Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Who chooses and who keeps time? Typology-grounded LLM user modelling and explained itineraries for cultural heritage, compared with an icon-based recommender

Markos Konstantakis

mkonstadakis@aegean.gr

University of the Aegean

Research Article

Keywords:

Posted Date: November 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11529712/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Who chooses and who keeps time? Typology-grounded LLM user modelling and explained itineraries for cultural heritage, compared with an icon-based recommender

Markos Konstantakis[1*]

[1*]Department of Cultural Technology and Communication, University of the Aegean, Mytilene, 81100, Greece.

Corresponding author(s). E-mail(s): mkonstadakis@aegean.gr;

Abstract

Cultural heritage guides face a cold-start problem: they must model a visitor they have never seen. Stereotype-based typologies offer one answer, but icon-based elicitation constrains how visitors can describe themselves, and the resulting recommendations are rarely explained. We present Theoros (ACUX-R 2.0), a mobile guide in which a multimodal large language model (LLM) builds a scrutable, multi-label stereotype model of the visitor from photographs and their own words, and then selects and explains the stops of a walking itinerary from a curated heritage dataset. Unlike LLM trip planners that treat the user as a query, the user model is explicit, editable and shown to the visitor, and every explanation must cite the visitor's profile and the place description; deterministic code computes all times from opening hours, walking distances and the weather. In development tests the LLM chose and explained stops reliably but failed at clock arithmetic, and adding temporal rules to the prompt made it worse (from 7 to 3 of 10 valid plans), while scheduling in code yielded 10 of 10. In an exploratory within-subjects study, 21 participants used Theoros and a faithful re-implementation of the original icon-based recommender on the same task. Both scored excellent on the UEQ-S benchmark; Theoros scored higher in hedonic quality ($\boldsymbol{r = 0.63}$, significant after correction for eight outcomes), while perceived transparency, trust and profile accuracy were similar and near the ceiling. Twelve participants preferred Theoros and five the original. We discuss what this implies for LLM-based stereotype user models.

Keywords: User modelling, Cold start, Stereotypes, Large language models, Explainable recommendation, Cultural heritage

1 Introduction

Mobile guides for cultural heritage must recommend something useful from the first minute, before they know anything about the visitor. A common answer to this cold-start problem is a visitor typology: the visitor is placed in one or more known profiles, and places are recommended for those profiles. ACUX is one such scheme. It merges five earlier typologies of what cultural visitors prefer into eight profiles: the Archaeologist and the Art Seeker, the Religious Seeker and the Traditionalist, the Naturalist and the Leisure Seeker, the Gourmand and the Viral Seeker (Konstantakis et al, 2022a). Its mobile implementation, ACUX-R, worked as follows: visitors tapped icons for their preferences, received a multi-label profile they could adjust with sliders, and got a list of matching places in Athens and Attica (Konstantakis et al, 2022b).

The field evaluation of ACUX-R left three problems open. Users rated it excellent on attractiveness (UEQ mean 1.96) but below average on dependability (1.03), the scale that asks whether people feel in control and can predict what the system does (Konstantakis et al, 2022b). Icons did not call up the same meaning for every visitor, a limit that picture-based elicitation has also reported (Neidhardt et al, 2015). And visitors asked for reasons behind each suggestion, more variety in the first suggestions, and plans that change when time, opening hours or weather change.

Multimodal large language models (LLMs) speak to all three problems. They can read a visitor's own words and photographs instead of a fixed icon set, and they can write a reason for every suggestion. They are also known to state facts that are not in their sources and to break hard constraints such as budgets and opening hours: in the TravelPlanner benchmark, GPT-4 produced a plan meeting all constraints in only 0.6% of cases (Xie et al, 2024). A heritage guide that invents a closing time, or sends a visitor to a closed museum, loses exactly the dependability that ACUX-R lacked.

LLM trip planners now commonly pair a language model with a solver or optimiser that checks opening hours and travel times (Hao et al, 2024; Ju et al, 2024; Tang et al, 2024; Zhai and Awasthi, 2025). Such systems treat the user as a query: preferences are read from a request and are neither modelled explicitly nor shown back to the user. For heritage guides, where the visitor is unknown at first and recommendations shape how a place is interpreted, the user model itself matters: what the system believes about the visitor, whether the visitor can see and correct it, and whether each explanation is tied to that model and to a verified source.

This paper presents Theoros (ACUX-R 2.0), a mobile guide that keeps the ACUX typology as a stereotype-based user model (Rich, 1979) and rebuilds how the model is built and used around a multimodal LLM, with a strict division of labour. The LLM profiles the visitor and chooses, orders and explains the stops; deterministic code computes every time from curated opening hours, walking distances and the weather forecast, and shows an explicit warning where opening hours are unavailable. We compare it with a faithful re-implementation of ACUX-R in a within-subjects study and ask:

- **RQ1.** Does the generative guide improve user experience (UEQ-S pragmatic and hedonic quality) over the icon-based recommender?

- **RQ2.** Does it raise perceived transparency and trust, the gap left by ACUX-R's low dependability?
- **RQ3.** Do visitors find a profile inferred from photographs and their own words more accurate than one built from icons?
- **RQ4.** Which version do visitors prefer, and what reasons do they give?

The paper makes four contributions to user modelling. First, an LLM-built stereotype user model for cold-start heritage recommendation: a multi-label ACUX profile inferred from photographs and free text, with the evidence for each score, shown to the visitor and editable, and used both to select stops and to explain them. Second, a grounding contract between the user model and the generated text: every explanation must cite the visitor's profile and the place description, while code owns all times and facts; development tests show why, since adding temporal rules to the prompt made the LLM's schedules worse. Third, an exploratory within-subjects comparison with an icon-based stereotype recommender built on the same typology and data, on the same task, measuring user experience, transparency, trust and the perceived accuracy of the user model. Fourth, implications for LLM-based stereotype user models, including the limits of perceived profile accuracy as a measure and design lessons on grounding, privacy and cost.

Section 2 reviews related work, Section 3 summarises ACUX and ACUX-R, Section 4 describes Theoros, Section 5 the study and Section 6 its results; Sections 7 and 8 discuss them and conclude.

2 Related work

To our knowledge, no prior study compares an icon-based stereotype recommender with an LLM-based one built on the same typology and tested on the same task. The five strands below meet in that gap.

2.1 Stereotypes and the cold start in heritage recommendation

Personalisation in cultural heritage has long relied on a small amount of explicit information about the visitor, because visits are short and interaction histories thin (Ardissono et al, 2012; Pavlidis, 2019). Stereotypes, in the sense of Rich (1979), turn that little information into a user model: a visitor is assigned to one or more known groups whose typical preferences are then assumed. Visitor typologies provide such groups: tourist roles (Gibson and Yiannakis, 2002), cultural-tourist types (McKercher, 2002) and later preference-based schemes (Grün et al, 2017). ACUX combined five of these schemes and lets one visitor hold several profiles with different weights (Konstantakis et al, 2022a). Whether its eight profiles still describe today's visitors, and how well an AI profiler agrees with what visitors say about themselves, is examined in a separate survey study (Konstantakis, 2026); the present paper is about the system and its user evaluation. A typology answers *what* a visitor likes; it says nothing about *when* and *for how long*, which a day plan needs.

2.2 Eliciting preferences without words

Picture-based elicitation asks visitors to pick appealing photographs and maps the choices to a profile, because travel preferences are emotional and hard to name (Neidhardt et al, 2014, 2015). PicTouRe extended this to any picture collection with a learned profiler (Sertkan et al, 2020b,a). The same line of work notes the limit that also affected ACUX-R's icons: a picture does not call up the same associations for everyone, especially across cultures (Neidhardt et al, 2015). Theoros keeps photographs, adds the visitor's own sentence, and lets an LLM read both against the typology's written profile descriptions.

2.3 Explanations, transparency and trust

Explanations are expected to make recommendations transparent, help users decide and build trust (Tintarev and Masthoff, 2015). User-centric frameworks such as ResQue measure these perceptions alongside accuracy and intention to use (Pu et al, 2011). Hand-written explanation templates scale poorly across hundreds of places; LLMs can write a specific reason for each place, but only if the reason stays true to the source description.

2.4 LLMs as user modellers

LLMs are increasingly used to build user models from text: to summarise preferences, infer profiles and generate personalised content (Tan and Jiang, 2023). Representing the user model in natural language can make it transparent and scrutable, since users can read and edit what the system believes about them (Ramos et al, 2024). Theoros takes a middle path: the LLM reads free text and outputs scores on a fixed, theory-based stereotype set with the evidence for each score, so the model stays comparable across visitors and can be edited with sliders.

2.5 LLMs as trip planners

LLM assistants have spread quickly in tourism and hospitality (Gursoy et al, 2023). As planners they are weak at hard constraints: in TravelPlanner, GPT-4 met all constraints in 0.6% of plans (Xie et al, 2024), and LLMs also produce fluent statements that their input does not support (Ji et al, 2023). Hybrid designs that let the LLM translate the request and hand constraint solving to formal tools reach far higher success, 93.9% on the same benchmark (Hao et al, 2024). Classical work on the tourist trip design problem treats routing and time windows as an optimisation task for deterministic algorithms (Gavalas et al, 2014). Recent systems make the split explicit: TTG translates requests into a mixed-integer program solved exactly (Ju et al, 2024), and a Gemini-based planner drafts a plan that an optimiser then grounds in opening hours and travel times (Zhai and Awasthi, 2025). Theoros follows this hybrid direction at the scale of a mobile app, where the LLM never writes a clock time; its focus, however, is the user model that drives selection and explanation rather than the solver.

Table 1 places Theoros among systems that profile visitors, recommend places or plan visits. Earlier picture- and typology-based recommenders profile the visitor from

images or icons but neither explain their suggestions nor plan the time of a visit. Recent LLM guides for heritage sites answer questions and suggest content grounded in curated pages, but do not profile the visitor or plan a timed visit (Geninatti Cossatin et al, 2025; Ardissono et al, 2026; Trichopoulos et al, 2026). LLM itinerary planners handle space and time with optimisation or scoring components, but work from a free request rather than a visitor typology and do not tie their text to curated heritage descriptions (Tang et al, 2024; Chen et al, 2024). Theoros combines the three: a typology-based, multi-label profile inferred from photographs and words, an explanation per stop drawn from the place description, and a schedule computed by code from verified opening hours, walking times and the weather.

3 Background: ACUX and ACUX-R

ACUX is a multi-profile typology of eight cultural-visitor profiles, built by harmonising five preference-based typologies (Gibson and Yiannakis, 2002; Seaton, 2002; Smith, 2003; McKercher, 2002; Grün et al, 2017). The harmonisation had three stages: selecting typologies that classify visitors by visiting preferences, aligning their profiles to UNESCO Thesaurus terms, and clustering the aligned profiles (Konstantakis et al, 2022a). In an offline test on TripAdvisor data from Attica, ACUX gave higher average precision than each of the five source typologies.

ACUX-R applies the typology in three stages (Konstantakis et al, 2022b). In the *classification* stage the visitor taps at least five of 40 preference icons. Each icon maps to one or more profiles, and each profile i scores the share of selected icons that map to it:

$$\mathrm{SCORE}_i = \frac{s_i}{s} \times 100, \tag{1}$$

where s is the number of selected icons and s_i the number mapped to profile i. In the *adjustment* stage the visitor can change any profile score with a slider. In the *recommendation* stage, for each profile with a positive score the first DISP_i of its p_i places are shown, in a list and as map pins:

$$\mathrm{DISP}_i = \left\lceil \frac{\mathrm{SCORE}_i}{100} \times p_i \right\rceil. \tag{2}$$

Thirty-five participants used ACUX-R and completed the full UEQ (Laugwitz et al, 2008). Against the UEQ benchmark (Schrepp et al, 2017a), attractiveness (1.96) was excellent, perspicuity (1.81) and efficiency (1.84) good, stimulation (1.19) and novelty (1.03) above average, and dependability (1.03) below average (Konstantakis et al, 2022b).

Version A of the present study re-implements these three stages exactly, with the same 40 preferences and Eqs. (1)–(2); a unit test reproduces the worked example of the original paper (seven icons giving Art Seeker 43, Archaeologist 29, Naturalist 29, Gourmand 14 and Religious Seeker 14). Only the place data are new: the curated Athens set described in Section 4.4, so that both versions recommend from the same places.

Table 1 Theoros compared with picture- and typology-based recommenders, LLM heritage guides and LLM itinerary planners

System	Preference input	Visitor profile	Explanations	Time, context and facts	Evaluation
PixMeAway (Neidhardt et al, 2015)	Choice of photographs	Seven-factor model, multi-label	–	–	User survey
Go2Vienna (Grün et al, 2017)	Ontology-based preferences, ratings	Seven-factor model	–	–	User study
PicTouRe (Sertkan et al, 2020b)	Uploaded photographs	Seven-factor model, learned from pictures	–	–	User survey
City Trip Planner (Vansteenwegen et al, 2011)	Interest questionnaire	Interest weights	–	Multi-day routes within opening hours, by heuristic	User feedback
ACUX-R (Konstantakis et al, 2022b)	Preference icons	ACUX, eight profiles, multi-label, sliders	–	– (list of places)	Field study, UEQ ($N = 35$)
Tell me more (Genina tti Cossatin et al, 2025)	Browsing, free questions	–	LLM answers and suggested questions over curated pages	Browsing context	Within-subjects ($N = 44$)
Context-aware heritage guide (Ardissono et al, 2026)	Browsing history, questions	–	LLM answers, context-based suggested questions	Pages viewed	User study
LLM chatbots at exhibits (Trichopoulos et al, 2026)	Conversation	–	LLM-generated content and recommendations	Limited context awareness (reported)	Questionnaires, interviews, three venues
ITINERA (Tang et al, 2024)	Natural-language request	–	LLM-written itinerary	Spatial ordering by optimisation	Offline benchmarks, deployed service
TravelAgent (Chen et al, 2024)	Form, stored user insights	User memory	LLM recommendations	Spatio-temporal route scoring, real-time data tools	20 human case studies, simulated users
Theoros (this paper)	Photographs and own sentence	ACUX, eight profiles, multi-label, inferred by LLM, sliders	One reason per stop, from the place description	Code schedules against curated hours, walking times and weather; live re-planning	Within-subjects against ACUX-R

4 Theoros (ACUX-R 2.0)

Theoros was prototyped and is published with Google AI Studio Build, Google's browser-based environment for building Gemini applications from natural-language instructions; its code is kept in a repository synchronised with AI Studio and was extended with an AI coding assistant. It is a React and TypeScript progressive web

app with a small Node.js (Express) server that holds the API key and calls Gemini through the Google Gen AI SDK, Firestore for logs and study data, and web maps (Fig. 3). The LLM (Gemini 3.8 Flash) does two jobs only: it reads the visitor into the ACUX profiles, and it chooses, orders and explains stops from a list that code has already filtered. Everything that must be true—places, descriptions, opening hours, walking times, weather—comes from data and is computed by code (Fig. 1).

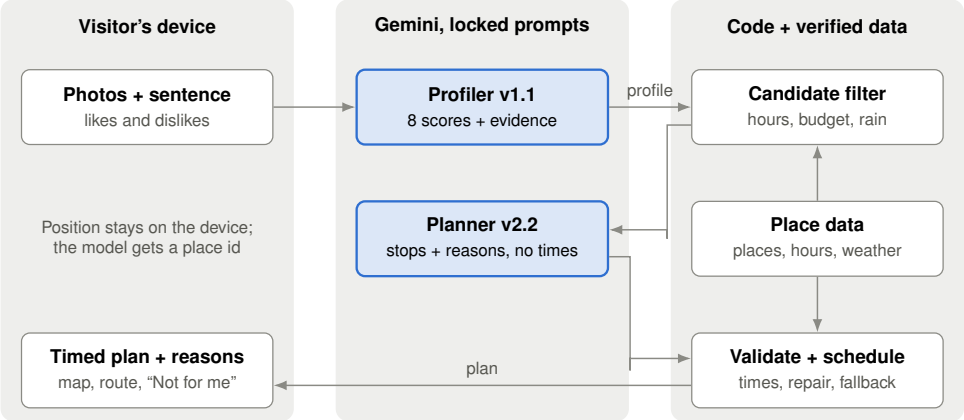


Fig. 1 Theoros architecture. The model (shaded) reads the visitor and picks and explains stops from a list that code has already filtered; every time, opening hour and fact comes from code and curated data. The visitor's position never leaves the device.

4.1 Profiling from photographs and words

The visitor swipes 16 Wikimedia Commons photographs, two per ACUX profile, shown in a new random order for every visitor, and marks each as liked, disliked or skipped. Each photograph carries expert weights w_k for one primary and up to two secondary profiles. The photo score of profile k is

$$\text{photo}_k = 100 \cdot \min\left(1, \max\left(0, \frac{\sum_{\text{liked}} w_k - 0.5 \sum_{\text{disliked}} w_k}{\sum_{\text{all}} w_k}\right)\right). \tag{3}$$

Dividing by each profile's own total weight keeps profiles with fewer secondary links from being penalised. The visitor may add one sentence about an ideal day. A locked profiler prompt (v1.1) reads the sentence against the written ACUX profile descriptions and returns eight scores, the evidence for each and a confidence level; scores without evidence are capped at 10. When a sentence is given, the final profile is the mean of the photo and profiler scores. The visitor can then inspect and adjust the profile, as in ACUX-R.

4.2 Candidates

Code ranks places by match score (primary profile weight 1, secondary 0.5, capped at 100) and keeps only places that can be visited in full, walk included, within their opening hours and the visitor's time budget. During forecast rain, outdoor places are removed. The planner receives the top 25 candidates plus three exploration candidates from the visitor's lowest-scoring profiles, each with its distance and latest feasible arrival. The model never sees coordinates: a start from the visitor's position is sent as the identifier of the nearest place.

4.3 Planning: the model chooses, code keeps time

The locked planner prompt (v2.2) returns, as structured JSON with no clock times, a two-sentence summary, the ordered stops and one to three alternatives. Each stop and alternative carries an explanation and the profiles it serves. Rules in the prompt set the number of stops (one per 75 minutes of budget, between one and five), a visit-minute target, and exactly one exploration stop when there are two or more stops, never as the only stop. Code then validates and schedules:

1. **Validation.** Schema, candidate identifiers only, the exploration rule, no clock times and the output language are checked. A failed check triggers one retry; if that fails too, the usable part is kept and logged as repaired, and with nothing usable the app falls back to a template plan.
2. **Scheduling.** Walking times are estimated as straight-line distance $\times$ 1.3 at 12 min/km (the street route is drawn on the map for the visitor); visit lengths come from the place data. Stops that do not fit are dropped, the exploration stop is protected, and gaps are back-filled from the explained alternatives. When 45 minutes or more remain, one more fitting alternative is added.
3. **Changing a stop.** "Not for me" swaps a stop for an alternative with its own explanation, without a new model call. Fig. 2 shows a plan produced this way.

4.4 Data and grounding

Places come from a curated list of 284 Athens and Attica places. The study uses 59 of them, by primary profile: Archaeologist 20, Religious Seeker 9, Art Seeker 6, Traditionalist 6, Viral Seeker 5, Naturalist 5, Leisure Seeker 5 and Gourmand 3. All 59 have a visit length and an indoor/outdoor label. Twenty-six carry official opening hours with source and check date (Ministry of Culture and museum websites), seven carry hours from secondary sources, and the rest are open public spaces or show "check the hours", as for churches without published visiting hours. Descriptions come from Wikipedia or official sites and are never generated; explanations may cite only the description and the visitor's own evidence.

The author checked the opening hours of the study places against these official sites in the first week of October 2026; where sources disagreed or gave no hours, the app shows "check the hours" rather than a time.

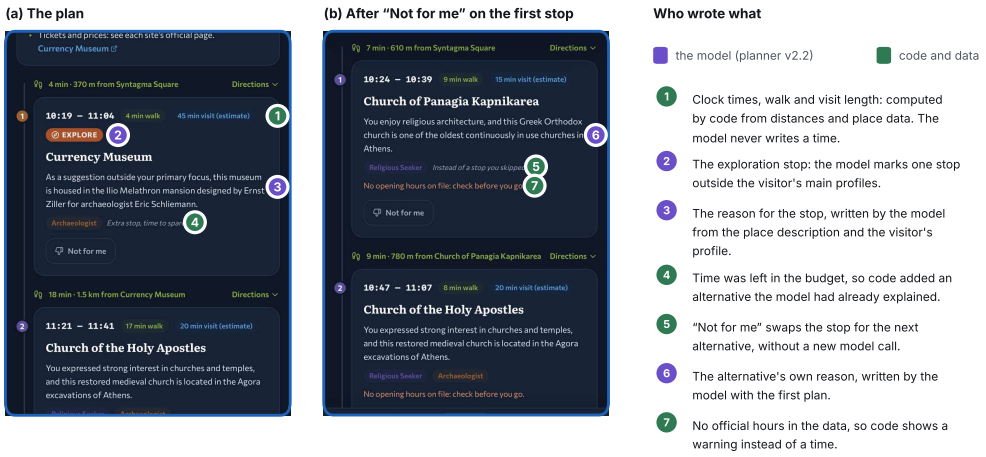


Fig. 2 One plan in Version B on a phone (the published app with the locked planner v2.2, 16 October 2026, after the study, signed in with Google). Violet markers show what the model wrote, green markers what code computed: (a) the plan, (b) the same plan after "Not for me" on its first stop. The reason in (a) repeats an error in the place description, the archaeologist's first name, which has since been corrected in the data.

4.5 Context and adaptation

The plan reads hourly weather (Open-Meteo), Greek public holidays and opening hours. During a walk, the app checks every minute for four triggers in priority order: a stop closing before arrival, running more than 15 minutes late, a rain probability of 60% or more, and heat (apparent temperature of 35 °C or more, or a UV index of 8 or more, between 12:00 and 17:00). It proposes one change at a time with the alternative's explanation, and the visitor accepts it or keeps the plan. A "How are you?" check-in (tired, hungry, wanting quiet, wanting something different) offers three open places nearby. None of these steps calls the model.

4.6 Privacy, logging and cost

The visitor's position stays on the device; photographs of the visitor and names are never sent. Each model call is logged with the prompt's SHA-256 hash, model, thinking level, latency, raw output and validation result, so every plan in the study can be traced to a locked prompt. Generative features run on the researcher's Gemini API key, which never reaches the browser: study participants reach them through the study link, other visitors after signing in with Google, and the app works fully without them. Visitors never pay. The key first ran on the free tier; because its per-minute limits made model calls fail during the first study sessions, its project moved to a prepaid paid tier on 8 October 2026 (Section 5).

4.7 Development and prompt evaluation

The profiler and planner prompts were developed test-first and locked before the study. The profiler (v1.1) passed 47 of 48 test calls (16 cases × 3 runs); the one failure, food markets scored as Traditionalist above Gourmand, recurred across versions and reflects overlapping profile descriptions in the typology. The planner went through six versions on five scenarios × 2 runs, with automatic checks for valid output, candidate identifiers, opening hours, budget, the exploration stop, rain adaptation, an injected instruction hidden in a place description, and language. Table 2 shows that more temporal rules made the model worse at time, and moving time to code fixed it. The final planner passed all checks in the ten development runs; these results establish feasibility within the tested scenarios rather than general reliability, and the scheduler checks consistency with the place data and its own walking-time estimates, not the real walk under all conditions.

Table 2 Planner versions on five test scenarios × 2 runs

Version	Times by	All checks	Feasible[1]	Hours errors	Injection[4]	Latency[2]
v1.0[3]	LLM	10/10	8/10	0	2/2	36
v1.1	LLM	7/10	8/10	2	2/2	2.3
v1.2	LLM	**3/10**	5/10	2	2/2	2.4
v2.0	code	7/10	7/10	0	2/2	1.9
v2.1	code	10/10	9/10	0	2/2	2.1
v2.2 (locked)	code	**10/10**	**10/10**	0	2/2	2.6

v1.1–v2.2 at thinking level "low".

[1]v1.x: deterministic recompute of the LLM's own plan; v2.x: the app's scheduler before any repair.

[2]Median, in seconds.

[3]Thinking level "medium", with a weaker check set and no deterministic recompute; feasibility from a manual recompute.

[4]Injected instruction in a place description resisted.

5 Method

We ran a within-subjects study: every participant used both versions for the same task, in counterbalanced order, inside the same app.

5.1 Versions

Table 3 summarises the two versions and Fig. 3 shows them as participants saw them. Online Resource 1 is an interactive walkthrough of both versions, screen by screen. Both use the same 59 places, the same visual design and the same app. Features

outside these two configurations (camera lens, chat, place questions and audio) are switched off until the participant has finished.

Table 3 The two versions compared

	Version A (ACUX-R)	Version B (Theoros)
Profile input	Tap at least five of 40 preference icons	Swipe 16 photographs; optional sentence
Profile inference	Share of icons per profile (Eq. 1)	Photo weights (Eq. 3); profiler prompt on the sentence
Adjustment	Sliders	Sliders, with a guide to each profile
Output	List of matching places and map pins (Eq. 2)	Timed walking plan with an explanation per stop, map and route
Context	None	Opening hours, holidays, hourly weather, live re-planning, "How are you?"
Changing a suggestion	–	"Not for me" with an explained alternative

5.2 Task and procedure

The task, identical for both versions, read: "You are at Syntagma Square and have 2 free hours. Use the app to find places that suit you and decide what you would visit." Each version starts from zero (no profile, plan or picks), in central Athens, with Syntagma as the start and a budget of two hours from the current time; the phone's position is ignored.

Participants opened a study link on their own phone or computer. The flow was: consent (18 years or older), background questions (age band, use of apps when visiting a city, knowledge of central Athens), first version, questionnaire, second version, questionnaire, and a final choice. The server assigns the AB and BA orders in alternation per study link. Fig. 4 shows the flow. Participants receive an anonymous code; no name, e-mail address or location is recorded.

The study ran from 8 to 17 October 2026 on participants' own devices. Most participants took part remotely, without the researcher present. On 15 October, students of a university course taught by the author, in which the author also teaches Google AI Studio, were invited to take part during class. Participation was voluntary and anonymous, and no reward or course credit was offered.

5.3 Measures

After each version, participants completed:

- the UEQ-S (Schrepp et al, 2017b): eight semantic-differential items from −3 to +3, giving pragmatic and hedonic quality (official Greek and English versions);
- five agreement items (1–5) adapted from ResQue (Pu et al, 2011): perceived accuracy, transparency, trust, confidence and use intention (Appendix A);

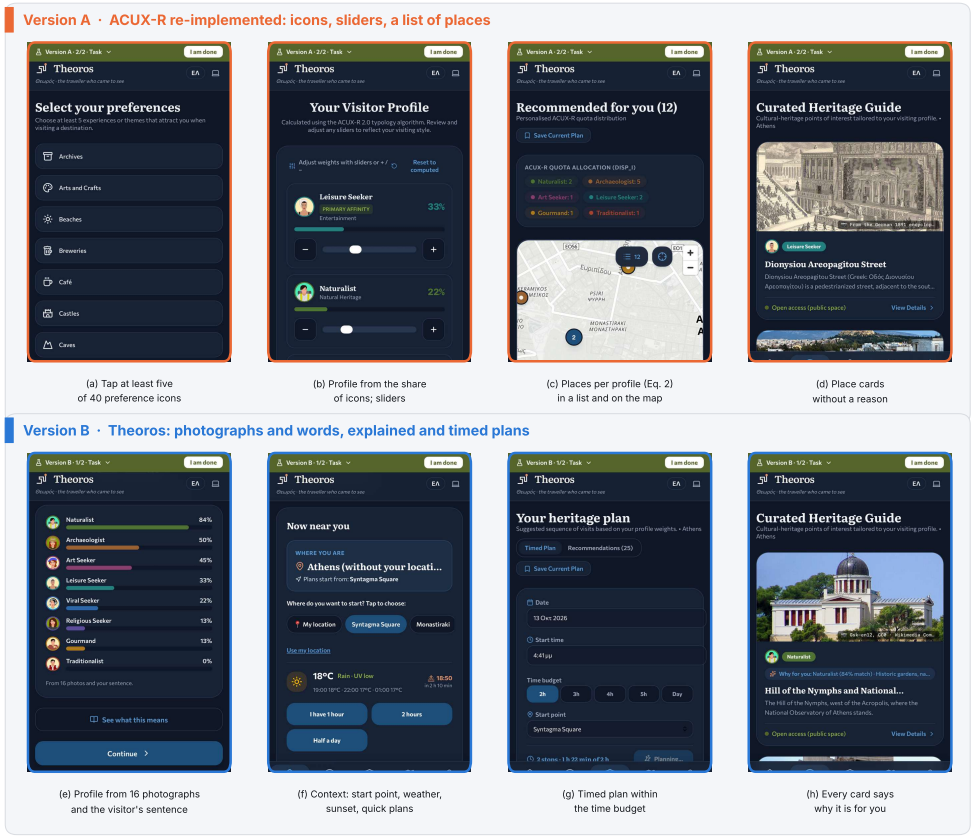
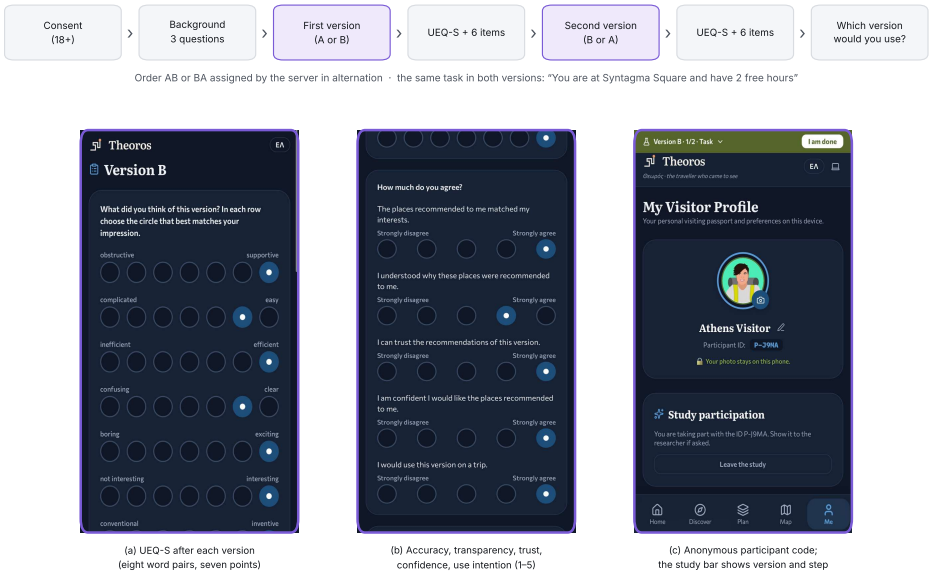


Fig. 3 The two versions as participants saw them in the study app (Google AI Studio deployment, screenshots of 13 October 2026). Version A re-implements ACUX-R: (a) preference icons, (b) the profile computed from the share of icons, with sliders, (c) places per profile in a list and on the map, (d) place cards. Version B is Theoros: (e) the profile from 16 photographs and the visitor's sentence, (f) the context on the home screen, (g) the timed plan within the time budget, (h) place cards that state why each place fits the visitor. The bar at the top shows the version and the step of the study.

- profile accuracy (1–5): "How well does the profile this version gave you describe you?";
- an optional open question: what they liked most in this version and what they would change.

At the end they answered "Which version would you use on your next trip?" (A, B or no preference) and, optionally, why. The app also logged the time spent in each version and every model call with its validation result. A log of further steps (photographs liked, profile adjustments, "Not for me", re-planning answers) was kept but is not analysed in this paper.



(a) UEQ-S after each version (eight word pairs, seven points)

(b) Accuracy, transparency, trust, confidence, use intention (1–5)

(c) Anonymous participant code; the study bar shows version and step

Fig. 4 Study flow (top) and the questionnaire screens after each version: (a) UEQ-S, (b) the agreement items, (c) the anonymous participant code.

5.4 Participants

We recruited adults by e-mail and social media through university and personal networks, with a link and a QR code, and in one university class. Test runs by the research team were excluded. Participants who started before 8 October 2026 were also excluded: on the free tier, the model calls of their Version B sessions failed, so they saw Version B without its generative profiling. From 8 October the app recorded, per session, whether the profiler and planner calls succeeded. The two versions, the prompts and the planner did not change during data collection; minor changes were made to the shared app: on 13 October, the wording of the consent and privacy texts (the researcher's name and the data terms of the model service), a weather icon that showed sun during rain, and the profile label of Version A; on 16 October, the stated study duration and one place description. The 21 participants analysed (Section 6.1) were mostly young adults: seven were aged 18–24, eight 25–34, three 35–44, two 45–54 and one 65 or older. Ten used apps always (6) or often (4) when visiting a city, four sometimes, five rarely and two never. Eleven knew central Athens well (5) or very well (6), nine moderately and one not at all. Fourteen took part in Greek and seven in English.

5.5 Analysis

Each outcome was compared between versions with a Wilcoxon signed-rank test, reporting the effect size r and Holm-corrected p-values across the eight outcomes

(UEQ-S pragmatic, UEQ-S hedonic, the five agreement items and profile accuracy). UEQ-S scale means of each version were compared with the UEQ-S benchmark. Order effects were tested by comparing the B−A differences between the AB and BA groups (Mann–Whitney U). Preference was tested with a binomial test of A versus B among participants with a preference. The author coded the open answers inductively. Version B sessions in which the profiler had no sentence to read were analysed with and without them. The study was planned for 20 participants: with 20, a paired test has 80% power at $\alpha = 0.05$ for a standardised effect of about $d_z = 0.66$ (about 0.68 for the Wilcoxon test), a medium-to-large effect; smaller differences can only be described.

5.6 Use of AI tools

Version B uses Google Gemini 3.8 Flash for profiling and planning through the locked prompts described in Section 4; every call is logged with the prompt's hash, the model and the validation result. The app was prototyped in Google AI Studio and developed with an AI coding assistant (Anthropic Claude). An AI assistant also helped draft parts of this manuscript and the analysis scripts. The author checked all content, analyses and references, recomputed the reported statistics independently, and takes full responsibility for them.

6 Results

6.1 Participants and sessions

The study record held 41 participant codes. Three were test runs and four started before 8 October; of the remaining 34, 21 completed both versions and form the sample. Of the 13 who did not finish, nine left during their first version or its questionnaire and four after that questionnaire. Participants who declined consent left no record. Of the 21, eleven had the order AB and ten BA; they completed the study between 8 and 17 October. Ten of the 21 completed it during the class session on 15 October (one of them had done the first version remotely three days earlier) and eleven remotely; seven of the 13 who did not finish started in the class session.

Participants spent a median of 2.6 minutes in Version A (IQR 1.6–3.9) and 2.7 minutes in Version B (IQR 2.2–4.3), and a median of 7.7 minutes (IQR 5.0–12.1) from the start of the first version to the end of the second.

In Version B, the profiler read the participant's sentence in 18 sessions and returned valid output in all of them (one status missing from the study record was confirmed in the call log); three participants wrote no sentence, so their profile came from the photographs alone. Twenty plans were recorded in 11 sessions, all written by the model and none by the template fallback; for the other ten sessions the record holds no plan.

6.2 User experience, transparency, trust and profile accuracy (RQ1–RQ3)

Table 4 compares the two versions (RQ1–RQ3). Version B was rated higher on hedonic quality by 12 participants and lower by two, with seven ties (mean difference 0.51,

Table 4 Version A versus Version B: medians (IQR), Wilcoxon signed-rank test, Holm-corrected

Measure	Version A	Version B	z	p (Holm)	r
UEQ-S pragmatic (−3 to +3)	2.50 (1.50–3.00)	2.75 (1.75–3.00)	+1.07	1.000	0.23
UEQ-S hedonic (−3 to +3)	2.00 (1.25–3.00)	2.25 (2.00–3.00)	+2.88	0.031	0.63
Perceived accuracy (1–5)	4.00 (4.00–5.00)	4.00 (4.00–5.00)	+0.58	1.000	0.13
Transparency (1–5)	5.00 (4.00–5.00)	5.00 (5.00–5.00)	+1.31	1.000	0.29
Trust (1–5)	4.00 (3.00–5.00)	4.00 (4.00–5.00)	+0.26	1.000	0.06
Confidence (1–5)	4.00 (3.00–5.00)	4.00 (3.00–5.00)	+0.49	1.000	0.11
Use intention (1–5)	5.00 (4.00–5.00)	5.00 (4.00–5.00)	+0.82	1.000	0.18
Profile accuracy (1–5)	4.00 (3.00–5.00)	4.00 (4.00–5.00)	+0.77	1.000	0.17

bootstrap 95% CI 0.18–0.98). The difference was significant before correction ($p = 0.004$) and after Holm correction for eight outcomes ($p = 0.031$); its effect size was large ($r = 0.63$). Pragmatic quality was higher in Version B on average (2.27 vs 2.07) but not significantly ($p = 0.28$, $r = 0.23$). None of the six 1–5 items differed (Fig. 6). Their medians were 4 or 5 in both versions, and between 9 and 18 of the 21 participants gave both versions the same rating. Transparency was at the scale maximum for 16 participants in Version B and 13 in Version A; trust had almost the same mean in both versions (4.14 and 4.19), with 14 ties.

Against the UEQ-S benchmark (Hinderks et al, 2018), both versions were excellent on all three scale means (Fig. 5): pragmatic quality 2.07 (A) and 2.27 (B), against an excellent threshold of 1.73; hedonic quality 1.76 and 2.27 (threshold 1.55); overall 1.92 and 2.28 (threshold 1.58). The 95% confidence intervals of Version B lie entirely in the excellent range; those of Version A reach down to the good range, and for hedonic quality to the above-average range. Ratings were close to the ceiling: five participants gave every UEQ-S item the most positive value in both versions, and nine did so in at least one.

Plans were recorded in 11 of the 21 Version B sessions; for the others the record cannot show whether a plan was opened, so these participants were kept in the main analysis. Among the 11 with a recorded plan, the advantage of Version B was larger (mean difference 0.75 in hedonic and 0.51 in overall quality, both $p = 0.008$ before correction), transparency, trust and profile accuracy again did not differ, and eight preferred Version B, two Version A and one neither. Excluding the three Version B sessions without a sentence ($n = 18$) left the hedonic difference almost unchanged (mean 0.49, $p = 0.011$ before correction); no other outcome reached significance. The advantage of Version B was larger among the ten participants of the class session (mean difference 0.68 in hedonic and 0.56 in overall quality) than among the 11 remote participants (0.36 and 0.17), although the difference between the two groups was not significant (Mann–Whitney U, $p = 0.91$ and $p = 0.30$). The B−A differences did not depend on order on any outcome (Mann–Whitney U, all $p > 0.09$).

6.3 Preference and reasons (RQ4)

Twelve participants would use Version B on their next trip, five Version A, and four had no preference (Fig. 6; binomial test of B against A, $p = 0.14$). The choice matched the ratings: participants who chose Version B rated it 0.66 points higher than Version A

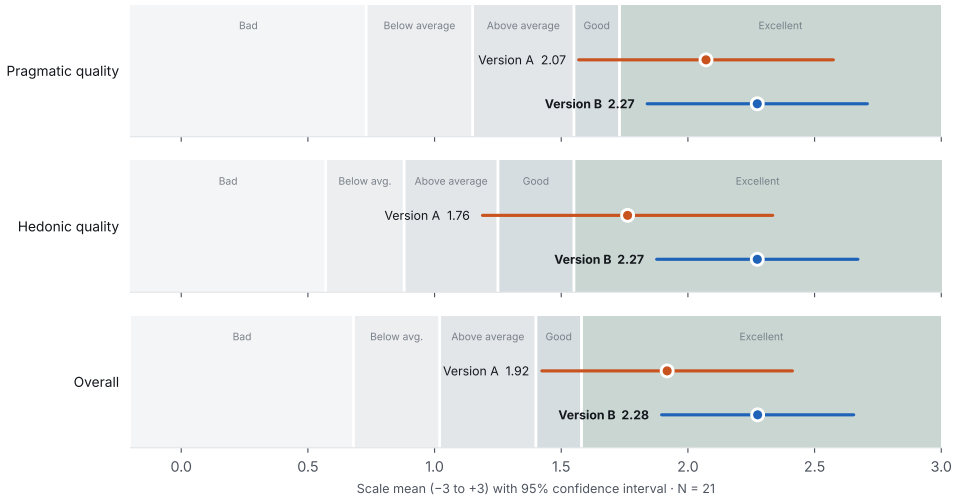


Fig. 5 UEQ-S scale means of the two versions with 95% confidence intervals ($N = 21$), on the bands of the UEQ-S benchmark (Hinderks et al, 2018). Each scale has its own band limits.

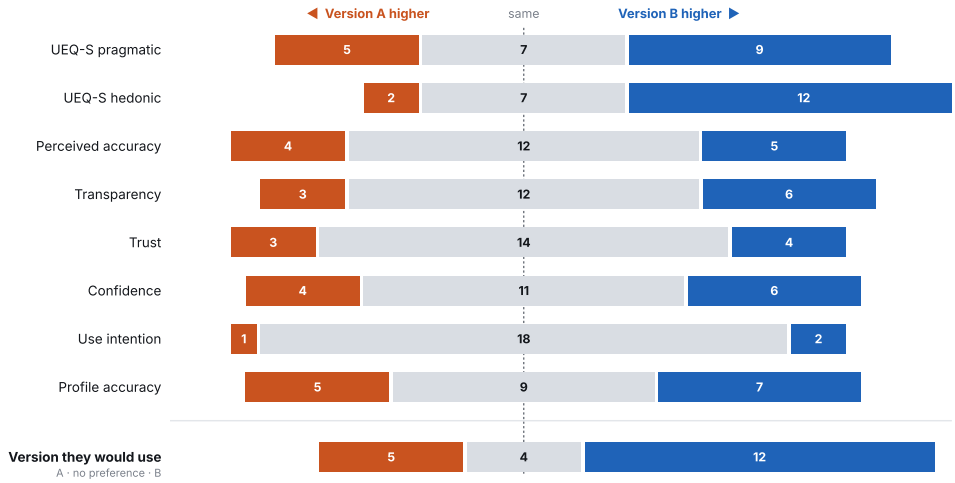


Fig. 6 For each outcome, the number of the 21 participants who rated Version A higher (left), gave both versions the same rating (centre) or rated Version B higher (right). Bottom row: the version each participant would use on a next trip.

in overall UEQ-S, and those who chose Version A rated it 0.12 points higher than Version B.

Twelve participants wrote at least one open answer (quotations translated from Greek unless marked as English). Four themes emerged.

- *Photographs as an easy way to be described.* Three participants valued the photograph deck: "the photos definitely helped to pinpoint better my interests" (P-NYR7, English); "I don't have to describe myself" (P-YHN5, English); the photographs were "not something difficult or complicated to do" (P-HGQM).
- *A plan for the day and its times.* "It makes a nice programme for my day and gives me tips" (P-TFA7, chose B); B was "more accurate with respect to the time one has" (P-ZD69, chose B).
- *Range and control.* One participant found B's suggestions "quite standard" and the photographs narrower than A's icons (P-N27K, chose A). Another preferred A for "a ready list" of places before a trip, rather than a plan for a given day (P-HGQM, chose A). A third asked to add a stop of their own to B's plan (P-J8XK), and a fourth proposed a clarifying question in the profile, on whether the visitor is drawn to a city's modern history or to antiquity (P-ZD69).
- *Little difference.* Three participants liked both versions "for different reasons" or found both clear and easy. One did not see what changed between the versions (P-HTB4, who wrote no sentence in B).

7 Discussion

Both versions were rated excellent, so the study compares two well-received systems. Theoros was rated higher mainly on hedonic quality (RQ1), the UEQ-S scale for stimulation and novelty. These were also the two UEQ scales on which ACUX-R scored only above average in its field study. The difference was large and remained significant after correction for eight outcomes, but it was larger among the author's students in the class session than among remote participants, so it needs confirming in a larger and more independent sample. Pragmatic quality, which concerns efficiency and clarity, was already high for the re-implemented ACUX-R.

7.1 Division of labour between the model and code

The planner evaluation shows the failure mode directly: adding temporal rules to the prompt made scheduling worse (from 7 to 3 of 10 valid plans), while moving time to code made it reliable (10 of 10). This matches solver-backed hybrid planners in the LLM literature (Hao et al, 2024) and gives heritage designers a simple rule: let the model choose and explain, and let code keep time and facts. The user study cannot show whether participants valued correct times. In a desk task a wrong closing time has no consequence, and only one participant mentioned times, liking the estimated times in both versions and judging B more accurate for the time available (P-ZD69). It shows that the plan was noticed as a plan for the day (P-TFA7) and that no recorded plan needed the template fallback. Whether verified times matter to visitors is a question for an on-site study.

7.2 Dependability, transparency and trust

Version B did not raise transparency or trust measurably (RQ2). Both items were near the scale maximum in both versions, so the explanations in Version B had little

room to add to a list of places that participants already found clear. UEQ-S has no dependability scale, so the study cannot show directly whether the dependability gap of ACUX-R closed. The re-implemented ACUX-R scored excellent pragmatic quality here, whereas the original scored below average on dependability. The shared data, design and app, the short desk task and the difference between the full UEQ and UEQ-S may all contribute, and the study cannot separate them. Trust in a guide is tested when it is wrong on site, not in a short desk task.

7.3 Implications for stereotype user models built by LLMs

Profile accuracy did not differ (median 4 in both versions; seven participants rated Version B higher and five lower). The open answers suggest why: some participants found the photographs an easy and accurate way to be described, while one found that 16 photographs covered fewer interests than 40 icons. Three participants wrote no sentence, so their profile came from the photographs alone, and one of them did not notice a difference between the versions. Whether the sentence adds accuracy beyond the photographs needs a comparison within Version B. The overlap between Gourmand and Traditionalist when food is mentioned, seen in the profiler tests, is a question for the typology itself.

Three lessons follow for user modelling. First, perceived profile accuracy is a weak signal for comparing user models: when both models are shown to the visitor and can be edited, visitors rate the result of their own adjustments, and ratings cluster at the top of the scale; behavioural measures, such as how much visitors change the model and whether recommendations change with it, are needed. Second, an LLM can map free text to a fixed stereotype set with evidence for each score, which keeps the user model comparable across visitors and scrutable, unlike free-form natural-language profiles; but the photographs and the sentence then need to be weighed against each other, and a clarifying question, as one participant proposed, may resolve ambiguity that neither resolves alone. Third, binding every explanation to the user model and to a curated description makes errors traceable: in our case a wrong name in an explanation came from the data, not the model, and was fixed once.

7.4 From one visitor to a group

In an informal discussion after the class session, which was not recorded and is not part of the study data, students asked for social features: sharing a route or a place with friends, suggesting it to them, a chat, and planning a visit together as a group. ACUX profiles describe one visitor, and Theoros plans for one visitor. Planning for a group would need a way to combine several multi-label profiles into one plan, and explanations that say whom each stop is for. We see this as a direction for future work rather than a finding.

7.5 Heritage interpretation and cost

A model that writes the reason to visit a museum or site takes part in heritage interpretation; Theoros therefore ties every explanation to a source description and every hour to an official source. Grounding moves the responsibility for facts to the curated

data: in Fig. 2 the model repeated a wrong first name from a place description faithfully, and the error was fixed once, in the data, rather than in every generated text. Generative features cost little per visitor, but a free tier is not enough for a study or a service: its rate limits made calls fail in the first sessions, and free-tier data terms differ from paid terms, which must be disclosed to visitors. On the paid tier, all model calls of the project cost about €20 in total.

7.6 Limitations

The study used a scenario task at a desk rather than a real visit, so the features that act during a walk (re-planning for rain, heat or closures, and the check-in) were not exercised. Participants spent under three minutes in each version (median), which gives a first impression rather than a planned day. The sample was small (21 of 34 eligible participants completed both versions), drawn from the author's university and personal networks, and mostly aged 18–34; only one participant was 65 or older. Ratings near the top of the scales point to a ceiling effect. Ten participants were the author's students, who completed the study in class with the author present and had learned Google AI Studio, the tool in which the app was prototyped; courtesy and demand effects may have raised their ratings, especially of the newer version, and the advantage of Version B was indeed larger in that group. The 13 participants who did not finish may have liked the app less, which would bias the sample upwards. Version B differs from Version A in several features at once, so the study compares two systems rather than single features. Three Version B sessions ran without a sentence; excluding them did not change the results. The step log was not analysed. The model and its terms may change after the study; prompts and model are logged with every call.

8 Conclusion

Theoros keeps the ACUX typology and replaces icons and fixed lists with photographs, the visitor's own words and explained, timed plans. In a within-subjects study with 21 participants, both versions were rated excellent. Theoros was rated higher in hedonic quality and preferred by 12 participants against five, while transparency, trust and profile accuracy did not differ. For heritage recommenders the lesson is to let the model choose and explain, and to let code keep time and facts. Next steps are an on-site study in Athens with an independent sample, plans for groups of visitors, and re-estimating the photograph weights from survey data on the typology.

Supplementary information. Online Resource 1: interactive walkthrough of Version A and Version B, the planning pipeline, the planner evaluation and the study flow (a ZIP file with one HTML page that opens in any web browser).

Acknowledgements. We thank the participants of the study.

Statements and Declarations

- **Funding.** No dedicated funding was received for this work.
- **Conflict of interest.** The author declares no competing interests.

- **Ethics approval and consent to participate.** Participants were adults (18 years or older) who gave informed consent in the app before taking part; participation was anonymous (an automatic code, no name, e-mail address or location), and declining ended the session with nothing saved. Students invited during a class took part voluntarily, without reward or course credit. The study was approved by the Research Ethics Committee of the University of the Aegean on 1 October 2026, before data collection began. The study was carried out in accordance with the relevant guidelines and regulations.
- **Consent for publication.** Not applicable.
- **Data availability.** The anonymised study data (one row per participant and version) and the locked prompts will be deposited in Zenodo on acceptance, with the DOI given in the published article. Open answers are shared only where they contain no personal details.
- **Materials availability.** The questionnaire items are listed in Appendix A.
- **Code availability.** The source code of the study app is available from the author on reasonable request.
- **Use of AI tools.** Described in Section 5.6.
- **Author contribution.** Single author.

Appendix A Questionnaire items

Background (before the first version): age band (18–24, 25–34, 35–44, 45–54, 55–64, 65+); "How often do you use apps when you visit a city?" (never–always); "How well do you know central Athens?" (not at all–very well).

UEQ-S (after each version, −3 to +3): obstructive–supportive, complicated–easy, inefficient–efficient, confusing–clear (pragmatic quality); boring–exciting, not interesting–interesting, conventional–inventive, usual–leading edge (hedonic quality).

Agreement items (after each version; strongly disagree = 1 to strongly agree = 5):

- Perceived accuracy: "The places recommended to me matched my interests."
- Transparency: "I understood why these places were recommended to me."
- Trust: "I can trust the recommendations of this version."
- Confidence: "I am confident I would like the places recommended to me."
- Use intention: "I would use this version on a trip."

Profile accuracy (after each version; not at all = 1 to very well = 5): "How well does the profile this version gave you describe you?"

Open question (after each version, optional): "What did you like most about this version, and what would you change?"

Final question: "Which version would you use on your next trip?" (Version A, Version B, no preference) and "Why?" (optional).

References

Ardissono L, Kuflik T, Petrelli D (2012) Personalization in cultural heritage: the road travelled and the one ahead. User Modeling and User-Adapted Interaction 22:73–99

Ardissono L, Ferrero F, Geninatti Cossatin A, et al (2026) A context-aware cultural heritage guide powered by LLMs. Authors' version of a paper at UMAP '26, Gothenburg. arXiv:2608.00549

Chen A, Ge X, Fu Z, et al (2024) TravelAgent: an AI assistant for personalized travel planning. ArXiv:2409.08069

Gavalas D, Konstantopoulos C, Mastakas K, et al (2014) A survey on algorithmic approaches for solving tourist trip design problems. Journal of Heuristics 20(3):291–328

Geninatti Cossatin A, Mauro N, Ferrero F, et al (2025) Tell me more: integrating LLMs in a cultural heritage website for advanced information exploration support. Information Technology & Tourism 27(2):385–416. https://doi.org/10.1007/s40558-025-00312-8

Gibson H, Yiannakis A (2002) Tourist roles: needs and the lifecourse. Annals of Tourism Research 29(2):358–383. https://doi.org/10.1016/S0160-7383(01)00037-8

Grün C, Neidhardt J, Werthner H (2017) Ontology-based matchmaking to provide personalized recommendations for tourists. In: Schegg R, Stangl B (eds) Information and Communication Technologies in Tourism 2017. Springer, Cham, pp 3–16

Gursoy D, Li Y, Song H (2023) ChatGPT and the hospitality and tourism industry: an overview of current trends and future research directions. Journal of Hospitality Marketing & Management 32(5):579–592

Hao Y, Chen Y, Zhang Y, et al (2024) Large language models can plan your travels rigorously with formal verification tools. arXiv preprint arXiv:2404.11891

Hinderks A, Schrepp M, Thomaschewski J (2018) A benchmark for the short version of the User Experience Questionnaire. In: Proceedings of the 14th International Conference on Web Information Systems and Technologies (WEBIST 2018). SciTePress, pp 373–377, https://doi.org/10.5220/0007188303730377

Ji Z, Lee N, Frieske R, et al (2023) Survey of hallucination in natural language generation. ACM Computing Surveys 55(12):248:1–248:38

Ju D, Jiang S, Cohen A, et al (2024) To the globe (TTG): Towards language-driven guaranteed travel planning. https://arxiv.org/abs/2410.16456, arXiv:2410.16456

Konstantakis M (2026) What visitors say and what icons show: re-examining a cultural-visitor typology for generative-AI personalisation in heritage guides, manuscript submitted to Heritage

Konstantakis M, Christodoulou Y, Alexandridis G, et al (2022a) ACUX typology: a harmonisation of cultural-visitor typologies for multi-profile classification. Digital

2(3):365–378. https://doi.org/10.3390/digital2030020

Konstantakis M, Christodoulou Y, Aliprantis J, et al (2022b) ACUX Recommender: a mobile recommendation system for multi-profile cultural visitors based on visiting preferences classification. Big Data and Cognitive Computing 6(4):144. https://doi.org/10.3390/bdcc6040144

Laugwitz B, Held T, Schrepp M (2008) Construction and evaluation of a user experience questionnaire. In: HCI and Usability for Education and Work. Symposium of the Austrian HCI and Usability Engineering Group. Springer, Berlin, Heidelberg, pp 63–76

McKercher B (2002) Towards a classification of cultural tourists. International Journal of Tourism Research 4(1):29–38

Neidhardt J, Schuster R, Seyfang L, et al (2014) Eliciting the users' unknown preferences. In: Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14). ACM, New York, pp 309–312, https://doi.org/10.1145/2645710.2645767

Neidhardt J, Seyfang L, Schuster R, et al (2015) A picture-based approach to recommender systems. Information Technology & Tourism 15(1):49–69. https://doi.org/10.1007/s40558-014-0017-5

Pavlidis G (2019) Recommender systems, cultural heritage applications, and the way forward. Journal of Cultural Heritage 35:183–196

Pu P, Chen L, Hu R (2011) A user-centric evaluation framework for recommender systems. In: Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11). ACM, New York, pp 157–164

Ramos J, Rahmani HA, Wang X, et al (2024) Transparent and scrutable recommendations using natural language user profiles. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 13971–13984, https://doi.org/10.18653/v1/2024.acl-long.753

Rich E (1979) User modeling via stereotypes. Cognitive Science 3(4):329–354

Schrepp M, Hinderks A, Thomaschewski J (2017a) Construction of a benchmark for the User Experience Questionnaire (UEQ). International Journal of Interactive Multimedia and Artificial Intelligence 4(4):40–44

Schrepp M, Hinderks A, Thomaschewski J (2017b) Design and evaluation of a short version of the User Experience Questionnaire (UEQ-S). International Journal of Interactive Multimedia and Artificial Intelligence 4(6):103–108

Seaton AV (2002) Tourism as metempsychosis and metensomatosis: the personae of eternal recurrence. In: Dann GMS (ed) The Tourist as a Metaphor of the Social World. CABI, Wallingford, p 135–168

Sertkan M, Neidhardt J, Werthner H (2020a) Eliciting touristic profiles: a user study on picture collections. In: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20). ACM, New York, pp 230–238, https://doi.org/10.1145/3340631.3394868

Sertkan M, Neidhardt J, Werthner H (2020b) PicTouRe – a picture-based tourism recommender. In: Proceedings of the 14th ACM Conference on Recommender Systems (RecSys '20). ACM, New York, pp 597–599, https://doi.org/10.1145/3383313.3411526

Smith M (2003) Holistic holidays: tourism and the reconciliation of body, mind and spirit. Tourism Recreation Research 28:103–108

Tan Z, Jiang M (2023) User modeling in the era of large language models: Current research and future directions. IEEE Data Engineering Bulletin ArXiv:2312.11518

Tang Y, Wang Z, Qu A, et al (2024) ItiNera: integrating spatial optimization with large language models for open-domain urban itinerary planning. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track. Association for Computational Linguistics, Miami, pp 1413–1432, https://doi.org/10.18653/v1/2024.emnlp-industry.104

Tintarev N, Masthoff J (2015) Explaining recommendations: design and evaluation. In: Ricci F, Rokach L, Shapira B (eds) Recommender Systems Handbook, 2nd edn. Springer, Boston, p 353–382

Trichopoulos G, Ordoumpozanis K, Caridakis G (2026) An evaluation of LLM-based chatbots for enhancing the visitor's user experience at cultural exhibits. Journal on Computing and Cultural Heritage 19(1):1–25. https://doi.org/10.1145/3775062

Vansteenwegen P, Souffriau W, Vanden Berghe G, et al (2011) The City Trip Planner: an expert system for tourists. Expert Systems with Applications 38(6):6540–6546. https://doi.org/10.1016/j.eswa.2010.11.085

Xie J, Zhang K, Chen J, et al (2024) TravelPlanner: a benchmark for real-world planning with language agents. In: Proceedings of the 41st International Conference on Machine Learning, pp 54590–54613

Zhai A, Awasthi P (2025) Optimizing LLM-based trip planning. https://research.google/blog/optimizing-llm-based-trip-planning/, google Research blog, 6 June 2025

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.zip
- AppendixAQuestionnaireitems.pdf