. Experimental design

Our synthetic passengers, all travellers between San Francisco and New York, faced 17 choice tasks with four alternative tickets each, one of which a baseline option—representing a no-frills ticket for comparison. A random design was chosen for its simplicity and ability to approach an efficient design for large sample sizes (Bliemer & Rose, 2024). Unlike human respondents, LLMs are not subject to fatigue or declining attention, which typically limit the number of attributes in stated-preference designs. This enabled testing more complex configurations and assessing whether LLM-derived WTPs remain stable as dimensionality increases—an issue long noted in the DCE literature (Hensher, 2006a). Accordingly, we tested five designs comprising four, six, nine, eleven, and thirteen attributes, respectively, using the latter for the final analysis.

Design 1 includes the four basic attributes, all of which have been examined in the wider DCE literature on air travellers' preferences (Adler et al., 2005; Garrow et al., 2007; Hess et al., 2011; Crosby et al., 2024):

- Price: Continuous variable between US$150[baseline] and US$800. The price range was chosen to reflect typical fares across booking classes on the New York–San Francisco route.
- Travel Time (3 levels): "5.5 h" [baseline], "8 h (1 stop)", "11 h (2 stops)".
- Legroom (2 levels): "31 inches" [baseline], "33 inches". We chose 33 inches as a modest improvement over the baseline option, reflecting an extra-legroom seat—a tangible upgrade that flyers would notice (Balcombe et al., 2009).
- CO2 Emissions (3 levels): "90″ [baseline], "120″, "150 g per pax-km". These levels reflect realistic operational performance: the baseline of 90 g CO_2 per passenger-kilometre was reported by ICCT (2020), while values up to 150 g fall within the range observed across aircraft types (d-carbonize, 2025). Expressing emissions in grams per passenger-kilometre decouples fuel burn from distance, allowing assessment of whether travellers prefer more efficient aircraft independent of trip length.

Design 2 adds:

- On-Time Performance – OTP (3 levels): "60 %" [baseline], "75 %", "90 %". The 75 % level approximates the North American industry average, while 60 % represents subpar performance and 90 % indicates exceptional reliability (Cirium, 2024).
- Baggage (2 levels): "No Free Checked Bag" [baseline], "One Free Checked Bag".

Design 3 adds more attributes related to cabin experience:

- Wifi (2 levels): "Not Available" [baseline], "Available". The inclusion of Wi-Fi reflects its integration into ticket search platforms: Google Flights, for instance, routinely displays Wi-Fi availability alongside core attributes, underscoring its role as a driver of ticket choice rather than a secondary amenity (Crosby et al., 2024).
- In-flight Entertainment – IFE (2 levels): "Not Available" [baseline], "Available".
- Meals (2 levels): "Not Included" [baseline], "Included".

Design 4 adds:

- Frequent Flyer Benefits (2 levels): "Not Available" [baseline], "Available". This binary framing distinguishes flights on airlines outside a traveller's loyalty network from those that credit miles or offer status benefits—factors known to foster strong loyalty even when fares are higher (Hartmann & Viard, 2008).
- Seat Location (2 levels): "Port side" [baseline], "Starboard side". This attribute is largely inconsequential to travel experience—unlike window vs. aisle, port vs. starboard is essentially arbitrary. We included it to test rationality: a consistent respondent (or LLM) should ignore it when making choices. In SP experiments, adding such irrelevant attributes can help verify focus on meaningful factors, as evidence shows their presence does not necessarily bias outcomes (Hess, 2014).

Design 5 finally adds two novel attributes:

- Noise-cancelling headphones (2 levels): "Not Available" [baseline], "Available". Some airlines (e.g., Emirates) provide noise-cancelling headphones in first class, as a quieter, more immersive cabin is part of a luxury service. Although no study has measured WTP for this attribute specifically, it falls within the broader comfort-amenity category. We included it to test the model's sensitivity: if synthetic respondents emulate human preferences, they should value it at least non-negatively, since reduced noise can only enhance the flight experience.
- Children (2 levels): "Children allowed" [baseline], "Child-free section". The latter refers to a cabin zone where young children are not allowed, catering to adults seeking a quieter environment. Corendon Airlines, for example, has charged €45–100 for such seating (Dailymail.co.uk, 2023), suggesting that some passengers—such as solo business travellers or child-free adults—are willing to pay for it. We included this attribute to assess demographic segmentation of WTP, as it is one where differences between traveller groups are likely to be pronounced.

The baseline option, which is present in all choice tasks, always shows the same baseline levels for the featured attributes.

Table 1
Demographics of synthetic respondents.

Gender	%	Income	%
		Under $50,000	18.13
Male	46.75	$50,000-$100,000	30.19
Female	52.95	$100,001-$150,000	19.71
Non-binary	0.30	Over $150,000	31.96
Age	**%**	**Home Location**	**%**
		US	80.07
18–24	11.15	International	19.93
25–34	24.97	**Trip Purpose**	**%**
35–44	20.01	Business	29.22
45–54	16.57	Holiday	41.17
55–64	15.76	Visiting Friends or Relatives (VFR)	22.26
65+	11.54	Other	7.35

Source: SFO (2019).

3.3. Survey administration

We simulated 150 passengers for each design. To make our synthetic sample representative of airline travellers, we generated random demographic profiles based on the 2018 SFO Airport Customer Survey

Table 2
Surveyed LLMs.

Model Name	*Code*	*Provider*	*Release Date*	*Size (assumed)*	*Used as*
DeepSeek-V3.1-Terminus (Thinking)	dsr3	DeepSeek	21/08/2025	Large	Reasoner
DeepSeek-R1	dsr	DeepSeek	25/03/2025	Large	Reasoner
DeepSeek-V3.1-Terminus	dsc3	DeepSeek	21/08/2025	Large	General
DeepSeek-V3-0324	dsc	DeepSeek	20/01/2025	Large	General
gemini-2.5-flash-preview-05-20	g25f	Google	20/05/2025	Large	Reasoner
gemini-2.0-flash	g20f	Google	05/02/2025	Small	General
gemini-1.5-flash	g15f	Google	24/05/2024	Small	General
gemini-1.5-flash-8b	g15f8	Google	03/10/2024	Small	General
gpt-5	5x	OpenAI	07/08/2025	Large	Reasoner
gpt-5-mini	5m	OpenAI	07/08/2025	Small	Reasoner
gpt-5-nano	5n	OpenAI	07/08/2025	Small	Reasoner
gpt-4.5-preview-2025-02-27	45x	OpenAI	27/02/2025	Large	General
o1-2024-12-17	o1x	OpenAI	17/12/2024	Large	Reasoner
o3-2025-04-16	o3x	OpenAI	16/04/2025	Large	Reasoner
gpt-4o-2024-08-06	4o	OpenAI	06/08/2024	Large	General
gpt-4.1-2025-04-14	41x	OpenAI	14/04/2025	Large	General
o4-mini-2025-04-16	o4m	OpenAI	16/04/2025	Small	Reasoner
o3-mini-2025-01-31	o3m	OpenAI	31/01/2025	Small	Reasoner
o1-mini-2024-09-12	o1m	OpenAI	12/09/2024	Small	Reasoner
gpt-4.1-mini-2025-04-14	41m	OpenAI	14/04/2025	Small	General
gpt-3.5-turbo-0125	35t	OpenAI	25/01/2024	Small	General
gpt-4o-mini-2024-07-18	4om	OpenAI	18/07/2024	Small	General
gpt-4.1-nano-2025-04-14	41n	OpenAI	14/04/2025	Small	General

Sources: OpenAI, Gemini, DeepSeek.

(SFO, 2018). The probabilities are shown in Table 1. This sociodemographic prompting has been shown to improve LLM performance in subjective tasks (Beck et al., 2024).

Once generated, the demographic profile is passed to the LLM with the following prompt (see Appendix A): *"You are a traveller at San Francisco International Airport (SFO). Your persona has been pre-generated based on demographic probabilities. Answer the following questions about a San Francisco - New York flight based on this persona. Your preferences regarding the travel attributes must remain constant through the survey. When making choices, apply the principle of bounded rationality: use simplified decision-making heuristics, focus on key attributes, and aim for satisficing rather than optimising outcomes."*

If the passenger is travelling for business purposes, an additional prompt about the trip being paid by the employer is added.

These "zero-shot" prompts do not provide information about traveller preferences or any choice examples, forcing the LLM to rely solely on its pre-trained knowledge (Xian et al., 2019). To ensure it always returns its most likely response based on that knowledge, we set the temperature parameter to zero. As Chen et al. (2023) noted, higher temperature settings may yield more invalid choice outputs, even if they do not fundamentally change the model's underlying rationality.

The synthetic passenger is first asked a travel frequency question: "*How many times have you flown out of SFO in the past 12 months?*" and then the 17 choice tasks. Besides the chosen alternative, the LLM is requested to provide choice probabilities for each option. All responses were saved and carried forward so that subsequent prompts included the full survey history. We used the OpenAI API within a Python environment rather than the web interface, and because the API does not automatically retain conversational context, memory was maintained by recursively appending prior questions and answers. This procedure promotes internal consistency across the LLM's responses—although it cannot guarantee it.

Table 2 summarises the 23 large language models (LLMs) included in our simulations, spanning three major providers: OpenAI, Google, and DeepSeek. The models vary in release date (from September 2024 to August 2025), reflecting rapid iteration in the field. Model size, categorised as large or small, broadly corresponds to parameter scale and performance. Larger models tend to exhibit stronger reasoning abilities and more stable behaviour in complex tasks but are also more resource-intensive to deploy. Since parameter scales are often not published, we used API token prices as a guideline.

Models were also categorised by functional role: either general or reasoner. General models are optimised for efficiency and broad coverage, typically used in applications such as chatbots or productivity tools. In contrast, reasoner models are designed for more demanding tasks requiring multi-step logic, inference, or contextual trade-offs—making them particularly relevant for simulating choice behaviour. Our sample includes both types: for instance, OpenAI's o3 and DeepSeek's R1 are reasoners, while models like gpt-4o and gemini-1.5-flash represent general-purpose systems. This functional diversity allows us to evaluate how reasoning capabilities affect the accuracy of our WTPs. Including a mix of small and large models also enables comparisons across computational cost and performance trade-offs, relevant for practical deployment in consumer behaviour modelling.

3.4. Direct elicitation

For all models above, we complemented our DCE with an additional run in which the LLM-subjects were directly asked to state their WTPs for the flight attributes. This direct elicitation, conducted with zero-shot prompting and zero temperature, enabled us to compare the LLM's inferred versus declared preferences. As shown by Manning et al. (2024), LLMs often possess implicit knowledge that emerges only through structured reasoning tasks, not through direct questioning—particularly when numerical trade-offs are involved. While direct elicitation tests what the model is willing to report, DCE inference reveals what the model can simulate, offering a deeper view of its internal representation of consumer preferences.

3.5. Benchmarks

We audited the LLMs' responses across five dimensions—rationality, selective attention, discriminatory power, plausibility, and class variability. The eleven corresponding benchmarks are summarised in Table 3. While the first three dimensions primarily assess internal behavioural coherence and processing style—revealing how different models reason, simplify, and prioritise information—plausibility and class variability directly connect LLM outputs to human benchmarks, providing the strongest evidence of alignment between model-derived and empirically observed traveller preferences.

First, we tested *rationality* by expanding our DCE with three additional modules: a repeated question (stability check), a task with dom-

Table 3
Key performance indicators.

Dimension	*Benchmark*	*Definition / Formula*
Rationality	Dominance	Share of dominant choices (%)
	Transitivity	Share of transitive decisions among all AB, BC, AC tasks (%)
	Stability	Share of consistent choices across repeated tasks (%)
	Minimum Cost Index-probability	Smallest choice probability adjustment to restore full transitivity
Selective Attention	Attribute Importance (%)	β_attr × sd_attr / Σ(β × sd); indicates each attribute's contribution (%)
	HHI of Attribute Importance	HHI = Σ(attribute share^2); a higher value indicates more focused attention
Discriminatory Power	Discriminatory Power	Low attribute share of "Seat Location"
Plausibility	Expected Signs	Share of WTPs with expected signs, excl. Children and Seat Location (%)
	Plausibility Score	WTPs within 95% CI of past studies / total reference WTPs (%)
	Plausibility Gains	Plausibility_DCE – Plausibility_Direct Elicitation (%)
Class Variability	Class Variability	Max % Travel Time WTP difference across demographic groups (%)

inated alternatives (dominance check), and a triadic set of ticket comparisons (A vs. B, B vs. C, and A vs. C) for transitivity. The reported KPIs are the shares of stable, dominant, and transitive responses across each LLM's choices. The latter can be viewed as a local analogue of the Houtman–Maks Index (HMI; Houtman & Maks, 1985), which measures the fraction of choices that can be rationalised without internal contradiction. In our triadic design, this means checking whether preferences over three options are consistent with one another; any inconsistency would manifest as a preference cycle (e.g., A preferred to B, B to C, but C to A). As a fourth rationality test, we developed a probability-based version of the Minimum Cost Index (MCI; Dean & Martin, 2016). Our MCI_{prob} index leverages the fact that the LLMs provided not only a discrete choice but also a probability distribution over the available alternatives. This allowed us to calculate, for each triad, the minimum probability mass that would need to be reassigned across options to restore transitivity (See the formulation in Appendix B). Averaging MCI_{prob} across respondents then provides a measure of the severity of rationality violations, capturing the distance from rational choice rather than treating all violations equally.

Secondly, prior research cautions that LLMs may behave more rationally than humans (Hagendorff et al., 2023). To address this, we examined the distribution of attribute importance as a proxy for *selective attention* under our bounded-rationality prompt. Human decision-makers typically use satisficing or adaptive strategies rather than strictly optimal ones, reflecting cognitive limits (Simon, 1955). Thus, if LLMs allocate attention unevenly—focusing on a few salient attributes—this would signal boundedly rational, human-like processing rather than hyper-rational optimisation.

A pilot run (Appendix C) strongly suggests that LLMs exhibit behaviour consistent with bounded rationality. Both GPT-4o-mini, a smaller general-purpose model, and o3, a large reasoning model, showed a consistent decline in WTP values as the number of ticket attributes increased, stabilising beyond nine attributes. This parallel trend suggests that even advanced reasoning models simplify decision processes as cognitive load rises, mirroring the satisficing behaviour observed in human DCEs. Hensher (2006a) similarly found lower mean WTPs under more complex designs (wider attribute ranges and more levels per attribute), warning that neglecting attribute-processing rules risks undervaluing WTPs. Hensher (2006b) also showed that mean WTPs shift systematically with the number of attributes and alternatives, while Hensher (2006c) demonstrated that respondents manage information load by applying relevancy-based heuristics and omitting attributes, reshaping WTP distributions. Rose et al. (2004) further noted that such exclusion may reflect either genuine irrelevance or coping behaviour; in our case, the latter is more likely, since the LLM was explicitly prompted to apply simplified heuristics and prioritise key attributes.

To measure each attribute's influence on choices, we standardise its utility coefficient (β) and express it as a share of the total "utility movement" across all attributes. These relative-importance shares reveal how much weight each attribute carries in a given model. To gauge *selective attention*, we then compute a Herfindahl–Hirschman Index (HHI) over these shares: a high HHI indicates that attention is concentrated on a few key features, whereas a low HHI suggests a more evenly distributed focus across many attributes.

Thirdly, the importance shares also serve to test the LLM's *discriminatory power* between relevant and irrelevant features. In particular, we expect the 'Seat Location' attribute to carry a negligible weight in the LLM's choices.

Fourthly, we check whether the estimated WTPs are *plausible*. At the very minimum, they should have the *expected signs*—all positive except for 'Travel Time' and 'CO2 emissions', including the 'Noise-cancelling headphones' attribute that has not been researched before. The ambiguous 'child-free' and 'seat location' attributes will be excluded from this test. Incidentally, the negative sign of the price coefficients that enables correctly signed WTPs also provides a test of downward-sloping demand—a fundamental rationality condition whereby demand for a good decreases as its price increases (Chen et al., 2023).

We also check if our segment-specific, mean WTP estimates fall within the 95 % confidence intervals from previous DCE studies on US air traveller preferences (Pels et al., 2003; Adler et al., 2005; Warburg et al., 2006; Garrow et al., 2007; Hess & Adler, 2011; Hess et al., 2011; Seelhorst & Liu, 2015; Crosby et al., 2024). The confidence intervals were calculated as mean ± 1.96 × standard deviation, using the original summary statistics reported in each study. When an attribute appeared in multiple studies, we treated each estimate as a separate reference rather than pooling them, reflecting differences in sample sizes, estimation methods, and DCE framings across studies. Reporting them separately has the advantage of covering a broader range of WTP estimates observed in human respondents, thus providing a more comprehensive benchmark against which to compare the LLM outputs.

The reference values, adjusted for inflation to 2025, are shown in Table 4. In total, five different attributes are covered (Travel Time, OTP, IFE, Legroom, and Wifi), as well as business and leisure segments—the latter bundling holiday and VFR travellers. The plausibility ratio, in percentage, shows the number of confidence-interval matches between the LLM's implicit WTPs and the 17 reference items in Table 4. Following Manning et al. (2024), we also discuss the *plausibility gains* between the DCE approach and the directly-elicited WTPs.

Finally, we examined whether the synthetic WTPs exhibit *class variability*. Results from the pilot run with GPT-4o-mini suggest they do (see latent-class regression in Appendix C). Consistent with prior work (Hess & Adler, 2011), LLM-simulated business travellers displayed higher WTPs for service improvements than holiday travellers. Within each segment, we also identified two behavioural classes—one relatively price-sensitive and another less so—echoing patterns reported by Crosby et al. (2024). Class heterogeneity also appeared for the child-free attribute, significant only among holiday travellers and manifesting as a disutility, likely reflecting the trade-offs faced by families when travelling. These results suggest that synthetic WTPs capture not only mean preferences but also subgroup variation. WTPs also rose with travel frequency, aligning with evidence that frequent fliers are less price-sensitive (Adler et al., 2005), while female and higher-income travellers

Table 4
Reference WTPs from the literature (inflation-adjusted).

Article	*Attribute*	*Segment*	*mean*	*s.d.*	*CI 95 %-low*	*CI 95 %-high*
Crosby et al. (2024)	IFE	Leisure	48.27	19.59	9.87	86.67
Crosby et al. (2024)	Legroom (inch)	Leisure	120.59	39.63	42.91	198.27
Crosby et al. (2024)	Legroom (inch)	Business	176.23	20.46	136.13	216.33
Warburg et al. (2006)	OTP (1 %)	Business	1.49	3.14	−4.66	7.63
Hess and Adler (2011)	OTP (1 %)	Leisure	0.87	0.06	0.76	0.98
Hess and Adler (2011)	OTP (1 %)	Business	1.49	0.36	0.79	2.19
Hess et al. (2011)	OTP (1 %)	All	2.94	0.47	2.03	3.86
Warburg et al. (2006)	Time (h)	Business	59.40	51.15	−40.85	159.65
Garrow et al. (2007)	Time (h)	Leisure	32.41	11.85	9.19	55.63
Garrow et al. (2007)	Time (h)	Business	113.80	33.61	47.92	179.68
Hess and Adler (2011)	Time (h)	Leisure	39.47	2.26	35.04	43.90
Hess and Adler (2011)	Time (h)	Business	100.39	14.30	72.36	128.41
Hess et al. (2011)	Time (h)	All	68.36	11.68	45.47	91.26
Seelhorst and Liu (2015)	Time (h)	Leisure	45.00	25.31	−4.61	94.61
Seelhorst and Liu (2015)	Time (h)	Business	67.50	34.06	0.74	134.26
Crosby et al. (2024)	Wifi	Leisure	85.43	18.85	48.49	122.37
Crosby et al. (2024)	Wifi	Business	112.52	20.11	73.11	151.93

exhibited higher WTPs (Warburg et al., 2006)—patterns GPT-4o-mini reproduced without explicit prompting.

To assess class variability across all LLMs, we used travel time as a representative attribute and calculated the maximum percentage difference in WTP across demographic segments—gender, income group, home location (domestic vs. international), and travel purpose (business vs. leisure). The resulting metric reflects the sensitivity of WTPs to demographic distinctions: higher values indicate greater responsiveness to passenger characteristics.

4. Results

4.1. Estimated WTPs

Table 5 presents the results of our DCE approach (13-attribute design). The corresponding mixed-logit coefficients are provided in Appendix D.

The highest average WTPs were found for the 'Baggage' and 'Wifi' attributes ($118.8 and $97.0, respectively), both also showing a low coefficient of variation across models (CV between 0.35 and 0.40), highlighting consistent preferences for essential travel utilities and connectivity. Meals ($37.2, CV = 0.49), IFE ($55.5, CV = 0.43), Legroom ($56.2, CV = 0.32), Frequent Flyer Benefits ($52.6, CV = 0.45) and Headphones ($47.3, CV = 0.51) followed closely. Lower values were found for Children-free cabins ($36.0, CV = 1.11) and On-time performance (OTP) ($7.7 per 1 % improvement, CV = 0.51), indicating these are more context-dependent and interpreted less uniformly across LLMs. In particular, the high CV for the child-free cabin suggests considerable disagreement on its relevance and appeal, likely due to differing assumptions encoded in the models. While the scarcity of published data on that novel feature could also explain such inconsistency, the 'Headphone' WTPs show a degree of variability similar to the legacy attributes.

As expected, CO2 emissions and Travel Time showed negative WTPs (-$1.0 per g/pax-km, CV = 0.66 and -$89.9 per hour, CV = 0.31, respectively). The relatively low variation of the latter suggests that the models consistently agree on its disutility—making it a stable anchor for segment comparisons.

4.2. Rationality, selective attention, and discriminatory power

The negative price coefficients indicate rational behaviour, consistent with the principle of downward-sloping demand in consumer choice theory (Chen et al., 2023).

Across the 23 models, dominance was consistently achieved at 100 %, confirming that all LLMs reliably avoided choices of dominated alternatives. Stability, however, varied more widely between 52 % and 90 % (mean = 76 %), suggesting some inconsistency in repeated responses. As noted by Gu et al. (2025), such variation likely reflects LLMs' context sensitivity—small prompt changes can subtly alter their preference ordering. Transitivity remained high (average 96 %), indicating that the models' internal preference structures were largely cycle-free. Our probability-based Minimum Cost Index (MCI_{prob}) quantified the severity of these deviations at about 2 %, indicating that when inconsistencies occurred, they were very small in magnitude—the LLM's internal choice probabilities deviated from full rationality by only about 2 % on average.

Next, we quantified selective attention using the HHI of attribute importance shares—a baseline value of 769 implies all 13 attributes carry equal weight. Our models averaged an HHI of 1723, suggesting that LLMs, like humans, focus on only a few salient features. Most models gave weight to 4–6 attributes, indicating relevance-based heuristics similar to those used by human decision-makers (Hensher, 2006c). This is clearly seen in Fig. 1, which shows the mean attribute importance across all models and disaggregated by provider. Travel Time and Price captured around 20–30 % of decision weight in all cases, with Google's models 'caring' less about travel time savings, and DeepSeek placing more emphasis on cheap fares. Interestingly, the Chinese LLMs gives less importance to CO_2, IFE, and Frequent-Flyer Benefits, while giving more weight to checked-in baggage. This agrees with the cultural bias literature (Lu et al., 2025), suggesting that, despite being prompted to simulate US travellers, DeepSeek mirrors travel priorities more typical among Chinese users, who may value tangible benefits over environmental considerations or elite perks (McKinsey, 2018). Indeed, recent work shows that Chinese-trained models display distinct value biases compared to Western ones—down-weighting "self-expression" values (such as environmentalism), thus reflecting more collectivist, utilitarian preferences (Segerer, 2025).

In view of this, applying LLMs to simulate air travellers in regions beyond their cultural training may yield valuations shaped by external cultural biases rather than genuine local behaviour. Incorporating culturally aligned training and calibration procedures could enhance the reliability of such models (Tao et al., 2024); without such adaptation, LLM-based passenger simulations risk misrepresenting demand patterns across diverse markets.

Comparing attribute importance by model type also delivers interesting insights (Fig. 2). Reasoning models appear to focus more heavily on core features such as travel time and on-time performance, while general-purpose models spread attention more evenly. This reflects architectural training: reasoning models prioritise fewer, high-utility attributes, consistent with their design for structured inference. Theoretical work shows that chain-of-thought can induce sparse, selective attention (Wen et al., 2024), and reasoning-oriented architectures embed

Table 5
Estimated WTPs and KPIs (discrete choice experiment).

Attribute\Model	*35t*	*41m*	*41n*	*41x*	*45x*	*4o*	*4om*	*5m*	*5n*	*5x*	*dsc*	*dsc3*	*dsr*	*dsr3*	*g15f*	*g15f8*	*g20f*	*g25f*	*o1m*	*o1x*	*o3m*	*o3x*	*o4m*
WTP_Baggage	91.6	178.9	63.3	83.1	85.0	159.7	140.8	206.2	120.4	90.7	121.8	85.0	157.2	85.0	183.1	86.1	103.3	101.8	117.6	92.4	94.8	78.0	205.5
WTP_Children	−14.3	59.5	5.0	31.3	50.1	55.8	−10.0	62.3	33.4	15.7	53.5	72.2	70.8	72.2	25.2	−51.6	14.2	158.1	5.9	32.3	26.0	22.9	36.7
WTP_CO2 (g-pax-km)	−1.7	−1.3	−2.1	−0.9	−1.0	−0.9	−1.3	−1.2	−1.4	−0.5	−0.8	−0.2	−0.3	−0.2	−2.5	−2.7	−1.3	−0.7	−0.5	−0.6	−0.3	−0.4	−0.9
WTP_FF_Benefits	51.5	81.6	30.6	37.2	42.1	59.5	71.5	69.0	64.5	27.8	26.9	31.5	12.2	31.5	95.8	60.8	80.4	57.5	106.5	42.3	40.3	25.6	63.5
WTP_IFE	76.9	40.7	47.6	43.6	67.3	68.3	89.1	84.7	77.4	28.4	31.1	14.7	17.6	14.7	96.5	56.7	44.6	68.7	88.3	45.9	52.2	46.7	74.1
WTP_Legroom (in)	51.3	50.7	69.9	65.3	44.7	91.5	74.5	81.2	74.1	54.2	56.0	53.3	11.8	53.3	78.8	26.6	73.6	34.5	55.8	46.0	37.4	51.2	58.0
WTP_Meals	73.3	120.2	37.9	54.7	45.5	99.8	151.0	97.3	81.0	46.5	50.4	39.2	13.8	39.2	45.0	62.8	61.0	26.2	108.3	56.4	93.0	41.9	101.9
WTP_Headphones	41.6	51.5	68.8	30.7	43.8	38.2	66.1	87.5	69.5	14.3	50.9	20.0	11.8	20.0	65.0	56.4	41.4	75.2	76.9	16.8	26.7	22.1	91.9
WTP_OTP (1 %)	4.5	5.2	8.8	3.0	5.6	4.9	9.7	10.8	12.3	7.0	5.7	7.3	5.8	7.3	6.0	0.8	7.2	17.4	13.4	4.7	15.8	5.3	8.9
WTP_Seat_Location	−5.3	3.1	4.0	−0.8	−4.1	−13.0	2.7	36.9	−3.7	7.6	−3.1	−11.7	−11.8	−11.7	1.5	−6.4	−15.5	4.6	−6.1	6.2	−22.3	−6.2	14.3
WTP_Travel_Time (h)	−67.0	−57.6	−135.1	−66.4	−97.2	−95.2	−72.0	−142.4	−100.8	−106.6	−71.5	−101.4	−59.6	−101.4	−74.4	−21.0	−77.7	−116.5	−73.6	−92.7	−123.6	−96.4	−118.6
WTP_WiFi	73.0	65.1	87.6	72.5	94.0	88.8	180.6	172.1	129.5	75.8	68.2	36.6	60.9	36.6	103.9	89.9	68.2	170.4	120.4	110.4	88.3	95.9	143.0
Stability	70 %	69 %	52 %	83 %	80 %	77 %	71 %	83 %	75 %	89 %	76 %	74 %	86 %	90 %	71 %	55 %	75 %	83 %	67 %	87 %	77 %	87 %	78 %
Dominance	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Transitivity	98 %	97 %	93 %	99 %	99 %	100 %	95 %	96 %	97 %	93 %	97 %	94 %	100 %	100 %	93 %	99 %	95 %	98 %	97 %	99 %	100 %	94 %	93 %
MCIprob	1 %	1 %	2 %	0 %	0 %	0 %	1 %	4 %	3 %	7 %	1 %	6 %	0 %	0 %	3 %	0 %	1 %	0 %	1 %	0 %	0 %	2 %	3 %
Plausibility	59 %	65 %	35 %	71 %	35 %	65 %	41 %	12 %	35 %	24 %	59 %	35 %	47 %	12 %	59 %	59 %	59 %	24 %	47 %	41 %	12 %	29 %	18 %
Plausibility gain	24 %	65 %	−12 %	36 %	0 %	24 %	35 %	12 %	29 %	−12 %	18 %	17 %	12 %	−12 %	41 %	18 %	18 %	0 %	41 %	6 %	−6 %	0 %	0 %
Expected signs	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
Beta Seat_Location	0.40 %	0.20 %	0.20 %	0.10 %	0.30 %	0.80 %	0.20 %	1.60 %	0.20 %	0.50 %	0.20 %	0.80 %	1.00 %	0.10 %	0.10 %	0.60 %	1.00 %	0.20 %	0.30 %	0.40 %	1.20 %	0.50 %	0.70 %
HHI attrib. importance	1552	1783	1353	2042	1777	1521	1268	1508	1484	2143	1660	2113	1946	2346	1301	1568	1556	1636	1416	1893	1959	2055	1519
gender_diff	14 %	2 %	9 %	6 %	26 %	23 %	2 %	3 %	5 %	63 %	26 %	6 %	19 %	22 %	20 %	2 %	5 %	13 %	3 %	26 %	3 %	22 %	11 %
income_diff	12 %	24 %	31 %	146 %	133 %	93 %	11 %	173 %	39 %	271 %	88 %	188 %	76 %	229 %	36 %	14 %	151 %	142 %	59 %	296 %	184 %	192 %	159 %
nationality_diff	3 %	8 %	3 %	12 %	13 %	20 %	18 %	48 %	4 %	8 %	21 %	27 %	7 %	67 %	12 %	12 %	25 %	23 %	7 %	20 %	19 %	28 %	4 %
purpose_diff	9 %	11 %	0 %	73 %	31 %	17 %	8 %	109 %	15 %	157 %	54 %	53 %	36 %	121 %	25 %	16 %	62 %	37 %	10 %	53 %	53 %	116 %	45 %

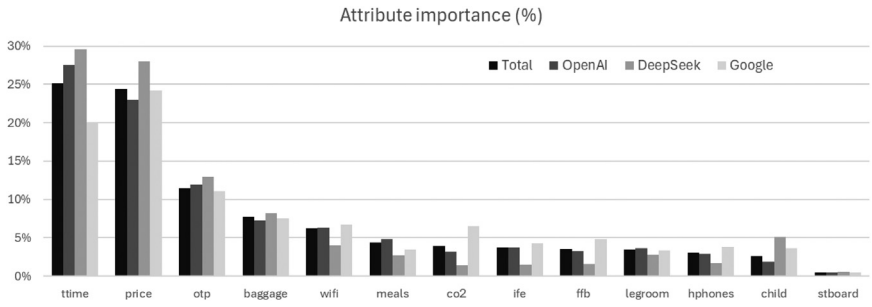

Fig. 1. Mean attribute importance (%) per LLM provider.

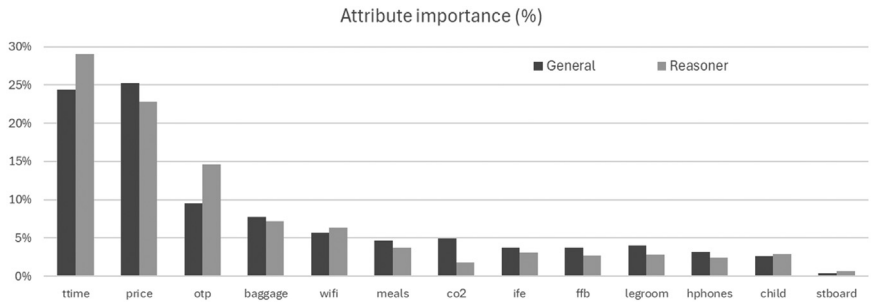

Fig. 2. Mean attribute importance (%) per model type.

this mechanism more deeply, leading to sharper focus compared to the broader patterns of generalist models.

Figs. 1 and 2 also illustrate the LLMs' discriminatory power—the decision weight of the 'Seat Location' feature (sitting starboard) is negligible in comparison to the other attributes.

4.3. Human-LLM alignment: plausibility and class variability

All the WTPs (except 'child-free' and 'seat location') have the expected signs across all models. Plausibility scores, which indicate the percentage of estimated WTPs falling within historical 95 % confidence intervals from prior studies, ranged from 12 % to 71 %, averaging around 41 % across the sample. Methodological differences between human stated-preference surveys and LLM-simulated DCEs had likely contributed to the deviations. Unlike human respondents, LLMs do not experience fatigue, bounded attention, or declining effort across repeated tasks, which may lead to more extreme or consistent valuations when many attributes are presented. In contrast, human SP studies show that greater dimensional complexity (more attributes or alternatives) can degrade response quality and alter WTP ranges due to cognitive constraints (Hensher, 2006b; Caussade et al., 2005).

In spite of that, the top-performing models, gemini-2.0-flash and gpt-4.1, matched the reference values in five out of six attributes and captured the expected WTP differential between business and leisure travellers. At the lower end, o4-mini and o3-mini scored lowest on plausibility, primarily due to unrealistically high valuations for travel time savings. Despite the evidence of cultural bias discussed above, DeepSeek models achieve plausibility levels comparable to US-trained LLMs, indicating that WTP deviations are small enough to fall within the confidence intervals reported in the literature for US travellers.

Plausibility gains—the improvement over direct elicitation values (see Appendix E)—were positive for most models, averaging 15 %. In general, DCEs produced higher WTPs than direct responses for nearly all attributes, with the exception of Travel Time (–$26.2) and OTP (–$10.4), which were slightly undervalued. On average, the difference across attributes was $27.2, with the largest gains observed for Baggage (+$89.4), WiFi (+$79.7), and Meals (+$48.0). These differences likely stem from the structured nature of DCEs, which anchor LLMs' judgments through explicit trade-offs, encouraging higher, more deliberate valuations—whereas direct elicitation often yields more conservative estimates due to its abstract and unanchored format. This pattern aligns well with the broader debate on hypothetical bias (HB) in choice experiments (e.g., Haghani et al., 2021). While some argue that choice experiments reduce HB with respect to direct methods (Penn & Hu, 2018), others find the opposite (Schmidt & Bijmolt, 2020). Our LLM results reflect the same ambiguity found for human-based studies.

Demographic sensitivity in travel time WTPs varies notably across segments. Differences by gender (*gender_diff*) and home location (*nationality_diff*) average around 14–18 %, whereas the gap between business and leisure travellers (*purpose_diff*) is substantially larger, averaging 48 %. When benchmarked against the published reference values in Table 5—where business–leisure WTP differentials range from 50 % to 251 % (Seelhorst & Liu, 2015; Garrow et al., 2007)—most LLMs underestimate the true magnitude of this variation. Direct elicitation responses exhibit greater demographic differentiation, although two sample models appear largely insensitive to travel purpose.

4.4. Discussion

In view of the above, commercial LLMs show promising capabilities as imitators of human decision-making in airline passenger contexts. Consistent with findings in cognitive science and economics (Chen et al., 2023; Gao et al., 2024; Strachan et al., 2024; Wang et al., 2025a), our results indicate that LLMs display structured, coherent decision patterns resembling human rationality, yet limited responsiveness to demographic cues. Their main strength lies in producing rational choices and realistic WTPs for key service attributes such as Travel Time and Legroom, aligning with evidence that LLMs can replicate core aspects of economic reasoning. However, plausibility remains limited, with only 41 % of WTPs within empirical confidence intervals. This supports prior observations that LLMs reproduce aggregate human-like behaviour but lack socio-demographic nuance (Yan et al., 2025). Most models in our study struggled to differentiate business and leisure valuations, underscoring the need for domain-specific fine-tuning to achieve behavioural realism.

Are LLMs at least improving in this area? Fig. 3 averages four key benchmarks according to the models' release year, size, and type. The results suggest clear progress, particularly in structured reasoning and behavioural segmentation. Stability rose from 71 % in 2024 to 79 % in 2025, indicating that newer models make more consistent choices across

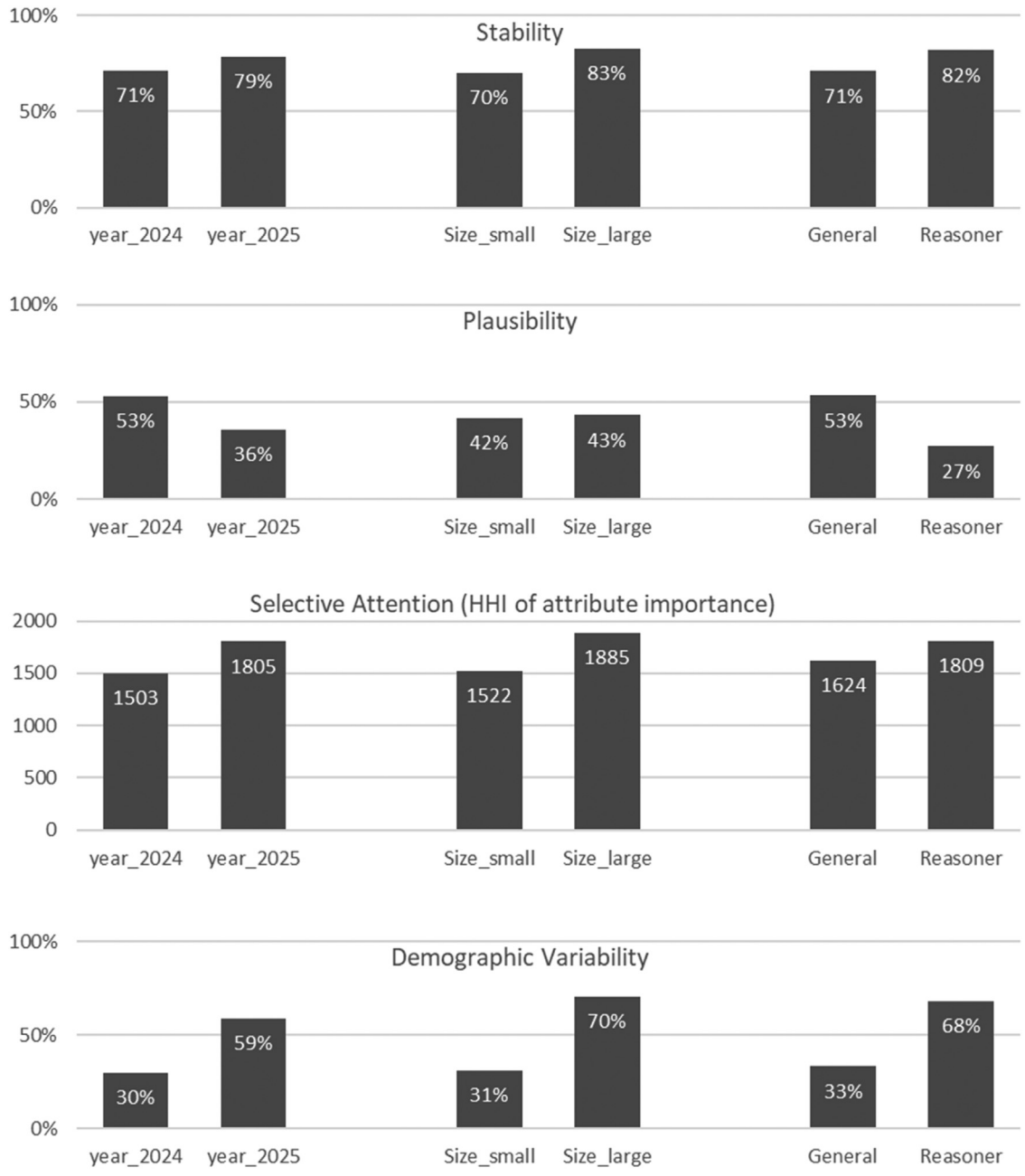

Fig. 3. Comparisons across model types.

repeated or logically related tasks. The effect is stronger by model size: large models outperform small ones on stability (83 % vs. 70 %) and selective attention (HHI 1885 vs. 1522), implying that scale enhances decision structure and focus. Reasoning models also show higher stability (82 %) and selective attention (1809) than general-purpose LLMs (71 % and 1624), suggesting that architectures trained for logical consistency are better suited to decision emulation. Demographic variability increases with both time and scale—from 30 % in 2024 to 59 % in 2025, and from 31 % in small models to 70 % in large ones—indicating greater sensitivity to segmentation cues. However, plausibility declines in newer (36 %) and reasoning-oriented (27 %) models, suggesting a trade-off: as structural reasoning improves, alignment with empirical preference data weakens. General-purpose models, though less specialised, outperform reasoners in plausibility (53 % vs. 27 %), implying that realistic behaviour may require balancing logical depth with domain alignment.

Based on our knowledge audit, we recommend that small and mid-sized airlines optimise their AI-driven marketing research costs by using large, general-purpose LLMs rather than reasoner models, which consume more tokens but deliver less realistic results. OpenAI's GPT-4.1 emerged as the best overall performer, offering strong balance across benchmarks at 20 % lower cost per million output tokens than GPT-4o ($8 vs $10). Among smaller, low-cost models ($0.40–$1.60), GPT-4.1-mini and Gemini-2.0-flash stand out, though with some trade-offs in capability.

The consistent travel time WTPs across models suggest that LLMs are particularly useful for modelling decisions where time savings are the key factor. Airlines could employ them to simulate passenger responses to adding nonstop flights on previously connecting routes, competing with high-speed rail, or introducing supersonic travel. For instance, Boom Supersonic's planned Overture jet, expected to halve long-haul flight times (Boom Supersonic, 2025), is a clear case where LLM-derived

WTPs could inform early pricing or demand forecasts. More broadly, LLMs can serve as virtual test markets for innovations by screening new service bundles or amenities. In our simulations, the models assigned meaningful value to relevant features such as noise-cancelling headphones while downplaying trivial ones like seat side. This aligns with evidence that LLMs detect impactful attributes while disregarding irrelevant factors (Argyle et al., 2023).

That said, LLMs should be viewed as complements rather than substitutes for human data, with reliability requiring careful validation. While GPT models can reproduce aggregate consumer preferences, they may misestimate niche features (Brand et al., 2024), highlighting the risks of unverified outputs (Dominguez-Olmedo et al., 2023). Airlines are well positioned to audit LLM-derived preferences against internal resources such as booking data, loyalty segmentation, and prior surveys. Without such checks, outputs may reflect cultural biases in training data (Andreev et al., 2025). Scholars likewise caution that "silicon samples" must be cross-validated across regions and contexts (Miranda & Balbi, 2025; Argyle et al., 2023). In practice, LLMs are best deployed as rapid, low-cost tools for scenario testing, hypothesis generation, and exploring variability in passenger preferences—tasks that can later be calibrated and verified through traditional DCE surveys.

5. Summary

This paper set out to evaluate whether commercial, off-the-shelf LLMs can reliably imitate airline passengers' decision-making. Motivated by the growing integration of AI into airline operations—particularly in pricing, personalisation, and service design—we examined whether LLMs could serve as practical tools for simulating consumer preferences. Drawing on decades of discrete-choice research, we implemented a flight-ticket choice experiment with 13 attributes, using responses from 23 LLMs prompted in zero-shot mode. These responses were analysed through logit-based regressions to extract WTP estimates, which we audited across five benchmarks: rationality, selective attention, discriminatory power, plausibility, and class variability.

Results indicate that while LLMs are not perfect proxies for human passengers, they display sufficient knowledge of air travel to make rational choices, correctly sign ticket attributes, and produce realistic WTPs for core features—particularly travel time. Large, general-purpose models, especially GPT-4.1, performed best across all criteria. Smaller or mid-sized airlines could leverage such accessible models to test service features, gauge price sensitivities, or simulate market reactions to innovations like supersonic travel or air–HSR competition, at a fraction of traditional survey costs.

Despite these promising findings, several limitations remain. First, we relied exclusively on zero-shot prompting. Using basic prompts without examples or fine-tuning may have constrained the richness of model responses. Prior research shows that prompt design strongly shapes model behaviour (Miranda & Balbi, 2025), and structured strategies such as chain-of-thought prompting often yield more reliable, nuanced outputs (Sfeir et al., 2025). Future work should therefore explore few-shot examples or light fine-tuning to enhance fidelity, particularly for modelling subtle trade-offs and realistic decision heuristics.

Second, our synthetic sample size was modest. We simulated 150 "respondents" (each with six choice tasks), which is small by conventional DCE standards. A robustness test with 1000 respondents produced nearly identical WTP estimates—albeit with narrower confidence intervals—suggesting reasonable stability. Nonetheless, larger synthetic samples or alternative approaches, such as extracting probability distributions from LLM outputs rather than single choices, could yield more efficient and reliable inference (Sun & Zhang, 2025).

Third, deviations from benchmark WTPs may reflect methodological asymmetries. Human respondents face cognitive limits when processing many attributes, and design dimensionality is known to affect WTPs and error variance (Hensher, 2006b; Caussade et al., 2005). LLMs, by contrast, can handle high-dimensional tasks effortlessly, potentially producing more extreme or consistent valuations. Future benchmarking could improve validity by constraining LLM tasks to better reflect human cognitive limits.

Finally, the generalisability of our findings remains uncertain. We focused on a single US long-haul route, and the models' knowledge may embed cultural or market-specific biases. Thus, our valuations may not transfer directly to short-haul or non-US contexts. Fine-tuning on regional data or testing across diverse markets could address this. Relatedly, the models underestimated business–leisure differences, suggesting that while LLMs capture average preferences, they underrepresent subgroup heterogeneity. Overall, LLM respondents should complement, not replace, human data, with outputs validated against booking records and survey evidence until their robustness across contexts is firmly established.

Declaration of competing interest

The author reports there are no competing interests.

CRediT authorship contribution statement

Augusto Voltes-Dorta: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization. **Pere Suau-Sanchez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Conceptualization.

Appendix A. Example prompt

You are a traveller at San Francisco International Airport (SFO). Your persona has been pre-generated based on demographic probabilities.

Persona: Gender: Female, Age: 35–44, Income: $50,000–$100,000, Home Location: San Francisco County, Trip Purpose: Visit Friends or Relatives.

Answer the following questions about a San Francisco - New York flight based on this persona. Your preferences regarding the travel attributes must remain constant through the survey. When making choices, apply the principle of bounded rationality: use simplified decision-making heuristics, focus on key attributes, and aim for satisficing rather than optimising outcomes.

How many times have you flown out of SFO in the past 12 months? Response: 3: 3–6 times

Please select your preferred flight option for Task_1. Also, provide the choice probabilities (in percentages) that you would assign to each alternative. Provide your answer in the following format:

Selected Alternative: [1, 2, 3, or 4]
Probability Alternative 1: [value] %
Probability Alternative 2: [value] %
Probability Alternative 3: [value] %
Probability Alternative 4: [value] %

Alternative 1: -Price: $220 -Travel_Time: 8 h *(1 stop) -OTP: 75 % -WiFi: Not Available -IFE: Available -Meals: Included -Baggage: One Free Checked Bag -FF_Benefits: Not Available -Noise-Cancelling headphones: Not Available -Children: Children allowed -Seat_Location: Starboard side -Legroom: 33 inches -CO2_Emissions: 90* g *CO2/passenger-km*

Alternative 2: -Price: $150 -Travel_Time: 11 h *(2 stops) -OTP: 60 % -WiFi: Not Available -IFE: Not Available -Meals: Not Included -Baggage: No Free Checked Bag -FF_Benefits: Not Available -Noise-Cancelling headphones: Not Available -Children: Children allowed -Seat_Location: Port side -Legroom: 31 inches -CO2_Emissions: 150* g *CO2/passenger-km*

Alternative 3: -Price: $460 -Travel_Time: 5.5 h *-OTP: 90 % -WiFi: Available -IFE: Available -Meals: Included -Baggage: One Free Checked Bag -FF_Benefits: Available -Noise-Cancelling headphones: Available -Children: Child-free section -Seat_Location: Starboard side -Legroom: 31 inches -CO2_Emissions: 120* g *CO2/passenger-km*

Alternative 4: -Price: $320 -Travel_Time: 8 h *(1 stop) -OTP: 75 % -WiFi: Available -IFE: Available -Meals: Not Included -Baggage: One Free*

Checked Bag -FF_Benefits: Available -Noise-Cancelling headphones: Available -Children: Children allowed -Seat_Location: Port side -Legroom: 33 inches -CO2_Emissions: 150 g CO2/passenger-km

Appendix B. Probability-based Minimum Cost Index (MCI_{prob}) formulation

Let $T=(A,B,C)$ be the three alternatives used in Task 20. For each two-alternative sub-task, we define:

p_{AB}=Pr(choose A|{A,B}), p_{BC}=Pr(choose B|{B,C}), p_{CA}=Pr(choose C|{C,A}).

A cycle (transitivity violation) occurs if either

$(p_{AB}>0.5 \wedge p_{BC}>0.5 \wedge p_{CA}>0.5)$ or $(p_{AB}<0.5 \wedge p_{BC}<0.5 \wedge p_{CA}<0.5)$

For each pairwise task, define the margin to indifference as the minimum probability mass that must be reassigned to reach the 0.5 boundary:

$m_{AB}=|p_{AB}-0.5|$, $m_{BC}=|p_{BC}-0.5|$, $m_{CA}=|p_{CA}-0.5|$.

The probability-based Minimum Cost Index (MCI_{prob}) for the triad is calculated as $\min\{m_{AB}, m_{BC}, m_{CA}\} \in [0,0.5]$, if a cycle is present, and 0 otherwise.

Appendix C. Pilot run

The pilot run (Tables C1-C3; Fig. C1) followed the same survey framework and demographic sampling procedure described in Section 3.3. Respondent profiles were drawn from the 2018 SFO Passenger Survey (see Table 1) to preserve demographic realism. For each design (ranging from 4 to 13 attributes), 150 synthetic respondents were generated for both the GPT-4o-mini and o3 models, resulting in 2550 choice tasks per model (17 tasks × 150 respondents) and 10,200 evaluated alternatives in total. This sample size ensured stable maximum-likelihood estimation in conditional and mixed logit models, balancing computational cost with demographic diversity in traveller profiles. All respondents completed the tasks under zero-shot, zero-temperature prompting (Appendix A). The pilot's main objectives were to test estimation feasibility, assess the stability of WTP patterns under increasing attribute dimensionality, and compare behavioural responses between a small general model (GPT-4o-mini) and a large reasoning model (o3) before expanding the analysis to the full LLM sample.

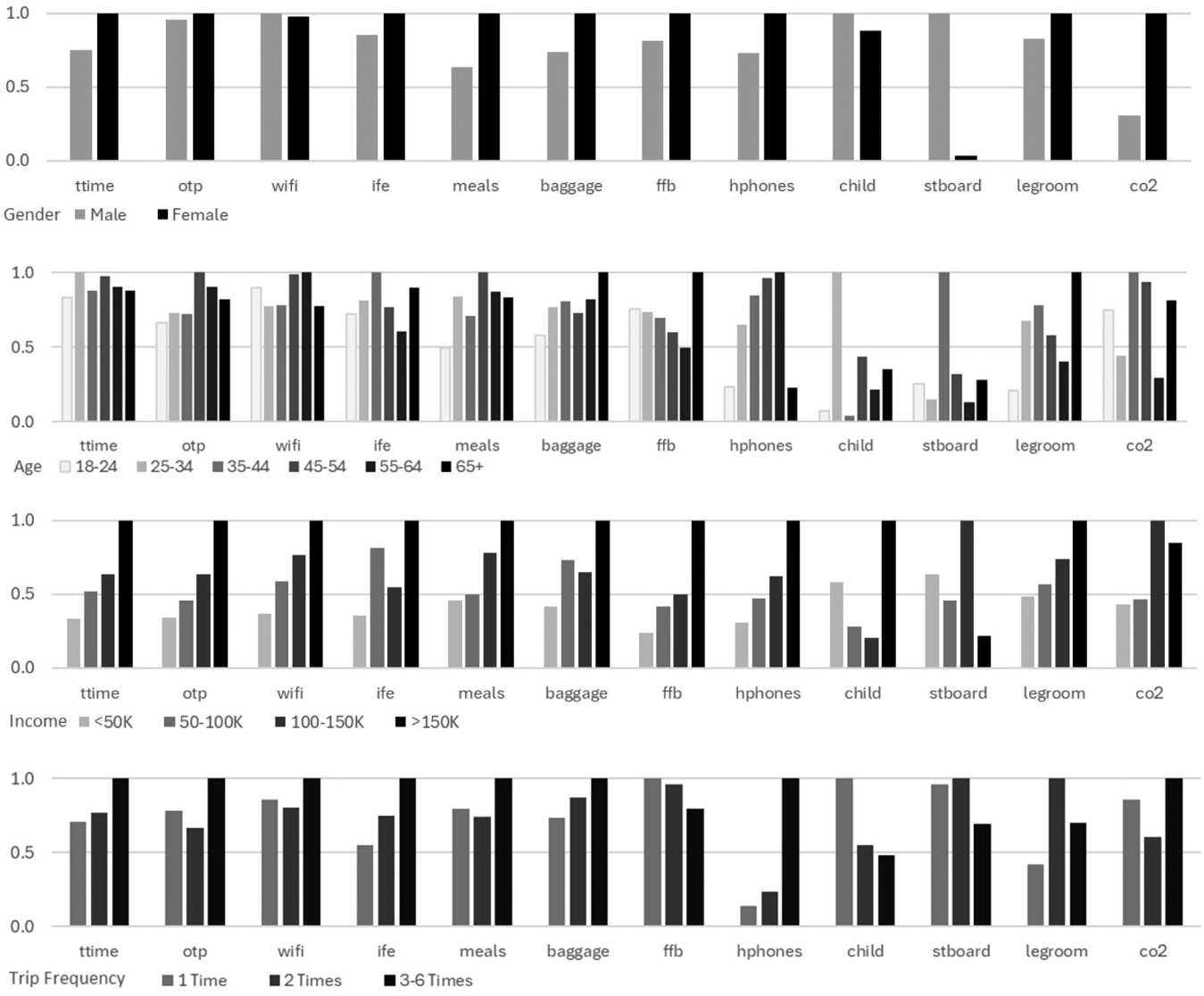

Fig. C1. Normalised WTPs by demographic characteristics (pilot run using gpt-4o-mini-2024-07-18).

Table C1
WTP estimates from conditional logit models (pilot run).

Model	*gpt-4o-mini-2024-07-18*					*o3-2025-04-16*				
Attribute	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*	*WTP ($)*
Travel_Time (h)	−86.3	−70.6	−63.0	−63.0	−66.0	−114.8	−110.2	−107.0	−106.8	−100.7
Legroom (in)	76.9	55.4	32.4	36.6	39.6	59.4	32.2	28.3	32.0	59.2
CO2_Emissions (kg)	−0.7	−0.5	−0.3	−0.3	−0.3	−1.5	−1.4	−0.5	−0.4	−0.4
OTP (1 %)		10.6	9.2	8.5	10.3		8.4	7.8	6.6	5.2
Baggage		241.8	165.8	183.6	177.6		102.7	94.0	94.1	86.4
WiFi			176.8	174.6	179.3			111.7	101.6	97.7
IFE			95.8	87.3	90.5			46.7	46.7	53.4
Meals			119.1	115.6	125.6			39.9	41.6	45.5
FF_Benefits				83.3	79.7				27.7	33.4
Seat_Location				10.2	−11.8				−5.8	−1.8
NC_Headphones					65.8					23.2
Child_free					−17.6					24.1

Table C2
Mixed logit and latent-class logit regression results (pilot run: gpt-4o-mini-2024-07-18).

Mixed Logit	**Business**				**Leisure**		
Mean	Coeff.	S.d.	Prob.	Mean	Coeff.	S.d.	Prob.
Price	−0.007	0.001	0.000	Price	−0.008	0.001	0.000
Travel_Time	−0.531	0.040	0.000	Travel_Time	−0.502	0.036	0.000
OTP	8.402	0.671	0.000	OTP	7.536	0.728	0.000
WiFi	1.360	0.145	0.000	WiFi	1.400	0.139	0.000
IFE	0.747	0.133	0.000	IFE	0.671	0.125	0.000
Meals	0.840	0.136	0.000	Meals	1.026	0.129	0.000
Baggage	1.519	0.161	0.000	Baggage	1.285	0.155	0.000
FF_Benefits	0.523	0.144	0.000	FF_Benefits	0.676	0.125	0.000
NC_Headphones	0.543	0.130	0.000	NC_Headphones	0.465	0.137	0.001
Children	0.016	0.132	0.906	Children	−0.322	0.121	0.008
Seat_Location	−0.137	0.128	0.285	Seat_Location	−0.071	0.121	0.557
Legroom	0.247	0.074	0.001	Legroom	0.345	0.072	0.000
CO2_Emissions	−0.007	0.003	0.023	CO2_Emissions	−0.009	0.002	0.000
S.d.				S.d.			
Price	−0.002	0.001	0.000	Price	0.002	0.001	0.000
Travel_Time	−0.001	0.069	0.987	Travel_Time	−0.087	0.048	0.069
OTP	−0.071	1.102	0.948	OTP	−3.152	0.841	0.000
WiFi	0.098	0.632	0.877	WiFi	−0.311	0.221	0.160
IFE	0.044	0.266	0.868	IFE	−0.194	0.185	0.293
Meals	−0.123	0.316	0.697	Meals	−0.156	0.353	0.659
Baggage	−0.377	0.209	0.071	Baggage	−0.597	0.193	0.002
FF_Benefits	0.497	0.199	0.013	FF_Benefits	0.063	0.233	0.787
NC_Headphones	−0.059	0.185	0.751	NC_Headphones	−0.426	0.233	0.068
Children	0.242	0.283	0.391	Children	−0.037	0.156	0.814
Seat_Location	−0.100	0.208	0.630	Seat_Location	0.155	0.268	0.564
Legroom	0.278	0.096	0.004	Legroom	−0.217	0.102	0.034
CO2_Emissions	−0.009	0.004	0.014	CO2_Emissions	0.000	0.005	0.922
Latent-class Logit	**Business**				**Leisure**		
Class 1 (54.7 %)	Coeff.	S.d.	Prob.	Class 1 (56.3 %)	Coeff.	S.d.	Prob.
Price	−0.004	0.001	0.000	Price	−0.009	0.001	0.000
FF_Benefits	0.607	0.166	0.000	OTP	5.432	0.782	0.000
Legroom	0.237	0.111	0.033	Baggage	1.147	0.172	0.000
CO2_Emissions	−0.008	0.003	0.017	Legroom	0.325	0.079	0.000
Class 2 (45.3 %)				Class 2 (43.7 %)			
Price	−0.009	0.001	0.000	Price	−0.005	0.001	0.000
FF_Benefits	0.389	0.206	0.059	OTP	8.757	0.935	0.000
Legroom	0.199	0.137	0.146	Baggage	1.168	0.192	0.000
CO2_Emissions	−0.003	0.004	0.545	Legroom	0.301	0.092	0.001
Membership Class 1				Membership Class 1			
Income	2.029	2.007	0.312	Income	−13.968	67.037	0.835
International	1.155	3.158	0.714	International	0.009	2.051	0.996

(*Continued on the next page*)

Table C2
(*Continued*).

Latent-class Logit	**Business**				**Leisure**		
Class 1 (54.7 %)	Coeff.	S.d.	Prob.	Class 1 (56.3 %)	Coeff.	S.d.	Prob.
TimesFlownSFOs	15.026	161.090	0.926	TimesFlownSFOs	12.112	67.038	0.857
Constant	−50.733	483.269	0.916	Constant	5.057	3.851	0.189
Fixed coefficients				Fixed coefficients			
Travel_Time	−0.494	0.036	0.000	Travel_Time	−0.462	0.029	0.000
WiFi	1.259	0.131	0.000	WiFi	1.356	0.119	0.000
IFE	0.710	0.125	0.000	IFE	0.607	0.110	0.000
Meals	0.752	0.124	0.000	Meals	0.948	0.115	0.000
NC_Headphones	0.499	0.125	0.000	NC_Headphones	0.436	0.112	0.000
Children	0.028	0.121	0.817	Children	−0.264	0.110	0.016
Seat_Location	−0.138	0.119	0.244	Seat_Location	−0.062	0.109	0.567
OTP	7.812	0.607	0.000	FF_Benefits	0.629	0.114	0.000
Baggage	1.385	0.141	0.000	CO2_Emissions	−0.008	0.002	0.001

Table C3
WTP estimates from the latent-class logit models (pilot run using gpt-4o-mini-2024-07-18).

	WTP Business ($)			***WTP Holiday ($)***		
Attribute	***All***	***Class 1***	***Class 2***	***All***	***Class 1***	***Class 2***
Travel_Time (h)	−79.1			−61.3		
Legroom (in)	36.8	56.2	22.8†	42.1	36.1	59.8
CO2_Emissions (kg)	−0.24	−0.47	−0.07†	−0.26		
OTP (1 %)	12.5			9.2	6.0	17.4
Baggage	226.6			157.0	127.2	231.9
WiFi	202.9			171.0		
IFE	111.4			82.0		
Meals	125.3			125.3		
FF_Benefits	78.0	143.6	44.6†	82.6		
Seat_Location	−20.4†			−8.7		
NC_Headphones	81.1			56.8		
Child_free	2.3†			−39.3		
Class Share		55 %	45 %		56 %	44 %

Note: † denotes non-significant coefficients at 5 %.

Attribute\Model	*35t*	*41m*	*41n*	*41x*	*45x*	*4o*	*4om*	*5m*	*5n*	*5x*	*dsc*	*dsc3*	*dsr*	*dsr3*	*g15f*	*g15f8*	*g20f*	*g25f*	*o1m*	*o1x*	*o3m*	*o3x*	*o4m*
Price	−0.010	−0.010	−0.004	−0.013	−0.010	−0.011	−0.006	−0.006	−0.007	−0.009	−0.009	−0.006	−0.007	−0.006	−0.007	−0.008	−0.008	−0.006	−0.005	−0.010	−0.007	−0.009	−0.006
Baggage	0.886	1.840	0.239	1.053	0.848	1.711	0.911	1.138	0.801	0.830	1.046	0.497	1.050	0.374	1.235	0.655	0.777	0.565	0.567	0.908	0.634	0.704	1.156
Children	−0.138	0.612	0.019	0.397	0.500	0.597	−0.065	0.344	0.222	0.144	0.459	0.422	0.473	0.480	0.170	−0.393	0.107	0.877	0.028	0.317	0.174	0.207	0.207
CO2 (g/pax-km)	−0.017	−0.013	−0.008	−0.011	−0.010	−0.009	−0.009	−0.007	−0.009	−0.005	−0.007	−0.001	−0.002	0.000	−0.017	−0.021	−0.010	−0.004	−0.002	−0.006	−0.002	−0.004	−0.005
FF_Benefits	0.499	0.839	0.116	0.471	0.420	0.637	0.463	0.380	0.429	0.255	0.231	0.184	0.082	0.076	0.646	0.463	0.605	0.319	0.514	0.416	0.269	0.231	0.357
Headphones	0.403	0.530	0.260	0.389	0.436	0.409	0.428	0.483	0.463	0.130	0.437	0.117	0.079	0.053	0.438	0.429	0.311	0.417	0.371	0.165	0.179	0.199	0.517
IFE	0.744	0.418	0.180	0.552	0.671	0.732	0.577	0.468	0.515	0.260	0.268	0.086	0.117	0.085	0.651	0.431	0.335	0.381	0.426	0.452	0.349	0.422	0.417
Legroom (in)	0.496	0.522	0.264	0.827	0.446	0.979	0.482	0.448	0.493	0.495	0.481	0.312	0.079	0.183	0.532	0.202	0.554	0.191	0.269	0.452	0.250	0.462	0.326
Meals	0.709	1.237	0.143	0.692	0.454	1.069	0.977	0.537	0.539	0.426	0.433	0.229	0.092	0.275	0.304	0.478	0.459	0.145	0.523	0.555	0.622	0.378	0.573
OTP	4.309	5.301	3.340	3.843	5.594	5.201	6.271	5.932	8.155	6.413	4.921	4.248	3.890	5.037	4.034	0.590	5.407	9.645	6.475	4.635	10.573	4.819	4.995
Seat_Location	−0.051	0.032	0.015	−0.010	−0.041	−0.139	0.018	0.204	−0.025	0.070	−0.027	−0.068	−0.079	−0.004	0.010	−0.049	−0.116	0.026	−0.029	0.061	−0.149	−0.056	0.080
Travel_Time (h)	−0.648	−0.592	−0.511	−0.841	−0.969	−1.019	−0.466	−0.785	−0.671	−0.975	−0.614	−0.593	−0.398	−0.634	−0.502	−0.160	−0.585	−0.647	−0.355	−0.911	−0.827	−0.871	−0.667
WiFi	0.707	0.669	0.331	0.917	0.937	0.951	1.168	0.950	0.862	0.693	0.586	0.214	0.407	0.259	0.701	0.684	0.513	0.946	0.581	1.086	0.590	0.865	0.804

Attribute\Model	*35t*	*41m*	*41n*	*41x*	*45x*	*4o*	*4om*	*5m*	*5n*	*5x*	*dsc*	*dsc3*	*dsr*	*dsr3*	*g15f*	*g15f8*	*g20f*	*g25f*	*o1m*	*o1x*	*o3m*	*o3x*	*o4m*
WTP_Baggage	38.5	22.1	28	20.3	25.6	25.5	34.2	33.0	47.1	27.1	28.4	27.5	26.1	30.3	26.9	36.4	31	40	32.7	21.7	20.6	24.3	26.9
WTP_Children	31.6	10	9.5	11.6	21	10.1	37	20.3	18.6	10.3	38.2	41.5	21.5	31.7	24	22.5	29	45.7	31.6	16.3	9.2	19.4	13.4
WTP_CO2 (g/pax-km)	−24.1	−1.5	−6.1	−0.2	−6.5	−4.3	−7.4	0.0	−10.0	0.0	−3.9	3.3	−1	1.8	−3	−4.9	−0.1	−9.6	−6.7	−2	−0.2	−0.1	0
WTP_FF_Benefits	43.5	24.3	30.5	12	25.9	24.4	48.3	50.0	57.9	17.3	29.6	20.7	9.4	11.5	0	25	24.4	56.6	53.2	18.3	22.6	18.8	26.3
WTP_Headphones	30.6	19	10.2	7.1	9.5	8.9	23.7	24.2	19.3	6.4	19.5	19.2	16.5	17.8	8.3	12	10	22.3	35	9.1	14	13.7	13.2
WTP_IFE	15.5	10.1	15.5	6.5	9.5	10.1	28	6.9	17.6	5.3	12.8	8.8	6.6	9.4	10.9	32.3	10.6	15.3	26.5	7.9	8.3	7.3	8.2
WTP_Legroom (in)	34.4	30.3	36.6	9.9	15.4	14.5	23.5	14.3	39.4	9.7	18.6	27.4	12.4	21.3	14.6	13	13.5	27.4	24.4	12.3	5.6	13.8	10.5
WTP_Meals	25	16.5	13.6	9.8	18.3	19.5	38.6	15.2	23.1	11.4	18.1	16.6	11.1	13.7	10.5	20.3	16.5	20.7	32.4	13.8	14.4	12.8	13.5
WTP_OTP (1 %)	30	5.2	27.8	3.6	10.5	9.6	80.7	4.0	76.2	2.3	7.8	12.9	10	9.2	21.7	20.8	5.7	19.6	39.6	4.6	9	3.2	2.5
WTP_Seat_Location	12.4	3.3	5.2	1.7	3.7	4.1	11.5	2.3	10.1	0.5	5.2	3.7	4.5	4.8	0	5.9	5.4	5.1	15.3	3.5	6.3	2.6	3.1
WTP_Travel_Time (h)	−50	−96.7	−39.3	−54.7	−61	−59.5	−127	−117.0	−157.9	−66.4	−66.1	82.5	−50.1	65.4	−91.5	−56.1	−48.6	−72.8	−113.1	−52.9	−80.8	−69.1	−82.3
WTP_WiFi	19.5	13.2	19	10.4	13.6	14.9	28.3	15.3	23.6	13.3	19.7	17.0	13.8	16.3	18.3	16.3	15.8	25.4	32.6	12.7	14.3	12.9	13
Plausibility	35 %	0 %	47 %	35 %	35 %	41 %	6 %	0 %	6 %	35 %	41 %	18 %	35 %	24 %	18 %	41 %	41 %	24 %	6 %	35 %	18 %	29 %	18 %
Expected signs	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %	100 %
gender_diff	0 %	6 %	16 %	14 %	10 %	13 %	9 %	21 %	11 %	16 %	20 %	7 %	7 %	2 %	5 %	22 %	15 %	10 %	28 %	0 %	17 %	11 %	16 %
income_diff	0 %	259 %	362 %	157 %	221 %	164 %	149 %	284 %	309 %	319 %	297 %	337 %	387 %	361 %	152 %	80 %	310 %	335 %	579 %	222 %	337 %	265 %	413 %
nationality_diff	0 %	9 %	32 %	30 %	4 %	19 %	8 %	15 %	9 %	8 %	17 %	36 %	12 %	19 %	12 %	15 %	11 %	11 %	3 %	15 %	25 %	15 %	30 %
purpose_diff	0 %	38 %	67 %	39 %	53 %	7 %	8 %	91 %	81 %	114 %	41 %	41 %	41 %	116 %	29 %	0 %	34 %	67 %	64 %	33 %	37 %	58 %	145 %

References

Adler, T., Falzarano, C., & Spitz, G. (2005). *Modeling service trade-offs in air itinerary choices', transportation research record: Journal of the transportation research board, no. 1915* (pp. 20–26). Washington, D.C.: Transportation Research Board of the National Academies.

Amadeus.. *Dynamic pricing* Available at: Accessed: 15 January 2025 https://amadeus.com/en/airlines/products/ancillary-dynamic-pricing .

Andreev, H., Kosmas, P., Livieratos, A., Theocharous, A., & Zopiatis, A. (2025). Destination (un)known: Auditing bias and fairness in LLM-based travel recommendations. *AI, 6*(9), 236.

Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., & Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis, 31*(3), 337–351.

Ashford, N., & Benchemam, M. (1987). *Passengers' choice of airport: An application of the multinomial logit model', transportation research record, 1147*. Washington, D.C: National Research Council.

Balcombe, K., Fraser, I., & Harris, L. (2009). Consumer willingness to pay for in-flight service and comfort levels: A choice experiment. *Journal of Air Transport Management, 15*(5), 221–226. https://doi.org/10.1016/j.jairtraman.2008.12.005.

Beck, T., Schuff, H., Lauscher, A., & Gurevych, I. (2024). Sensitivity, performance, robustness: Deconstructing the effect of sociodemographic prompting. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: 1* (pp. 2589–2615). Long Papers.

Bergantino, A., Capurso, M., & Hess, S. (2020). Modelling regional accessibility to airports using discrete choice models: An application to a system of regional airports. *Transportation Research Part A: Policy and Practice, 132*. https://doi.org/10.1016/j.tra.2019.12.012.

Bhat, C. (1997). An endogenous segmentation mode choice model with an application to intercity travel. *Transportation Science, 31*, 34–48. https://doi.org/10.1287/trsc.31.1.34.

Bliemer, M., & Rose, J. (2024). Designing and conducting stated choice experiments. In S. Hess, & A. Daly (Eds.), *Handbook of choice modelling* (pp. 172–205). Cheltenham: Edward Elgar Publishing.

Boom Supersonic.. *Airlines and passengers* Available at Accessed: 15 January 2025 https://boomsupersonic.com/airlines-and-passengers .

Brand, J. E., Dell'Acqua, F., & Maheswaran, D.. *Consumer behavior and AI: Comparing human and GPT survey responses. Harvard Business School Working Paper No. 23-062* Available at https://www.hbs.edu/ris/Publication%20Files/23-062_ed720ebc-ec4d-4bc3-a6ba-bad8cfbd9d51.pdf.

Brown, T., et al., (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature News, 538*(7623), 20.

Caussade, S., Ortúzar, J., Rizzi, L., & Hensher, D. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B: Methodological, 39*(7), 621–640. https://doi.org/10.1016/j.trb.2004.07.006.

Chen, Y., Liu, T., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences, 120*(51), Article e2316205120.

Cho, W., & Dresner, M. (2018). The impact of a baggage fee waiver on airline choice: Evidence from the Washington-Baltimore region. *Transportation Research Part A, 112*, 4–17.

Cirium. (2024). The on-time performance review 2023. *Cirium Ascend Consultancy*. Available at https://assets.fta.cirium.com/wp-content/uploads/2024/01/02023854/Cirium-OTP-2023-Annual-Report1-prmc.pdf .

Crosby, P., Thompson, D., & Best, R. (2024). Air travellers' attitudes towards carbon emissions: Evidence from the Google Flights interface. *Journal of Sustainable Tourism*, 1–24. https://doi.org/10.1080/09669582.2024.2412161.

d-carbonize.. *Carbon footprint of a flight: What is its impact?* Available at https://d-carbonize.eu/carbon-footprint/plane-flight/.

Dadoun, A., Defoin-Platel, M., Fiig, T., Landra, C., & Troncy, R. (2021). How recommender systems can transform airline offer construction and retailing. *Journal of Revenue and Pricing Management, 20*(3), 301–315. https://doi.org/10.1057/s41272-021-00313-2.

Daily Mail. (2023). Would you pay extra to travel on a child-free flight? Airline divides opinion as it launches adults-only section where babies and kids are banned from the rest of the aircraft. *DailyMail*. 31 August. Available at https://www.dailymail.co.uk/travel/article-12464581/Would-pay-extra-travel-child-free-flight-Airline-divides-opinion-launches-adults-section-babies-kids-BANNED-rest-aircraft.html .

Dean, M., & Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics, 98*, 524–534.

Dominguez-Olmedo, R., Hardt, M. & Mendler-Dünner, C. (2023) Questioning the survey responses of large language models. arXiv preprint

Eigner, E., & Händler, T.. *Determinants of LLM-assisted decision-making* arXiv. Available at https://arxiv.org/html/2402.17385v1.

Filippas, A., Horton, J., & Manning, D. (2024). Large language models as simulated economic agents: What can we learn from Homo Silicus? In *Proceedings of the 25th ACM Conference on Economics and Computation (EC '24)* (pp. 614–615). Association for Computing Machinery. https://doi.org/10.1145/3670865.3673513.

Fukushi, M., Delgado, F., & Raveau, S. (2024). Discrete choice model applications for air passenger airport choices: A literature review. *Transportation Research Record*. https://doi.org/10.1177/03611981241263809.

Gao, Y., Lee, D., Burtch, G. & Fazelpour, S. (2024) Take caution in using LLMs as human surrogates: Scylla Ex Machina. arXiv preprint arXiv:2410.19599.

Garrow, L. A., Jones, S. P., & Parker, R. A. (2007). How much airline customers are willing to pay: An analysis of price sensitivity in online distribution channels. *Journal of Revenue and Pricing Management, 5*, 271–290. https://doi.org/10.1057/palgrave.rpm.5160052.

Gu, Z., Wang, Q., & Han, S.. *Alignment revisited: Are large language models consistent in stated and revealed preferences?* arXiv preprint arXiv:2506.00751v1 [cs.AI]. Available at https://arxiv.org/html/2506.00751v1.

Guo, T. et al. (2024) Large language model based multi-agents: A survey of progress and challenges. arXiv preprint arXiv:2402.01680.

Hagendorff, T., Fabi, S., & Kosinski, M. (2023). Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science, 3*(10), 833–838.

Haghani, M., Bliemer, M. C. J., Rose, J. M., Oppewal, H., & Lancsar, E. (2021). Hypothetical bias in stated choice experiments: Part I. Macro-scale analysis of literature and integrative synthesis of empirical evidence from applied economics, experimental psychology and neuroimaging. *Journal of Choice Modelling, 41*, Article 100309. https://doi.org/10.1016/j.jocm.2021.100309.

Hartmann, W. R., & Viard, V. B. (2008). Do frequency reward programs create switching costs? A dynamic structural analysis of demand in a reward program. *Quantitative Marketing and Economics, 6*, 109–137. https://doi.org/10.1007/s11129-007-9035-3.

Hensher, D. (2006a). Towards a practical method to establish comparable values of travel time savings from stated choice experiments with differing design dimensions. *Transportation Research Part A: Policy and Practice, 40*(10), 829–840.

Hensher, D. (2006b). Revealing differences in willingness to pay due to the dimensionality of stated choice designs: An initial assessment. *Environmental and Resource Economics, 34*(1), 7–44.

Hensher, D. (2006c). How do respondents process stated choice experiments? Attribute consideration under varying information load. *Journal of Applied Econometrics, 21*, 861–878. https://doi.org/10.1002/jae.877.

Hess, S. (2014). Impact of unimportant attributes in stated choice surveys. *European Journal of Transport and Infrastructure Research, 14*(4), 349–361.

Hess, S., & Adler, T. (2011). An analysis of trends in air travel behaviour using four related SP datasets collected between 2000 and 2005. *Journal of Air Transport Management, 17*, 244–248. https://doi.org/10.1016/j.jairtraman.2010.11.003.

Hess, S., & Polak, J. W. (2005). Accounting for random taste heterogeneity in airport choice modeling. *Transportation Research Record, 1915*(1), 36–43. https://doi.org/10.1177/0361198105191500105.

Hess, S., Ryley, T., Davison, L., & Adler, T. (2011). Improving the quality of demand forecasts through cross nested logit: A stated choice case study of airport, airline and access mode choice. *Transportmetrica A: Transport Science, 9*(4), 358–384. https://doi.org/10.1080/18128602.2011.577758.

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

Houtman, M., & Maks, J. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve Methoden, 19*, 89–104.

Kazemi, S., Gerhardt, G., Katz, J., Kuria, C.I. & Pan, E. (2024) 'Cultural fidelity in large-language models: An evaluation of online language resources as a driver of model performance in value representation', arXiv preprint arXiv:2410.10489.

Ke, L., Tong, S., Cheng, P., & Peng, K. (2025). Exploring the frontiers of LLMs in psychological applications: A comprehensive review. *Artificial Intelligence Review, 58*, 305. https://doi.org/10.1007/s10462-025-11297-5.

Kim, S., & Mokhtarian, P. (2023). Finite mixture (or latent class) modeling in transportation: Trends, usage, potential, and future directions. *Transportation Research Part B: Methodological, 172*, 134–173.

Kitadai, A., Fukasawa, Y. & Nishino, N. (2025) 'Bias-adjusted LLM agents for human-like decision-making via behavioral economics', arXiv preprint arXiv:2508.18600.

International Council on Clean Transportation (ICCT).. *CO_2 emissions from commercial aviation: 2013, 2018, and 2019* Available at https://theicct.org/publication/co2-emissions-from-commercial-aviation-2013-2018-and-2019/.

Liu, T., Li, M., & Yin, Y. (2024). Can large language models capture human travel behavior? Evidence and insights on mode choice. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4937575.

Liu, T., Yang, J., & Yin, Y. (2025). Toward LLM-agent-based modeling of transportation systems: A conceptual framework. *Artificial Intelligence for Transportation, 1*, Article 100001. https://doi.org/10.1016/j.ait.2025.100001.

Lu, J.G., Song, L.L. & Zhang, L.D. (2025) 'Cultural tendencies in generative AI', Nature Human Behaviour, advance online publication. https://doi.org/10.1038/s41562-025-02242-1

Ludwig, J., Mullainathan, S. & Rambachan, A. (2024) 'Large language models: An applied econometric framework', arXiv preprint arXiv:2412.07031, revised Jan 2025.

McKinsey.. *Chinese tourists: Dispelling the myths* September. Available at https://www.mckinsey.com/i/media/mckinsey/industries/travel%20transport%20and%20logistics/our%20insights/huanying%20to%20the%20new%20chinese%20traveler/chinese-tourists-dispelling-the-myths.pdf.

Manning, B., Zhu, K. & Horton, J. (2024) 'Automated social science: Language models as scientist and subjects', arXiv preprint arXiv:2404.11794.

McFadden, D. (1974). The measurement of urban travel demand. *Journal of Public Economics, 3*(4), 303–328. https://doi.org/10.1016/0047-2727(74)90003-6.

Mei, Q., Xie, Y., Yuan, W., & Jackson, M. (2024). A turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences, 121*(9), Article e2313925121.

Miranda, F., & Balbi, P. (2025). Simulating public opinion: Comparing distributional and individual-level predictions from LLMs and Random forests. *Entropy, 27*(9), 923.

Mo, B., Xu, H., Zhuang, D., Ma, R., Guo, X. & Zhao, J. (2023) 'Large language models for travel behavior prediction', arXiv preprint arXiv:2312.00819.

Mukherjee, A. & Chang, H. (2024) 'Heuristic reasoning in AI: Instrumental use and mimetic absorption', arXiv preprint arXiv:2403.09404.

Nguyen, T. N., Jamale, K., & Gonzalez, C. (2024). Predicting and understanding human action decisions: Insights from large language models and cognitive instance-based learning. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing: 12* (pp. 126–136). https://doi.org/10.1609/hcomp.v12i1.31607.

Nie, T., Sun, J., & Ma, W. (2025). Exploring the roles of large language models in reshaping transportation systems: A survey, framework, and roadmap. *Artificial Intelligence for Transportation, 1*, Article 100003. https://doi.org/10.1016/j.ait.2025.100003.

Nikolenko, N. (2025). Transformers at a crossroads: From narrow specialization to new paradigms. *Medium*. Available at https://nikitanikolen.medium.com/transformers-at-a-crossroads-from-narrow-specialization-to-new-paradigms-47b447aec453 .

Pels, E., Nijkamp, P., & Rietveld, P. (2001). Airport and airline choice in a multiple airport region: An empirical analysis for the San Francisco Bay Area. *Regional Studies, 35*(1), 1–9.

Pels, E., Nijkamp, P., & Rietveld, P. (2003). Access to and competition between airports: A case study for the San Francisco Bay area. *Transportation Research Part A: Policy and Practice, 37*, 71–83. https://doi.org/10.1016/S0965-8564(02)00007-1.

Penn, J. M., & Hu, W. (2018). Understanding hypothetical bias: An enhanced meta-analysis. *American Journal of Agricultural Economics, 100*(4), 1186–1206.

Rose, J., Hensher, D., & Greene, W. (2004). *Accounting for exogenous information on decision maker processing strategies in models of discrete choice: Attribute exclusion strategies*. Institute of Transport and Logistics Studies.

Sabre.. *LATAM adopts Sabre's intelligent solution Air Price IQ* Available at https://www.sabre.com/insights/releases/latam-adopts-sabres-intelligent-solution-air-price-iq/.

Schmidt, J., & Bijmolt, T. H. (2020). Accurately measuring willingness to pay for consumer goods: A meta-analysis of the hypothetical bias. *Journal of the Academy of Marketing Science, 48*, 499–518.

Scotti, D., & Dresner, M. (2015). The impact of baggage fees on passenger demand on US air routes. *Transport Policy, 43*, 4–10.

Seelhorst, M., & Liu, Y. (2015). Latent air travel preferences: Understanding the role of frequent flyer programs on itinerary choice. *Transportation Research Part A: Policy and Practice, 80*, 49–61. https://doi.org/10.1016/j.tra.2015.07.007.

Segerer, R.. *Cultural value alignment in large language models: A prompt-based analysis of Schwartz values in Gemini, ChatGPT, and DeepSeek* arXiv preprint arXiv:2505.17112. Available at https://arxiv.org/abs/2505.17112.

Sfeir, G., Nova, G., Hess, S. & van Cranenburgh, S. (2025) 'Can large language models assist choice modelling? Insights into prompting strategies and current models' capabilities', arXiv preprint arXiv:2507.21790.

SFO.. *2018 SFO customer Survey* Available at https://data.sfgov.org/Transportation/2018-SFO-Customer-Survey/3w8r-nuxp/about_data.

Simon, H. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*(1), 99–118.

Singapore Airlines.. *Singapore Airlines and Salesforce collaborate on AI-powered customer service applications* [Online], 12 March. Available at https://www.singaporeair.com/en_UK/gb/corporate/newsroom/press-release/2025/January-March/singapore-airlines-and-salesforce-collaborate-on-ai-powered-cust/.

Strachan, J. W. A., Albergo, D., Borghini, G., et al., (2024). Testing theory of mind in large language models and humans. *Nature Human Behaviour, 8*, 1285–1295. https://doi.org/10.1038/s41562-024-01882-z.

Sun, H. & Zhang, J. (2025) 'From model choice to model belief: Establishing a new measure for LLM-based research', MIT Sloan Working Paper (SSRN 5399546).

Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus, 3*(9), 1–9. https://doi.org/10.1093/pnasnexus/pgae346.

Wandelt, S., Zheng, C., Wang, S., Liu, Y., & Sun, X. (2024). Large language models for intelligent transportation: A review of the state of the art and challenges. *Applied Sciences, 14*(17), 7455. https://doi.org/10.3390/app14177455.

Wang, L., Duan, P., He, Z., Lyu, C., Chen, X., Zheng, N., Yao, L., & Ma, Z. (2025a). Agentic large language models for day-to-day route choices. *Transportation Research Part C: Emerging Technologies, 180*, Article 105307. https://doi.org/10.1016/j.trc.2025.105307.

Wang, A., Morgenstern, J. & Dickerson, J.P. (2025b) 'Large language models that replace human participants can harmfully misportray and flatten identity groups', arXiv preprint arXiv:2402.01908.

Warburg, V., Bhat, C., & Adler, T. (2006). Modeling demographic and unobserved heterogeneity in air passengers' sensitivity to service attributes in itinerary choice. *Transportation Research Record, 1951*(1), 7–16. https://doi.org/10.1177/0361198106195100102.

Wen, K., Zhang, H., Lin, H. & Zhang, J. (2024) 'From sparse dependence to sparse attention: Unveiling how chain-of-thought enhances transformer sample efficiency', arXiv preprint arXiv:2410.05459.

Wendt, N., Kipperberg, G., & Lindhjem, H. (2024). Consumer willingness to pay for emission reduction in air travel: A meta-analysis. *Transportation Research Part D: Transport and Environment, 134*, Article 104347. https://doi.org/10.1016/j.trd.2024.104347.

Windle, R., & Dresner, M. (1995). Airport choice in a multi-airport region. *Journal of Transportation Engineering, 121*(4), 332–338.

Xian, Y., Lampert, C., Schiele, B., & Akata, Z. (2019). Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis & Machine Intelligence, 41*(9), 2251–2265. https://doi.org/10.1109/TPAMI.2018.2857768.

Yan, Y., Liu, T. & Yin, Y. (2025) 'Valuing time in silicon: Can large language models replicate human value of travel time', arXiv preprint arXiv:2507.22244.

Yoo, H. I. (2020). lclogit2: An enhanced command to fit latent class conditional logit models. *The Stata Journal, 20*(2), 405–425. https://doi.org/10.1177/1536867X20931003.

Zhou, L., Schellaert, W., Martínez-Plumed, F., et al., (2024). Larger and more instructable language models become less reliable. *Nature, 634*, 61–68. https://doi.org/10.1038/s41586-024-07930-y.

https://dspace.org/

Can large language models mimic airline passenger preferences?

Voltes-Dorta, Augusto; Suau-Sanchez, Pere

2025-11
Elsevier
https://doi.org/10.1016/j.ait.2025.100034
Voltes-Dorta A, Suau-Sanchez P. (2025) Can large language models mimic airline passenger preferences?. Artificial Intelligence for Transportation, Volume 3-4, November 2025, Article number 100034

Downloaded from a DSpace repository. https://dspace.org/