

Interpretable Recommendations and Parameter-Grounded LLM Explanations with Multi-Graph Attention: Electronic Companion

Appendix Contents

A. Literature Comparisons	A-1
A.1. Evaluation Quality Metrics	A-1
A.2. Comparison with Benchmarks	A-1
A.3. Comparisons with the ERS Literature	A-1
A.4. Comparative Analysis with Attention- and KG-based recommenders	A-4
B. Key Notations	B-1
C. Stage-2 Implementation Details	C-1
C.1. Stage 2 Computational Cost	C-1
C.2. Prompt Templates	C-1
D. Proofs	D-1
D.1. Proof for Proposition 1	D-1
D.2. Proof for Proposition 2	D-1
D.3. Proof for Proposition 3	D-1
D.4. Proof for Corollary 1	D-1
D.5. Proof for Proposition 4	D-2
E. Summary Statistics of the Yelp Data	E-1
E.1. Data Descriptions	E-1
E.2. Experimental Setting	E-2
F. Performance Evaluation Metrics	F-1
G. Experiment Setup	G-1
G.1. Branching Questionnaire and Scenario Assignment	G-1
G.2. On-Screen Text for Illustrative Path	G-1
G.3. Interface of the Experiment Conditions	G-2
G.4. Statistical Analysis of All Questions	G-2
G.5. Balance Check	G-4
G.6. Mediation Analysis	G-5

Appendix A. Literature Comparisons

A.1. Evaluation Quality Metrics

Table A1 summarizes the six explanation quality metrics introduced by Tintarev and Masthoff (2012a,b). We exclude scrutability because its assessment requires an interactive system in which users can modify inferred preferences and evaluate whether the system updates its recommendations in response, which is outside the scope of our experimental design.

Metric	Explanation
Effectiveness	This explanation is effective in helping your friend evaluate if he/she should visit “[restaurant]”.
Efficiency	This explanation helps your friend decide efficiently whether to visit “[restaurant]”.
Persuasiveness	This explanation persuades your friend to visit “[restaurant]”.
Satisfaction	This explanation would improve how easy it is to pick a recommendation.
Transparency	This explanation helps your friend to understand what the recommendation is based on.
Trust	This explanation increases your friend’s confidence in the recommender system.

Table A1 Evaluation Metrics for Explanations Proposed by Tintarev and Masthoff (2012b)

A.2. Comparison with Benchmarks

We summarize the input data and the characteristics of different benchmarks in Table A2.

	Data		Method char.
	Auxiliary	Network	Interpretability
SVD++ (Koren 2008)	N	N	N
NNMF (Dziugaite and Roy 2015)	N	N	N
F-EAE (Hartford et al. 2018)	N	N	N
IGMC (Zhang and Chen 2020)	N	N	N
sRGCNN (Monti et al. 2017)	N	Y	N
GraphRec (Fan et al. 2019)	N	Y	N
DGAN (Wu et al. 2019)	Y	Y	N
NGCF (Wang et al. 2019b)	N	Y	N
GC-MC (Berg et al. 2018)	Y	N	N
GRALS (Rao et al. 2015)	Y	Y	N
IFM (Pan et al. 2021)	Y	N	Y
KGAT (Shimizu et al. 2022)	Y	N	Y
HGCL (Chen et al. 2023)	Y	Y	N
LightGCL (Cai et al. 2023)	N	N	N
MG-GAT	Y	Y	Y

Table A2 Input data and method characteristics of the benchmarks. Note: “Network” indicates explicit user–user or business–business graphs beyond the user–item interaction graph; knowledge-graph/entity relations used by KG-based recommenders are counted under “Auxiliary”.

A.3. Comparisons with the ERS Literature

Table A3 positions our work in the explainable-RS literature along seven dimensions: whether the method uses a large language model, whether the paper reports a user study (and the number of subjects), the paper category, whether it uses network data, whether it uses auxiliary information, and the evaluation metrics employed. The comparison shows that most prior LLM-based ERS work relies on case studies or small-scale human-labeler validation rather than randomized user experiments, and few combine network data

with auxiliary information. Our paper combines a randomized user experiment with 1,363 participants, both network and auxiliary information, and parameter-grounded LLM-generated explanations—a combination that does not appear in any of the surveyed work.

Paper	Large language model	User study	Number of Subjects	Category	Network data	Auxiliary information	Metrics
Rahdari et al. (2024)	Yes	Human labeler	None	LLM	No	Yes	No
Liu et al. (2023)	Yes	Human labeler	3	LLM	No	Yes	BLEU; ROUGE
Li et al. (2023b)	Yes	Case study	None	LLM	No	Yes	Incorrect classification; Correct classification; insufficient explanation; Correct classification; acceptable explanation; Correct classification; satisfying explanation
Lei et al. (2024)	Yes	Case study	3	LLM	No	Yes	No
Li et al. (2023a)	Yes	Case study	None	LLM	No	Yes	BLEU; ROUGE
Wang et al. (2024)	Yes	No	None	LLM	No	Yes	BLEU; ROUGE
Peng et al. (2024)	Yes	No	None	LLM	No	Yes	BLEU; ROUGE
Gao et al. (2023)	Yes	Case study	None	LLM	No	No	No
Geng et al. (2022)	Yes	No	None	LLM	No	No	BLEU; ROUGE
Hada and Shevade (2021)	Yes	No	None	LLM	No	No	BLEU
Ni et al. (2019)	Yes	Controlled experiment	≥ 3	LLM	No	No	BLEU; human (Relevance; Informativeness; Diversity)
Geng et al. (2023)	Yes	No	None	LLM	No	Yes	BLEU; ROUGE
Wang et al. (2025)	Yes	Case study	None	LLM	No	No	Aspect score; Rating score
Lu et al. (2023)	No	Controlled experiment	39	NA	No	No	BLEU; ROUGE
Shimizu et al. (2022)	No	No	None	Knowledge graph paths	No	Yes	No
Li et al. (2021)	No	Case study; Human labeler	10	Pre-defined templates	No	No	No
Zhu et al. (2021)	No	Controlled experiment	20	Knowledge graph paths	No	No	No
Pan et al. (2021)	No	No	None	Item features	No	Yes	General preferences; Specific preferences
Sun et al. (2020)	No	Case study	None	Knowledge graph paths	No	No	BLEU; ROUGE
Huang et al. (2019)	No	No	None	Knowledge graph paths	Yes	Yes	No
Xian et al. (2020)	No	Case study	None	Knowledge graph paths	Yes	Yes	No
Wang et al. (2018b)	No	No	None	Item features	No	Yes	No
McInerney et al. (2018)	No	Controlled experiment	None	Item features	No	Yes	No
Chen et al. (2018a)	No	No	None	Ranked text	No	No	No
Chen et al. (2018b)	No	No	None	Deep learning	No	No	No
Seo et al. (2017)	No	No	None	Deep learning	No	No	No
Park et al. (2017)	No	No	None	Graph-based social explanation	Yes	No	No
Our paper	Yes	Controlled experiment	1363	LLM guided by model parameters	Yes	Yes	Perceived relevance; Future interency; Confidence/Trust; Transparency; Effectiveness; Efficiency; Persuasiveness; Satisfaction

Table A3 Paper Comparisons in the Explainable Recommender Systems Literature

A.4. Comparative Analysis with Attention- and KG-based recommenders

Table A4 situates MG-GAT relative to representative attention-based and knowledge-graph (KG) recommenders along four explanation-relevant dimensions that are central to our framing in Section 2. First, we distinguish the underlying relational structure each method operates on (single homogeneous graph, multi-graph/user-item plus social graph, or an explicit KG). Second, we record what is actually exposed as an explanation signal (e.g., attention weights over latent states, explicit KG paths, or text salience). Third, we indicate whether the method provides an explicit feature-level decomposition of attention or importance scores into observable attributes (as opposed to weights over latent embeddings). Fourth, we indicate whether any user-facing explanation text is *parameter-grounded* in prediction-time model signals, as defined in Section 4.

Table A4 shows how MG-GAT combines several known components into an interpretable interface: NIG exposes tie-level salience, FR provides an attribute-level decomposition of neighbor importance, and these prediction-time signals are reused as control inputs for Stage 2 explanation generation. NIG and FR therefore serve as inspectable signals for both model interpretation and explanation generation.

Table A4 Comparison of representative attention- and KG-based recommenders with MG-GAT along explanation-relevant dimensions.

Method family (examples)	Graph structure	Exposed explanation signal	Feature-level decomposition?	Parameter-grounded narrative?
GAT-style attention over a single graph (Veličković et al. 2018)	single homogeneous graph	attention weights over latent node states	No	No
Dual-/social-side graph attention recommenders (Wu et al. 2019, Fan et al. 2019)	user-item + social-side graph	latent attention weights (typically over embeddings)	No	No
KG attention recommenders (e.g., KGAT and variants) (Wang et al. 2019a, Shimizu et al. 2022)	knowledge graph	attention weights / path scores over KG embeddings (often visualized)	No	No
Path-based KG explanation methods (Huang et al. 2019, Xian et al. 2020, Pan et al. 2021)	knowledge graph	explicit multi-hop reasoning paths	No (path-level, not feature-level)	No (paths/visualizations rather than parameter-grounded text)
Review-/text-attention explainers (Chen et al. 2018a, Li et al. 2021)	user-item interactions + text	salient words, reviews, or aspects	No (text salience, not attention-logit decomposition)	No (no explicit grounding in prediction-time parameters)
MG-GAT (this paper)	user-user + business-business graphs (multi-graph)	NIG (tie salience) + FR (attribute contributions)	Yes (FR linear decomposition)	Yes (Stage 2 narratives grounded in NIG/FR)

Appendix B. Key Notations

We summarize the key notations used in this paper in Table B1. For notational consistency, we use the lower-case, bold-lower-case, bold-capital case to represent scalar, vector, and matrix, respectively. We use subscript or superscript u and b to represent users and businesses, respectively.

Table B1 Key notations

Notations	Definitions and Descriptions
$\mathbf{X} \in \mathbb{R}^{n \times m}$	User-business rating matrix with n users and m businesses
$\hat{\mathbf{X}} \in \mathbb{R}^{n \times m}$	Predicted user-business rating matrix
$\mathbf{U} \in \mathbb{R}^{n \times k_f}, \mathbf{B} \in \mathbb{R}^{m \times k_f}$	Inferred latent user and business embedding with dimension k_f
$\mathbf{S}_u \in \mathbb{R}^{n \times s_u}, \mathbf{S}_b \in \mathbb{R}^{m \times s_b}$	Auxiliary information about users and businesses with dimensions s_u and s_b
$\mathbf{G}_u, \mathbf{G}_b$	User friendship network and business network
$\mathbf{L}_u, \mathbf{L}_b$	Graph Laplacian of the user friendship network and business network
N_i^u, N_j^b	Neighbor set of user i or business j on the user or business graph
$\alpha_u^{k \rightarrow i}, \alpha_b^{l \rightarrow j}$	Neighbor importance for user and business
NIG_θ	Neighbor Importance Graph: attention weights over $\mathbf{G}_u$ and $\mathbf{G}_b$ used for prediction and explanations
FR_θ	Feature Relevance coefficient vectors $\{\text{FR}_{\text{self}}^u, \text{FR}_{\text{nb}}^u, \text{FR}_{\text{self}}^b, \text{FR}_{\text{nb}}^b\}$ decomposing attention logits over observable attributes
k_f	Dimension of latent user/business embeddings
K, F	Number of high-salience neighbors (top- K) and features (top- F) exposed to Stage 2
$S_\theta(u, j)$	High-salience neighbor set for user u and recommended item j selected via NIG_θ
$\text{score}(f)$	NIG -weighted instance-level contribution score used to select salient features for Stage 2
$\mathcal{F}_\theta^*(u, j)$	High-salience feature set for business j selected by the top- F score(f) values
$\mathbf{a}_{u,\text{self}}(\mathbf{a}_{u,\text{nb}}), \mathbf{a}_{b,\text{self}}(\mathbf{a}_{b,\text{nb}})$	Coefficients to compute feature relevance of focal (neighbor) nodes
$\Omega_{\text{training}} \in \mathbb{R}^{n \times m}$ and $\Omega_{\text{test}} \in \mathbb{R}^{n \times m}$	Indicator matrix representing entries for the training and test set
$\mathcal{R}(\cdot)$	Regularization term
$\mathcal{L}$	Loss function
$\mathbf{W}_u^{(q)}$ and $\mathbf{W}_b^{(q)} \forall q \in \{1, 2, 3\}, \mathbf{W}_{us}^{(2)}, \mathbf{W}_{bs}^{(2)}$	Learnable weights in the deep learning layer
$\mathbf{H}_u^{(q)}$ and $\mathbf{H}_b^{(q)} \forall q \in \{1, 2, 3, 4\}$	Intermediate embeddings for users and businesses
$\mathbf{b}_u^{(1)} \in \mathbb{R}^{d_u^1}, \mathbf{b}_u^{(x)} \in \mathbb{R}^n, \mathbf{b}_b^{(1)} \in \mathbb{R}^{d_b^1}, \mathbf{b}_b^{(x)} \in \mathbb{R}^m, b_x \in \mathbb{R}$	Learnable bias term, where d_1^u and d_1^b are the size of the latent dimension

Appendix C. Stage-2 Implementation Details

C.1. Stage 2 Computational Cost

Stage 2 generation dominates the computational cost of our two-stage design; Stage 1 inference is comparable to that of other GNN recommenders and can be batched and cached, whereas Stage 2 invokes an LLM multiple times (Algorithm 2). Three instantiations of Stage 2 represent distinct points on the cost–fidelity tradeoff curve; all three preserve the same evidence contract, namely the top- K neighbors and top- F features drawn from NIG/FR.

- **Deterministic instantiation.** Templates verbalize NIG/FR without any LLM call. This isolates the contribution of the structured evidence from any contribution of language-model generation and provides a worst-case lower bound on explanation fidelity.
- **Single-pass instantiation.** A single constrained generation call produces an explanation together with its cited-evidence declaration; no candidate plans are searched and no critic scoring is applied. This isolates the marginal effect of plan-and-critique search by holding the evidence contract fixed.
- **Plan-and-critique instantiation.** Algorithm 2 generates multiple candidate plans and explanations and selects among them using NIG-weighted critics. Compared with the single-pass instantiation, the additional calls trade compute for stronger filtering against support violations.

Table C1 Per-recommendation LLM call counts for the three Stage 2 instantiations. Let $K = |S_\theta(u, j)|$ be the number of neighbors in the support set and let N_p, N_e be the number of candidate plans and explanations.

Instantiation	# LLM calls	Description
Deterministic	0	Template-based verbalization of NIG/FR; no LLM call.
Single-pass	1	One constrained generation call producing the explanation and its cited-evidence declaration.
Plan-and-critique	$1 + KN_p + N_e + KN_e$	Algorithm 2: plan and explanation candidates scored by NIG-weighted critics.

Table C2 Illustrative Stage 2 call budget (per recommendation) under Algorithm 2. Let $K = |S_\theta(u, j)|$ be the number of neighbors surfaced and let N_p/N_e be the number of candidate plans/explanations generated.

Component	# LLM calls	Output-length target (approx.)
Plan generation	1	Short plan paragraph
Plan evaluation (critics)	$K \cdot N_p$	One paragraph ending with six metric scores
Explanation generation	N_e	2–3 sentences; ~ 60 – 68 words (~ 80 – 95 tokens)
Explanation evaluation (critics)	$K \cdot N_e$	One paragraph ending with six metric scores

C.2. Prompt Templates

For reproducibility, we provide the exact prompt templates used in Stage 2. These templates are parameter-grounded: all neighbor and feature content is selected via NIG/FR, and no additional context is injected. We also include an EVIDENCE.USED checklist line in the plan and explanation prompts to enable a lightweight, deterministic support check; this line is removed before user-facing display.

Prompt 1 (General Prompt Provided to All LLM Agents)

Here is the dining history of a customer, including only the restaurants they have highly rated:

- Neighbor $k(\in \{1, 2, 3\})$: {restaurant name} described with highly relevant features $\{F_{N_k}$ by $FR\}$.

Based on this user’s visiting history, the machine learning algorithm recommends the following restaurant:

- Focus business: {restaurant name} with relevant features $\{F_f$ according to $FR\}$.

Prompt 2 (Plan Generation)

You are a restaurant recommender system. {Provide the general prompt in Prompt 1.}

Step 1 (evidence checklist). *Output exactly one line in the form:*

EVIDENCE_USED: neighbors=[Neighbor 1, Neighbor 2, Neighbor 3]; features=[...]

List only neighbors and feature names that appear in the prompt. Do not introduce any restaurant or feature that is not provided above.

Step 2 (plan). *Write a short plan (one paragraph) for the explanation that will be shown to the consumer, linking the recommendation to the visited restaurants using only the listed evidence.*

Prompt 3 (Evaluation Prompt)

Consider you are a fan of {name of neighboring restaurants}. {Provide the general prompt in Prompt 1.}

Please evaluate the explanation (plan) for the recommendation based on the following criteria: If the plan/explanation mentions any restaurant or attribute that does not appear in the prompt, treat it as unsupported and assign very low scores (especially transparency and trust). Provide a brief analysis and end the paragraph with exactly this sentence with the ratings: “Thus, effectiveness score is {s}, efficiency score is {s}, persuasiveness score is {s}, satisfaction score is {s}, transparency score is {s}, and trust score is {s}.”

Prompt 4 (Explanation Generation)

You are a restaurant recommender system. {Provide the general prompt in Prompt 1.}

Step 1 (evidence checklist; internal). *Output exactly one line in the form:*

EVIDENCE_USED: neighbors=[Neighbor 1, Neighbor 2, Neighbor 3]; features=[...]

List only neighbors and feature names that appear in the prompt. Do not introduce any restaurant or feature that is not provided above.

Step 2 (user-facing text). *Then generate a paragraph of explanation (in 2–3 sentences) for the recommended restaurant to be shown to the consumer so that the consumer can understand the reasons behind the recommendation (relate to the customer’s visited restaurants). The EVIDENCE_USED line will be removed before display.*

Appendix D. Proofs

D.1. Proof for Proposition 1

Proof for Proposition 1 From Eq. (2), the first-layer embedding is linear in the observed attributes: $H^{(1)} = W^{(1)}x$. Writing the linear pre-activation attention score in Eq. (3) as the sum of a focal term and a neighbor term and substituting $H^{(1)}$ yields

$$\tilde{e}_{k \rightarrow i} = a_{\text{self}}^\top H_i^{(1)} + a_{\text{nb}}^\top H_k^{(1)} = a_{\text{self}}^\top W^{(1)}x_i + a_{\text{nb}}^\top W^{(1)}x_k.$$

By Definition 2, $\text{FR}_{\text{self}} = (W^{(1)})^\top a_{\text{self}}$ and $\text{FR}_{\text{nb}} = (W^{(1)})^\top a_{\text{nb}}$ (both column vectors), so

$$\tilde{e}_{k \rightarrow i} = (\text{FR}_{\text{self}})^\top x_i + (\text{FR}_{\text{nb}})^\top x_k.$$

Taking the difference between neighbors k and k' cancels the focal term and gives

$$\tilde{e}_{k \rightarrow i} - \tilde{e}_{k' \rightarrow i} = (\text{FR}_{\text{nb}})^\top (x_k - x_{k'}),$$

as claimed. The post-activation score $e_{k \rightarrow i} = \text{LeakyReLU}(\tilde{e}_{k \rightarrow i})$ (Eq. (4)) does not admit the same closed-form decomposition; FR is defined on the linear pre-activation score $\tilde{e}$. $\square$

D.2. Proof for Proposition 2

Proof for Proposition 2 By construction, the attention weights $\alpha_{k \rightarrow i}$ used in Eq. (6) are obtained by applying a softmax to the logits $e_{k \rightarrow i}$ (Eq. (5)), so $\alpha_{k \rightarrow i} \geq 0$ and $\sum_{k \in \mathcal{N}_i} \alpha_{k \rightarrow i} = 1$ for each i . Eq. (6) sets $H_i^{(2)} = \sum_{k \in \mathcal{N}_i} \alpha_{k \rightarrow i} H_k^{(1)}$, which is a convex combination of the neighbor embeddings. Therefore $H_i^{(2)}$ lies in the convex hull of $\{H_k^{(1)} : k \in \mathcal{N}_i\}$. $\square$

D.3. Proof for Proposition 3

Proof for Proposition 3 Algorithm 2 takes as input MG-GAT’s prediction-time business-side NIG signal $\text{NIG}_\theta^b(\cdot, j)$ (used to select $S_\theta(u, j)$ from $\mathcal{H}_u$), the business-side FR coefficient vectors $\{\text{FR}_{\text{self}}^b, \text{FR}_{\text{nb}}^b\}$, and the observed attributes of the recommended business x_j and the selected neighbors $X(S_\theta(u, j))$. The plan-selection step deterministically chooses a subset of neighbors and features from these inputs (using the NIG-weighted instance-level contribution score $\text{score}(f)$, yielding the high-saliency feature set $\mathcal{F}_\theta^*(u, j)$), and the prompt-construction step includes only this selected information. Critics then score candidate plans and explanations under a fixed decoding configuration. The only randomness arises from the LLM’s internal sampling given this fixed configuration. Therefore, there exists a (possibly randomized) mechanism Φ_θ mapping $(\text{NIG}_\theta^b(\cdot, j), \text{FR}_{\text{self}}^b, \text{FR}_{\text{nb}}^b, x_j, X(S_\theta(u, j)), \mathcal{F}_\theta^*(u, j))$ to the distribution of $E(u, j)$, so $E(u, j)$ is parameter-grounded by Definition 3. User-side NIG and FR are part of MG-GAT’s Stage-1 prediction architecture but are not exposed to the Stage-2 prompts, and therefore do not enter the explanation mechanism Φ_θ . $\square$

D.4. Proof for Corollary 1

Proof for Corollary 1 Because θ is fixed, the business-side FR coefficient vectors $\{\text{FR}_{\text{self}}^b, \text{FR}_{\text{nb}}^b\}$ are identical across all recommendation instances. We treat $X(S_\theta(u, j))$ as an ordered tuple under a canonical neighbor ordering (e.g., descending business-side NIG weight), so equality of $X(S_\theta(u, j))$ and $X(S_\theta(u', j'))$ is understood as equality of ordered tuples. Moreover, the instance-level contribution score $\text{score}(f)$ is a

deterministic function of $(\text{NIG}_\theta^b(:, j), x_j, X(S_\theta(u, j)))$ given fixed θ , so the high-salience feature set $\mathcal{F}_\theta^*(u, j)$ is itself determined by this tuple. Therefore, if the tuples stated in Corollary 1 coincide for (u, j) and (u', j') , then $\mathcal{F}_\theta^*(u, j) = \mathcal{F}_\theta^*(u', j')$ and the full input tuples to Φ_θ in Proposition 3 coincide. By Definition 3 and Proposition 3 we have

$$E(u, j) \sim \Phi_\theta(\cdot) \quad \text{and} \quad E(u', j') \sim \Phi_\theta(\cdot)$$

with the same argument. Therefore $E(u, j)$ and $E(u', j')$ have the same distribution, i.e., $E(u, j) \stackrel{d}{=} E(u', j')$.

D.5. Proof for Proposition 4

Proof for Proposition 4 Algorithm 2 takes as input the recommended business j , the selected neighbor set $S_\theta(u, j)$ (from **NIG**), and the selected feature set $\mathcal{F}_\theta^*(u, j)$ (top- F by the **NIG**-weighted instance-level contribution score $\text{score}(f)$). The prompts constructed for planning and explanation include only these entities and attributes. Moreover, we enforce the evidence contract explicitly: any candidate plan or explanation that cites restaurants or features outside $\{j\} \cup S_\theta(u, j)$ and $\mathcal{F}_\theta^*(u, j)$ is rejected before scoring/selection. Therefore, explanations that survive selection are support-limited to $\{j\} \cup S_\theta(u, j)$ and $\mathcal{F}_\theta^*(u, j)$, establishing the claimed support limitation. $\square$

Appendix E. Summary Statistics of the Yelp Data

E.1. Data Descriptions

We apply MG-GAT to the publicly available Yelp dataset (from 2009 to 2018), an online review platform where users may rate and post reviews on businesses (e.g., restaurants).⁹ Yelp shares data from eleven states and provinces; among them, we chose two representative locations in two countries: Ontario (ON) in Canada and Pennsylvania (PA) in the United States. These regions encompass 135,173 and 76,865 users, and 32,393 and 10,966 businesses, respectively. We summarize the statistics in Table E1 and in Figure E1.

Business. Yelp’s business information comprises self-uploaded data from business owners and user reviews. Their diversity and high dimensionality make the data ideal for evaluating the effectiveness of the proposed method. The model uses three types of auxiliary information about the businesses: basic information (attributes, categories, and operation hours); location data; and check-in information reflecting temporal popularity.

1. The basic information that Yelp collects consists of three elements: business attributes (amenities), business categories, and operation hours. The attribute information collected in Canada is different from that collected in the United States: It includes 84 attributes in ON and 93 attributes in PA. Business attributes include parking options, WiFi availability, and takeout service. Because most of the attributes are categorical variables, we adopt one-hot (i.e., one-of-K) encoding to indicate whether a business has a particular business attribute. The business category data (e.g., Mexican, burgers) include 953 and 946 categories in ON and PA, respectively. In this second type of information, each business may belong to multiple categories. Thus, we similarly adopt the one-hot encoding on business categories. Meanwhile, the operation hours contain information about when businesses open and close on each day of the week.

2. The location data, in latitude and longitude, provides spatial information on businesses (Figure E2).

3. The check-in information from users allows us to analyze the temporal patterns of the popularity of the businesses. We aggregate the check-ins into 144 hourly bins in a week to obtain one vector for each business.

We construct a business network $\mathbf{G}_b$ based on three key conditions among its members: (1) geographical proximity, (2) consumer co-visitations, and (3) shared business category.

Geographical proximity captures spatial dependencies between businesses, as recognized in the spatial econometrics literature. These dependencies may arise from positive demand spillovers, spatial competition, and cannibalization effects (Pancras et al. 2012). Co-visitation data provide additional predictive value, reflecting economic ties based on consumer behavior. Businesses in the same category or with similar attributes affect each other through shared consumer markets and competition, a principle adopted in prior studies (Grieser et al. 2022).

Because the network created through these three mechanisms is dense, we construct a k -nearest neighbor graph for each mechanism to optimize memory and computational efficiency. A common approach is to set $k \sim \log(m)$, where m is the total number of businesses. Here, k is set at ten for each edge type (geographically close, consumer co-visits, and business category), bounding the out-degree contribution of each component at ten. The resulting combination forms the business network $\mathbf{G}_b$.

⁹ The data was accessed via <https://www.yelp.com/dataset/>.

User. User information can be categorized into two types: the basic metadata and friendships on Yelp. The basic metadata on users includes whether they are elite reviewers in a certain year and detailed compliments about their work.¹⁰ This auxiliary information leads to 19 attributes for each user in ON and 33 attributes for each user in PA. Yelp also provides a list of each Yelp user’s friends. With this information, we can build a friendship network, where a connection on the network indicates that the two users are friends.

Implicit features. Our model incorporates both explicit and implicit user feedback. Explicit feedback is derived from user ratings, while implicit feedback captures users’ interactions with businesses. Implicit feedback assumes that users demonstrate stronger preferences for businesses they have engaged with. This assumption has been validated in several studies (Koren 2008, Chen et al. 2019) and through the Netflix competition (Rendle et al. 2019), where implicit feedback was shown to be highly predictive in RSs.

To generate implicit features, we binarize continuous user ratings to indicate whether a user-business interaction has occurred. To reduce complexity, we then apply singular value decomposition (SVD) on the binarized rating matrix. The resulting implicit feedback is represented as $\mathbf{X}_{\text{bina}} = \mathbf{U}^{(0)}\Sigma\mathbf{B}^{(0)}$, where $\mathbf{U}^{(0)} \in \mathbb{R}^{n \times k_i}$, $\Sigma \in \mathbb{R}^{k_i \times k_i}$, and $\mathbf{B}^{(0)} \in \mathbb{R}^{k_i \times m}$. Here, k_i represents the dimension of the implicit features and is treated as a hyperparameter.

Specifically, we have binarized the continuous user ratings by setting a threshold where any rating greater than zero (i.e., there exists a rating) is converted to a value of one ($X_{uv} > 0 \Rightarrow X_{\text{bina},uv} = 1$), and a rating of zero remains zero ($X_{uv} = 0 \Rightarrow X_{\text{bina},uv} = 0$). This binary feature therefore captures whether any interaction between the user and business occurred, regardless of the specific rating value.

The user and business implicit features are then calculated as $\mathbf{S}_{u,\text{imp}} = \mathbf{U}^{(0)}\Sigma^{\frac{1}{2}} \in \mathbb{R}^{n \times k_i}$ and $\mathbf{S}_{b,\text{imp}} = \mathbf{B}^{(0)T}\Sigma^{\frac{1}{2}} \in \mathbb{R}^{m \times k_i}$, respectively. These implicit features are combined with explicit features to form the auxiliary information, $\mathbf{S}_u$ for users and $\mathbf{S}_b$ for businesses.

If the implicit features are found to be non-predictive, the parameters governing their influence—specifically, the dimensions in $\mathbf{W}_b^{(1)}$, $\mathbf{W}_u^{(1)}$, $\mathbf{W}_{bs}^{(2)}$, and $\mathbf{W}_{us}^{(2)}$ —are regularized toward zero.

E.2. Experimental Setting

We split the rating data into training, validation, and test sets based on time. The data between 2009 and 2016 are used for training, the 2017 data are used for validation, and the 2018 data are used for testing. This sample split aligns with practical RSs to predict “future” ratings. We present the statistics for the three subsets in Table E2.

The location information, in terms of latitude and longitude, is shown in Figures E2a and E2b. Additionally, Figure E2c presents the spatial distances for the top 10 closest businesses.

We present the statistics about the training, validation, and test sets of the datasets in Table E2.

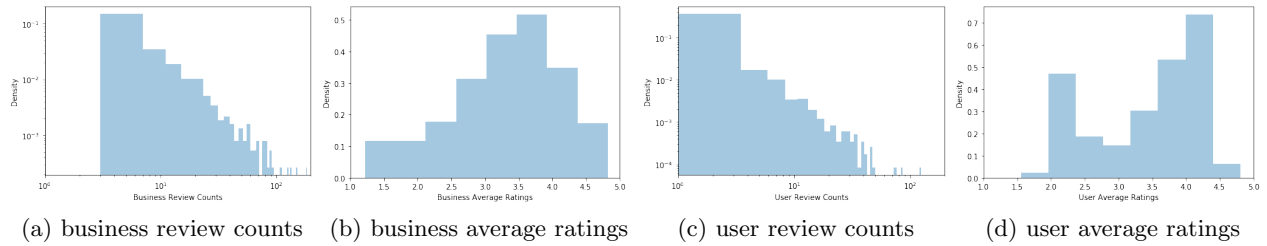
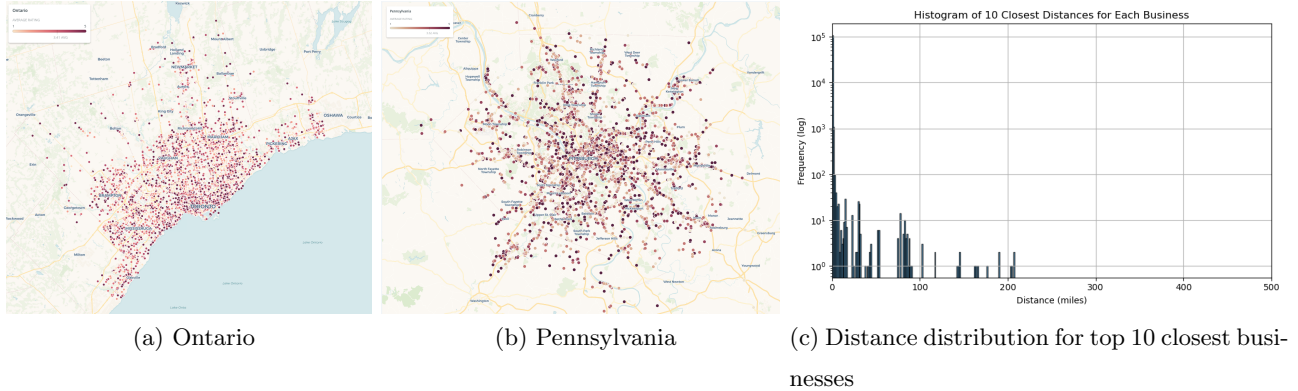
We summarize the statistics of the ratings for ON and PA in Table E1. We show the distributions of the number of reviews of each business and user in Figures E1a and E1c, respectively, both of which follow a power-law distribution. The distributions of the average ratings of businesses and users are shown in Figures E1b and E1d.

¹⁰ The compliments include the number of “useful,” “funny,” and “cool” reviews; the number of fans; the number of compliments that describe reviews as being “hot,” “cute,” “plain,” “cool,” “funny,” or “good writer”; and the number of compliments on the users’ profile, lists, notes, photos, and other information.

Table E1 Summary statistics of the data for Ontario and Pennsylvania.

Statistic	Ontario	Pennsylvania
Rating count	706,998	260,350
User count	135,173	76,865
Business count	32,393	10,966
Average rating (std. dev.)	3.556 (1.334)	3.728 (1.384)
Ratings per user (std. dev.)	5.230 (21.007)	3.387 (12.140)
Ratings per business (std. dev.)	21.826 (47.221)	23.742 (56.371)
Average degree per user (std. dev.)	7.077 (34.738)	5.557 (29.906)
Average degree per business (std. dev.)	22.009 (20.067)	23.105 (22.342)

Note: The business network aggregates three k-NN components ($k = 10$ each): a spatial graph (geographic nearest neighbors), a co-visitation graph (frequently co-visited businesses), and a same-category graph. The reported degree statistics are computed on the symmetrized aggregate adjacency $\mathbf{G}_b$ used by the model.

**Figure E1** Distributions of review counts and average ratings for businesses and users.**Figure E2** Spatial distributions of businesses in (a) and (b); the color code represents the average ratings of businesses. The spatial distributions between businesses and their ten closest businesses are shown in (c).**Table E2** Statistics about the training, validation and test set

	Time period	Metrics	Ontario	Pennsylvania
Training	2009 - 2016	Average user rating	3.468 (1.373)	3.641 (1.403)
		Average business rating	3.419 (0.995)	3.604 (1.049)
Validation	2017	Average user rating	3.463 (1.479)	3.665 (1.501)
		Average business rating	3.401 (1.266)	3.614 (1.285)
Test	2018	Average user rating	3.469 (1.502)	3.667 (1.535)
		Average business rating	3.395 (1.314)	3.616 (1.373)

Appendix F. Performance Evaluation Metrics

The RMSE is computed as $\sqrt{\frac{\|\mathbf{\Omega}_{\text{test}} \odot (\mathbf{X} - \hat{\mathbf{X}})\|_2^2}{\|\mathbf{\Omega}_{\text{test}}\|_1}}$, where $\mathbf{\Omega}_{\text{test}}$ is the indicator matrix, with one for the entries in the test set and zero otherwise, and where $\|\cdot\|_1$ represents the entry-wise ℓ_1 norm.

In addition to the RMSE, we use three ranking-based metrics: Spearman’s rank-order correlation (Spearman’s correlation), Bayesian personalized ranking (BPR) (Rendle et al. 2009), and Fraction of Concordant Pairs (FCP) (Koren and Sill 2013). These ranking-based metrics are defined on a rating pair level, instead of on a single-rating level. Spearman’s correlation is defined by comparing the predicted and the actual rankings for all predictions. In contrast, BPR and FCP are user-level pairwise ranking metrics, which compare the ranking pairs for each user. Therefore, these two metrics are evaluated only on users who have more than one rating in the test set (30% of users in PA and 37% of users in ON). BPR is a pairwise personalized ranking loss derived from the maximum posterior estimator, which is computed as $\text{BPR} = e^{\frac{1}{|D_{\text{test}}|} \sum_{i,j,j' \in D_{\text{test}}} \ln \sigma(\hat{X}_{ij} - \hat{X}_{ij'})}$, where D_{test} consists of strictly ordered pairs in the test set for which $X_{ij} > X_{ij'}$, for each user i (equal-rating pairs are excluded); and $\sigma(\cdot)$ represents the sigmoid function. The BPR represents the geometric mean of the data likelihood, and a higher BPR corresponds to a better personalized ranking. FCP measures the correctly ranked business pairs for each user in the RS, restricted to strictly ordered pairs (equal-rating pairs are excluded). The number of correctly ranked (concordant) business pairs for user i , based on predicted ratings, is $n_c^i = |\{(j, j') \mid X_{ij} < X_{ij'} \text{ and } \hat{X}_{ij} < \hat{X}_{ij'}\}|$. The number of discordant pairs for user i is similarly defined as $n_d^i = |\{(j, j') \mid X_{ij} < X_{ij'} \text{ and } \hat{X}_{ij} > \hat{X}_{ij'}\}|$. FCP is defined as $\frac{\sum_{i=1}^n n_c^i}{\sum_{i=1}^n n_c^i + \sum_{i=1}^n n_d^i}$, and a higher value represents more concordant pairs.

A number of CS studies use top- N metrics to evaluate the efficacy of RSs. Here, top- N metrics determine whether the recommended list contains the top- N items (such as recall@ N , precision@ N , and nDCG) from the groundtruth list for each user. In evaluating these top- N related metrics, many of these studies sample the test set to include only users that provide a minimum number of ratings—for example, choosing a threshold of ten (Chen et al. 2019). However, because of data sparsity issues, using top- N metrics also results in excluding a significant number of users from our data. For instance, in the Yelp dataset, the distribution of users rating a particular number of items is as follows: ON [1 (63%), 2 (15%), 3 (7%), 4 (4%), 5 (2%), 6+ (8%)]; PA [1 (70%), 2 (14%), 3 (5%), 4 (3%), 5 (2%), 6 and above (6%)]. This issue may elucidate why the management studies referenced above do not use this metric. Moreover, setting a value for N in the prediction task can seem arbitrary, but it also can markedly influence the results, depending on the distribution of ratings. If we choose to use a top- N metric, we are faced with two choices: We either must select a tiny value for N , rendering the evaluation task overly simplistic and less informative, or we must exclude users who have not visited a certain number of stores, thereby introducing a systematic bias into the sample. Neither option is suitable for our evaluation.

Appendix G. Experiment Setup

This section complements Section 6 by documenting the randomized controlled experiment in greater detail. We first present the user interface for each experimental condition, then report omnibus and pairwise statistics for every questionnaire item, assess randomization balance across demographic and behavioral covariates, and summarize mediation models linking explanation-quality dimensions (transparency, trust, effectiveness, efficiency, persuasiveness, and satisfaction) to perceived relevance and future interest. These materials are intended to facilitate replication and to document how the experimental evidence supports our claims about explanation quality and user acceptance.

G.1. Branching Questionnaire and Scenario Assignment

Participants completed a short branching questionnaire to elicit key preferences for their closest friend. We first asked whether the friend typically prefers fast food. Fast-food preferences routed respondents to one of two focal fast-food restaurants (*Wonton Chai Noodle* or *One2 Snacks*), with a follow-up on table service. If the friend did not prefer fast food, respondents chose between American and Italian cuisine to keep the branching structure tractable and aligned with Yelp coverage. For American cuisine, one of three follow-up questions (full bar, reservations, or event planning services) was randomly sampled to select one of two focal American restaurants (*Fanzone Wings & Ribs* or *Farm'r Eatery & Catering*); for Italian cuisine, a credit-card acceptance question routed to *Pastizza* or *Campo*. For each path, participants were shown two previously liked restaurants consistent with the elicited preferences, followed by a focal recommendation produced by MG-GAT for a matched user in Yelp. The full branching structure is shown in Figure G1.

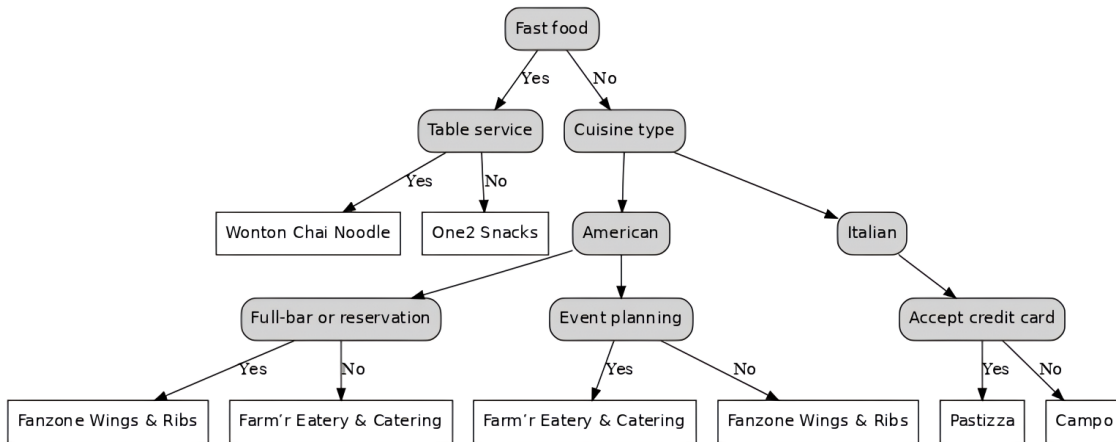


Figure G1 Preference-conditioned experiment design showing how survey questions lead to final recommendations.

G.2. On-Screen Text for Illustrative Path

For reproducibility, we record the exact on-screen text shown in the illustrative American/full-bar path (*Fanzone Wings & Ribs*). Screenshots for all conditions are in Figures G2–G7. The recommendation explanations below address the reader as “you” (the platform speaks to the imagined friend), while the questionnaire

items ask participants to evaluate on behalf of their friend. This second-person stimulus phrasing is identical across all conditions.

- **Context blurb (before the recommendation).** “Imagine that your friend has visited the following two restaurants, both of which he/she rated 4 out of 5. Each restaurant offers chicken wings, convenient bike parking, and a casual atmosphere, making them suitable for a relaxed meal. Whether for lunch or dinner, both establishments are equally ideal.”
- **Treatment I (Relevant Business).** “Fanzone Wings & Ribs is recommended because it closely resembles two restaurants that you already enjoy and have expressed a liking for: Church’s Chicken and Quik Chik. The similarities between these restaurants and Fanzone Wings & Ribs provide a strong basis for believing you would likely appreciate Fanzone’s offerings as well, making it a suitable recommendation based on your preferences”
- **Treatment II (Park et al. (2017) Social).** “Fanzone Wings & Ribs is recommended because you have enjoyed other similar dining establishments in the past, and we believe this restaurant aligns with your preferences. These establishments are considered similar because they are consistently favored by your friends or people you follow who share similar tastes and preferences with you. As a result, we are confident that you will have a delightful and satisfying experience at this recommended restaurant.”
- **Treatment III (SHAP features).** “Fanzone Wings & Ribs is a great match for your tastes, offering the takeout convenience and casual vibe you’ve enjoyed at Church’s Chicken and Quik Chik. With free WiFi, kid-friendly options, and ample parking, it caters to both comfort and practicality, just like your previous favorites. Plus, its lively bar scene adds a new layer of fun that complements your past dining preferences without straying too far.”
- **Treatment IV (MG-GAT parameter-grounded).** “If you loved the bike-friendly spots and delicious chicken wings at Church’s Chicken and Quik Chik, you’ll feel right at home at Fanzone Wings & Ribs, which shares these features. With its casual ambiance and suitability for both lunch and dinner, Fanzone Wings & Ribs perfectly matches your dining preferences. Don’t miss the chance to savor a meal that fits your tastes-give Fanzone Wings & Ribs a try!”

G.3. Interface of the Experiment Conditions

Figures G2–G7 present the interfaces for all experiment conditions grouped by restaurant.

G.4. Statistical Analysis of All Questions

We next present a detailed analysis of our experimental results in Table G1 across five experimental conditions: No Explanation (No EXP), Relevant Business Explanation (Relevant Business EXP), Park et al. (2017) Explanation (Park EXP), SHAP Explanation (SHAP EXP), and MG-GAT Explanation (MG-GAT EXP). Each condition’s performance was evaluated across several questions using the mean score and standard deviation (SD) on a 7-point Likert scale. An Analysis of Variance (ANOVA) was conducted to assess the statistical significance of differences in mean scores among these conditions, followed by pairwise comparisons.

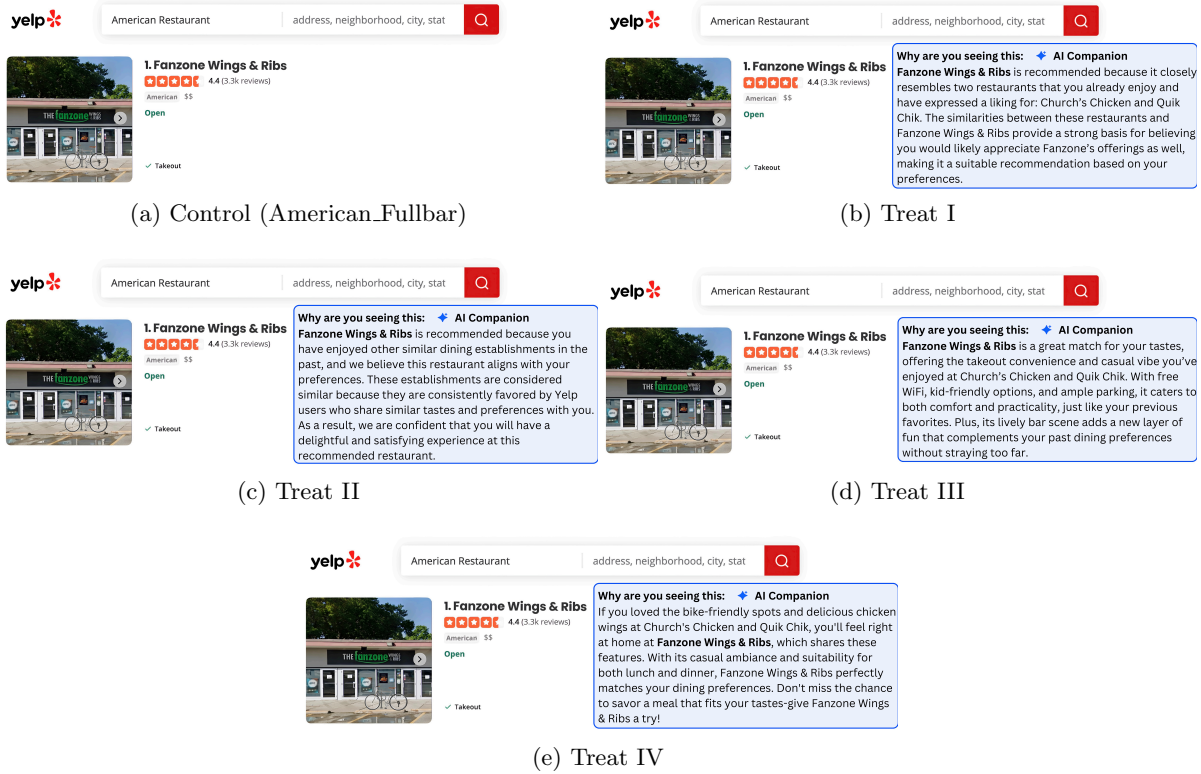


Figure G2 Interfaces for Fanzone Wings & Ribs (American_Fullbar)

- **Match with User Interests, Engagement, and Satisfaction:** For queries pertaining to the match between the recommended item and user interests, as well as the probability of users engaging with the recommendation, the MG-GAT-Based Explanation condition achieved the highest mean ratings across the five conditions. This pertains to the statements, “The recommended item, ‘Fanzone Wings & Ribs,’ matched your friend’s interests,” “Your friend would click on the recommendation, ‘Fanzone Wings & Ribs,’” and “Your friend would be satisfied with this recommendation.” The ANOVA test revealed a significant difference across conditions for all statements (with $p < 0.001$).
- **Understanding of Recommendation System and Confidence in Its Effectiveness:** The MG-GAT-Based Explanation condition also achieved the highest mean ratings in statements related to understanding the recommendation system and the user’s confidence in its effectiveness. These statements include: “This explanation helps your friend to understand what the recommendation is based on,” “I believe my friend has gained insight into why ‘Fanzone Wings & Ribs’ was recommended to him/her,” and “This explanation increases your friend’s confidence in the recommender system.” Once again, the ANOVA results demonstrated significant variances across the conditions for all statements (with p-values less than 0.001).
- **Usefulness of Explanations and Comfort in Relying on Them:** For items assessing usefulness, decision efficiency, persuasiveness, and ease of relying on explanations, the MG-GAT-Based Explanation condi-

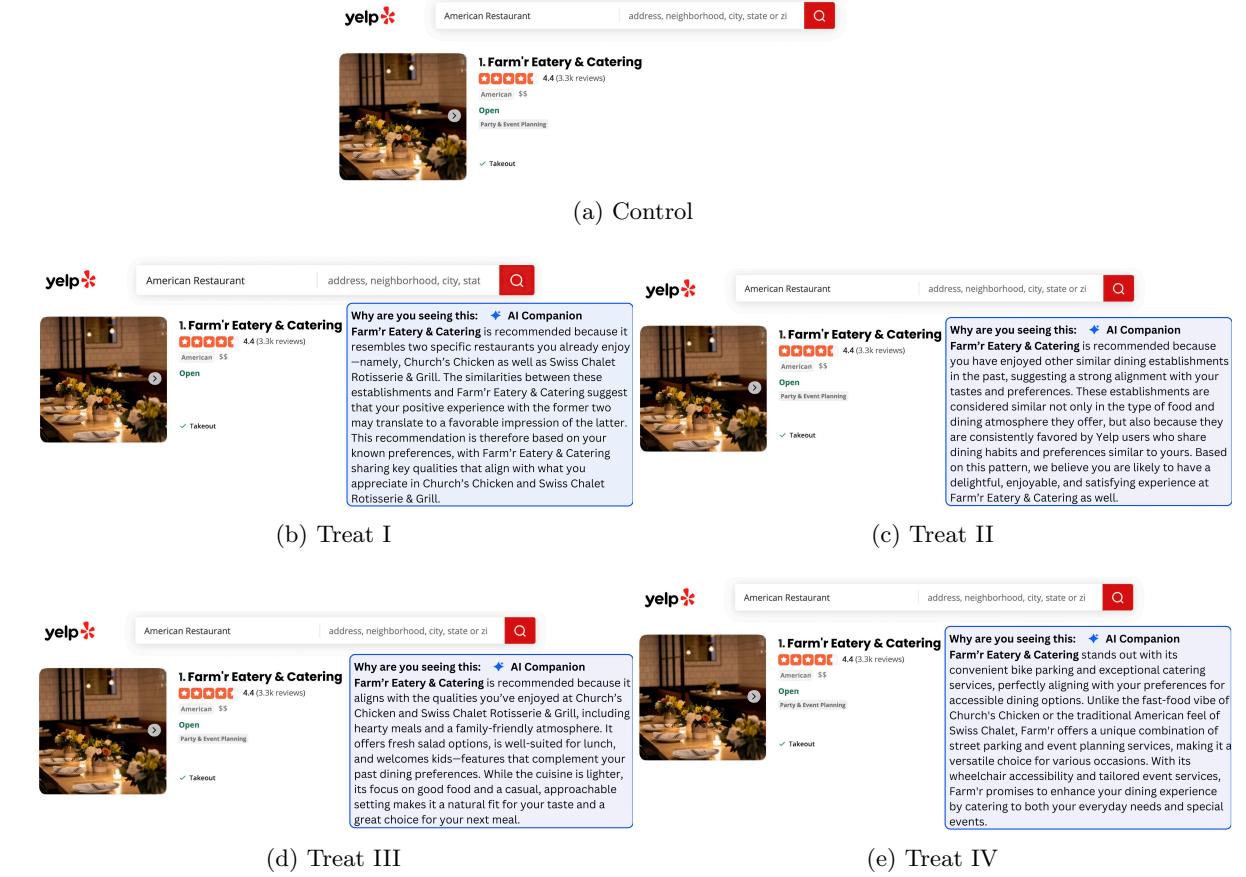


Figure G3 Interfaces for Farm'r Eatery & Catering (American_noFullbar/Reservation)

tion again achieved the highest mean ratings. ANOVA indicates significant differences across conditions for these statements ($p < 0.001$).

- **Future Engagement and Willingness to Explore:** Regarding the statement on future engagement, “Your friend would feel more inclined to try other recommended businesses provided by the same platform in the future,” the MG-GAT-Based Explanation condition displayed a significantly higher mean score; ANOVA and pairwise significance levels are reported in Table G1.

In summary, the MG-GAT Explanation condition achieves the highest mean ratings across engagement and explanation-quality measures. Pairwise comparisons show that, relative to the SHAP Explanation condition, the MG-GAT condition significantly improves perceived relevance, future interest, trust, persuasiveness, and satisfaction, while differences on transparency, efficiency, and effectiveness are not statistically distinguishable (Table G1).

G.5. Balance Check

We first affirmed balance across the five conditions through a chi-squared test for gender and race. The outcomes of these tests were not statistically significant, suggesting no differences in the distribution of these demographic variables across the groups. Consequently, this indicates that the groups were demographically balanced, and any variations in responses cannot be ascribed to these demographic factors. To further

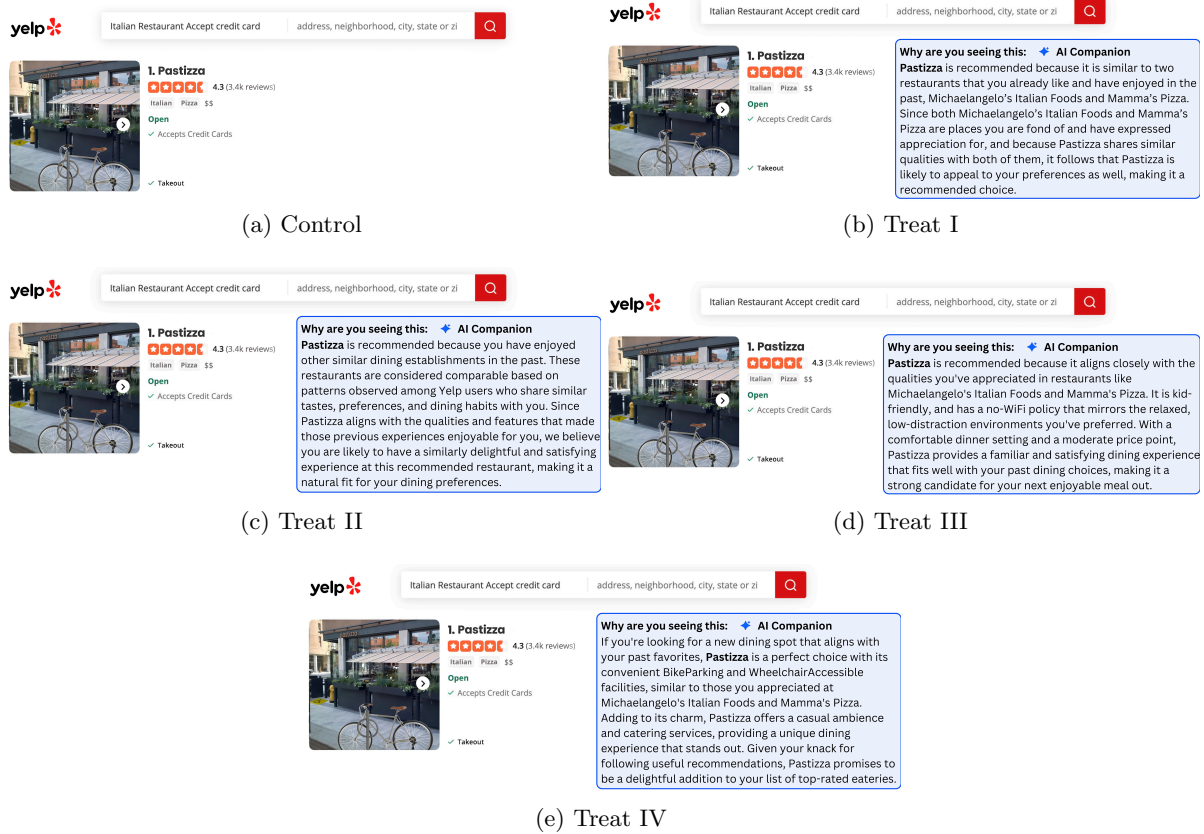


Figure G4 Interfaces for Pastizza (Italian.CreditCards)

ensure the equivalence of the groups, we conducted a one-way ANOVA for age, propensity to explore new restaurants, and years of Yelp usage. The results, being non-significant, suggest no statistically meaningful disparities among the groups concerning these variables. Therefore, any differences in users' perception of the recommendations cannot be ascribed to their age, willingness to explore new restaurants, or familiarity with Yelp. Furthermore, we employed the Tukey multiple comparisons of means to bolster the above findings. All comparisons turned out to be non-significant, providing robust evidence of no substantial differences among the conditions in terms of the variables we scrutinized.

G.6. Mediation Analysis: How Explanations Mediate Users' Perceived Relevance and Future Interest in the Recommendation

We conducted a mediation analysis to assess the underlying mechanisms through which different explanations affect Perceived Relevance and Future Interest. Specifically, we conducted three separate regression models to explore the mediation effects:

Model 1: $X \rightarrow Y$ (Total Effect): Regress the dependent variable (Y) on the independent variable (X). This step estimates the total effect of X on Y. Here, the independent variables correspond to the experimental condition indicators, namely the Relevant Business Explanation condition, the Park et al. (2017) Explanation condition, the MG-GAT condition, and the SHAP Explanation condition. The dependent variables in the analysis are Perceived Relevance and Future Interest. Regression results are displayed in the first column of

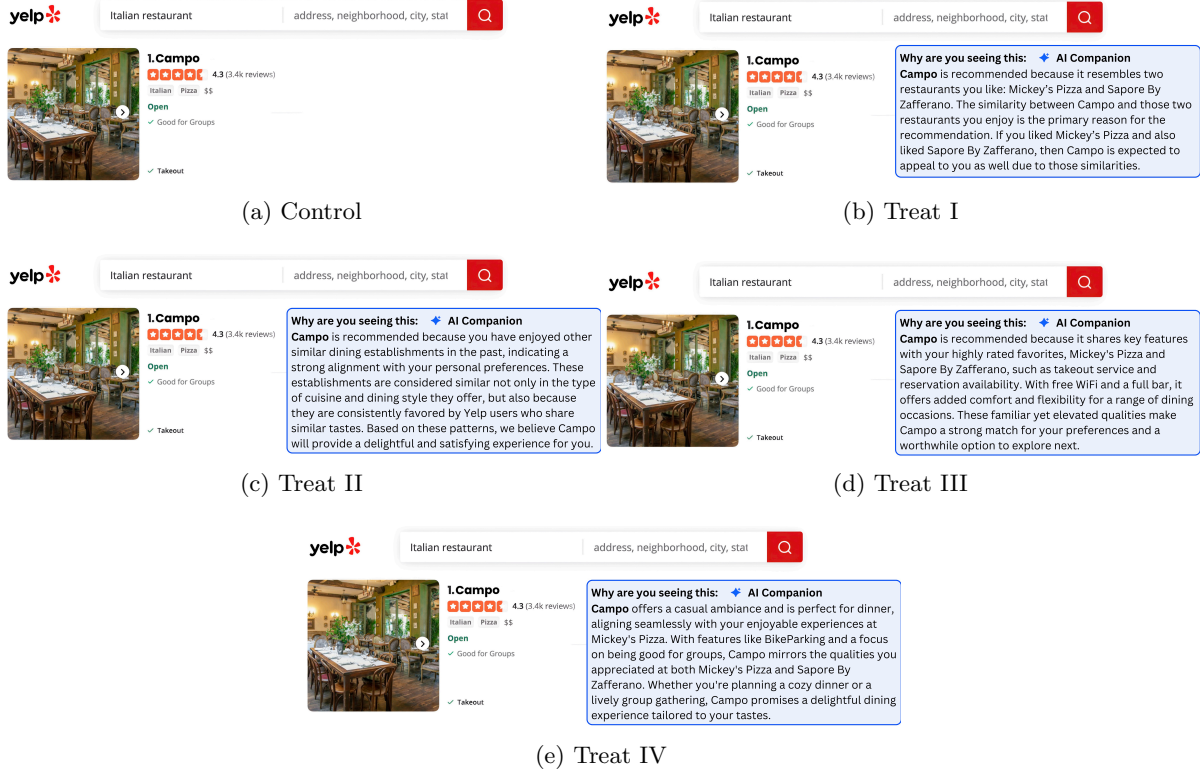


Figure G5 Interfaces for Campo (Italian_noCreditCards)

Table G2. The MG-GAT Explanation condition serves as the reference group for the baseline intercept term. Panels (1) to (6) present results for Perceived Relevance, while panels (7) to (12) provide results for Future Interest.

Model 2: $X \rightarrow M$: Regress the mediator (M) on the independent variable (X). This step estimates the effect of X on M. We consider six mediators: transparency, trust, effectiveness, efficiency, persuasiveness, and satisfaction. Results are presented in the second column of Tables G2.

Model 3: $X + M \rightarrow Y$: Regress the dependent variable (Y) on both the independent variable (X) and the mediator (M). This will estimate the effect of M on Y and the direct effect of X on Y. We report how the estimated treatment coefficients change when each explanation-quality dimension is included; we interpret coefficient attenuation patterns as exploratory evidence consistent with mediation-style pathways. The results are presented in the third column of Tables G2.

Across dimensions, including the mediator attenuates the treatment coefficients; we report these patterns as exploratory associations rather than claims of “full mediation.”

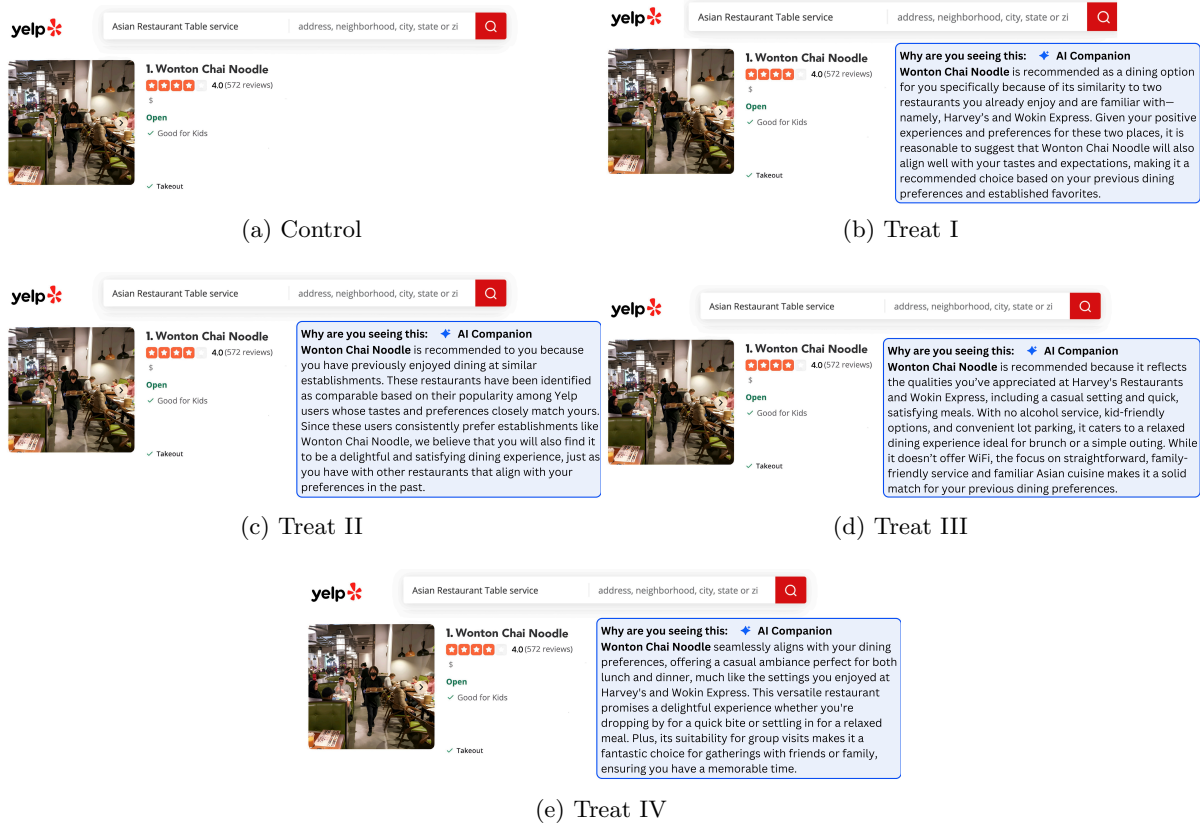
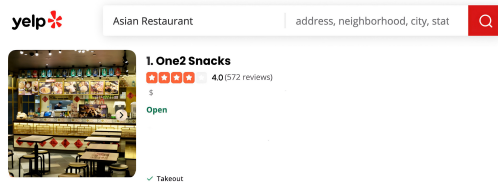
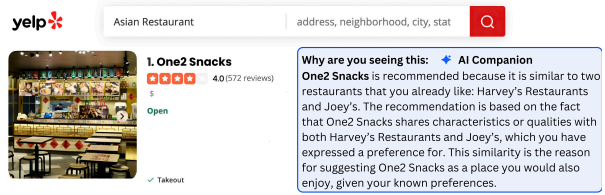


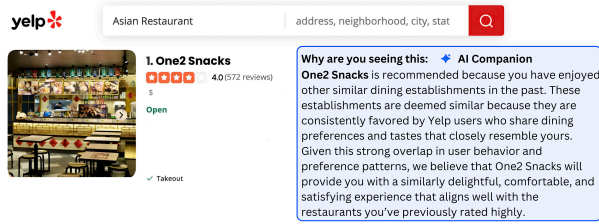
Figure G6 Interfaces for Wonton Chai Noodle (Fastfood_Table)



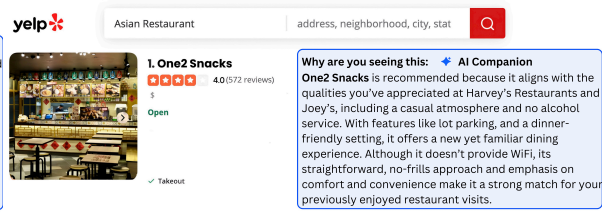
(a) Control (Fastfood_noTable)



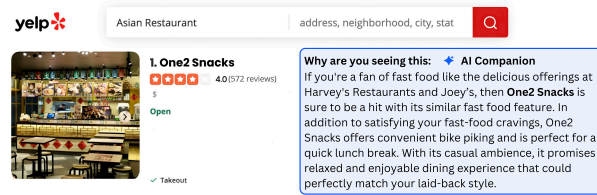
(b) Treat I



(c) Treat II



(d) Treat III



(e) Treat IV

Figure G7 Interfaces for One2 Snacks (Fastfood_noTable)

Table G1 Participants' ratings on their satisfaction, explanation effectiveness, and influence on their behavior

Questions	No EXP		RB EXP		Park EXP		SHAP EXP		MG-GAT EXP		ANOVA <i>p-value</i>	MG-GAT vs. No EXP	MG-GAT vs. RB	MG-GAT vs. Park	MG-GAT vs. SHAP
	M	SD	M	SD	M	SD	M	SD	M	SD					
I believe my friend has gained insight into why [recommended restaurant] was recommended to him/her.	3.27	1.59	4.91	1.51	5.12	1.53	5.45	1.37	5.71	1.12	<.001	<.001	<.001	<.001	<.05
Your friend would click on the recommendation, [recommended restaurant]	4.45	1.50	5.14	1.40	5.21	1.34	4.85	1.66	5.50	1.33	<.001	<.001	<.01	<.05	<.001
Your friend would feel more inclined to try other recommended businesses provided by the same platform in the future.	4.31	1.32	5.02	1.27	4.97	1.34	4.81	1.58	5.24	1.33	<.001	<.001	<.05	<.05	<.01
The recommended item, [recommended restaurant], matched your friend's interests.	4.77	1.55	5.47	1.25	5.48	1.19	4.83	1.82	5.82	1.19	<.001	<.001	<.001	<.01	<.001
Your friend would be satisfied with this recommendation.	4.48	1.52	5.14	1.44	5.11	1.49	4.80	1.77	5.52	1.36	<.001	<.001	<.01	<.001	<.001
The method behind the recommendation system is competent and effective.	4.26	1.49	4.88	1.62	4.95	1.63	4.70	1.75	5.42	1.31	<.001	<.001	<.001	<.001	<.001
This explanation helps your friend to understand what the recommendation is based on. (Transparency)	–	–	4.89	1.87	5.03	1.83	5.64	1.31	5.83	1.20	<.001	–	<.001	<.001	0.097
This explanation increases your friend's confidence in the recommender system. (Trust)	–	–	4.73	1.68	4.76	1.72	4.73	1.72	5.35	1.40	<.001	–	<.001	<.001	<.001
This explanation helps your friend decide efficiently whether to visit [recommended restaurant]. (Efficiency)	–	–	4.79	1.70	4.87	1.69	5.33	1.42	5.48	1.31	<.001	–	<.001	<.001	0.229
This explanation persuades your friend to visit [recommended restaurant]. (Persuasiveness)	–	–	4.66	1.66	4.73	1.67	4.55	1.81	5.23	1.30	<.001	–	<.001	<.001	<.001
This explanation is effective in helping your friend evaluate if he/she should visit [recommended restaurant]. (Effectiveness)	–	–	4.80	1.64	4.84	1.64	5.33	1.46	5.49	1.34	<.001	–	<.001	<.001	0.219
This explanation would improve how easy it is to pick a recommendation. (Satisfaction)	–	–	4.79	1.67	4.77	1.65	4.94	1.62	5.42	1.34	<.001	–	<.001	<.001	<.001

Note: RB stands for Relevant Business.

Table G2 Mediation Analysis Results

(1)	Y (Relevance)			M (Transparency)			Y (Relevance)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.356	0.118	0.003	-0.937	0.138	< 0.001	0.128	0.096	0.184
Park EXP	-0.344	0.120	0.004	-0.792	0.141	< 0.001	0.065	0.098	0.503
SHAP EXP	-0.993	0.125	< 0.001	-0.190	0.145	0.191	-0.895	0.099	< 0.001
M (Transparency)	—	—	—	—	—	—	0.517	0.021	< 0.001
Constant	5.821	0.086	< 0.001	5.825	0.101	< 0.001	2.809	0.142	< 0.001

(2)	Y (Relevance)			M (Trust)			Y (Relevance)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.356	0.118	0.003	-0.620	0.141	< 0.001	0.014	0.084	0.868
Park EXP	-0.344	0.120	0.004	-0.586	0.144	< 0.001	0.006	0.085	0.946
SHAP EXP	-0.993	0.125	< 0.001	-0.620	0.149	< 0.001	-0.623	0.088	< 0.001
M (Trust)	—	—	—	—	—	—	0.597	0.018	< 0.001
Constant	5.821	0.086	< 0.001	5.349	0.103	< 0.001	2.629	0.115	< 0.001

(3)	Y (Relevance)			M (Effectiveness)			Y (Relevance)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.356	0.118	0.003	-0.690	0.132	< 0.001	0.046	0.091	0.616
Park EXP	-0.344	0.120	0.004	-0.650	0.135	< 0.001	0.034	0.092	0.710
SHAP EXP	-0.993	0.125	< 0.001	-0.157	0.139	0.259	-0.902	0.095	< 0.001
M (Effectiveness)	—	—	—	—	—	—	0.582	0.021	< 0.001
Constant	5.821	0.086	< 0.001	5.492	0.097	< 0.001	2.624	0.133	< 0.001

(4)	Y (Relevance)			M (Efficiency)			Y (Relevance)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.356	0.118	0.003	-0.692	0.133	< 0.001	0.038	0.092	0.676
Park EXP	-0.344	0.120	0.004	-0.612	0.136	< 0.001	0.005	0.093	0.960
SHAP EXP	-0.993	0.125	< 0.001	-0.150	0.141	0.287	-0.908	0.095	< 0.001
M (Efficiency)	—	—	—	—	—	—	0.570	0.021	< 0.001
Constant	5.821	0.086	< 0.001	5.480	0.097	< 0.001	2.698	0.133	< 0.001

(5)	Y (Relevance)			M (Persuasiveness)			Y (Relevance)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.356	0.118	0.003	-0.574	0.140	< 0.001	-0.009	0.083	0.915
Park EXP	-0.344	0.120	0.004	-0.509	0.142	< 0.001	-0.037	0.085	0.667
SHAP EXP	-0.993	0.125	< 0.001	-0.689	0.147	< 0.001	-0.577	0.088	< 0.001
M (Persuasiveness)	—	—	—	—	—	—	0.605	0.018	< 0.001
Constant	5.821	0.086	< 0.001	5.234	0.102	< 0.001	2.657	0.114	< 0.001

(6)	Y (Relevance)			M (Satisfaction)			Y (Relevance)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.356	0.118	0.003	-0.628	0.136	< 0.001	-0.006	0.092	0.948
Park EXP	-0.344	0.120	0.004	-0.642	0.139	< 0.001	0.014	0.093	0.882
SHAP EXP	-0.993	0.125	< 0.001	-0.477	0.144	< 0.001	-0.727	0.096	< 0.001
M (Satisfaction)	—	—	—	—	—	—	0.557	0.021	< 0.001
Constant	5.821	0.086	< 0.001	5.417	0.100	< 0.001	2.803	0.130	< 0.001

(7)	Y (Future Interest)			M (Transparency)			Y (Future Interest)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.221	0.119	0.062	-0.937	0.138	< 0.001	0.247	0.099	0.012
Park EXP	-0.272	0.121	0.025	-0.792	0.141	< 0.001	0.123	0.100	0.216
SHAP EXP	-0.435	0.125	< 0.001	-0.190	0.145	0.191	-0.340	0.102	< 0.001
M (Transparency)	—	—	—	—	—	—	0.500	0.022	< 0.001
Constant	5.242	0.087	< 0.001	5.825	0.101	< 0.001	2.331	0.145	< 0.001

(8)	Y (Future Interest)			M (Trust)			Y (Future Interest)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.221	0.119	0.062	-0.620	0.141	< 0.001	0.138	0.087	0.113
Park EXP	-0.272	0.121	0.025	-0.586	0.144	< 0.001	0.067	0.088	0.448
SHAP EXP	-0.435	0.125	< 0.001	-0.620	0.149	< 0.001	-0.077	0.091	0.402
M (Trust)	—	—	—	—	—	—	0.579	0.019	< 0.001
Constant	5.242	0.087	< 0.001	5.349	0.103	< 0.001	2.146	0.119	< 0.001

(9)	Y (Future Interest)			M (Effectiveness)			Y (Future Interest)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.221	0.119	0.062	-0.690	0.132	< 0.001	0.176	0.092	0.057
Park EXP	-0.272	0.121	0.025	-0.650	0.135	< 0.001	0.102	0.094	0.277
SHAP EXP	-0.435	0.125	< 0.001	-0.157	0.139	0.259	-0.345	0.096	< 0.001
M (Effectiveness)	—	—	—	—	—	—	0.576	0.021	< 0.001
Constant	5.242	0.087	< 0.001	5.492	0.097	< 0.001	2.081	0.135	< 0.001

(10)	Y (Future Interest)			M (Efficiency)			Y (Future Interest)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.221	0.119	0.062	-0.692	0.133	< 0.001	0.170	0.093	0.066
Park EXP	-0.272	0.121	0.025	-0.612	0.136	< 0.001	0.074	0.094	0.432
SHAP EXP	-0.435	0.125	< 0.001	-0.150	0.141	0.287	-0.350	0.096	< 0.001
M (Efficiency)	—	—	—	—	—	—	0.566	0.021	< 0.001
Constant	5.242	0.087	< 0.001	5.480	0.097	< 0.001	2.141	0.135	< 0.001

(11)	Y (Future Interest)			M (Persuasiveness)			Y (Future Interest)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.221	0.119	0.062	-0.574	0.140	< 0.001	0.100	0.090	0.267
Park EXP	-0.272	0.121	0.025	-0.509	0.142	< 0.001	0.012	0.092	0.893
SHAP EXP	-0.435	0.125	< 0.001	-0.689	0.147	< 0.001	-0.050	0.095	0.601
M (Persuasiveness)	—	—	—	—	—	—	0.559	0.020	< 0.001
Constant	5.242	0.087	< 0.001	5.234	0.102	< 0.001	2.314	0.123	< 0.001

(12)	Y (Future Interest)			M (Satisfaction)			Y (Future Interest)		
	Coeff.	SE	p	Coeff.	SE	p	Coeff.	SE	p
Relevant Business EXP	-0.221	0.119	0.062	-0.628	0.136	< 0.001	0.128	0.092	0.165
Park EXP	-0.272	0.121	0.025	-0.642	0.139	< 0.001	0.085	0.094	0.368
SHAP EXP	-0.435	0.125	< 0.001	-0.477	0.144	< 0.001	-0.170	0.097	0.079
M (Satisfaction)	—	—	—	—	—	—	0.556	0.021	< 0.001
Constant	5.242	0.087	< 0.001	5.417	0.100	< 0.001	2.232	0.131	< 0.001

Table G3 Regression Results: Impact of Explanation Methods on User Responses

	Willingness to Click					Future Interest				
	Rel. Bus.	Park	MG-GAT	SHAP	Combined	Rel. Bus.	Park	MG-GAT	SHAP	Combined
Intercept	2.154*** (0.190)	2.649*** (0.212)	1.094*** (0.322)	0.918** (0.342)	1.966*** (0.119)	2.117*** (0.160)	1.957*** (0.188)	1.026** (0.317)	1.182*** (0.349)	1.780*** (0.109)
Transparency	0.101 (0.064)	0.147* (0.072)	0.246** (0.076)	0.124 (0.081)	0.078* (0.035)	-0.033 (0.054)	0.098 (0.064)	0.085 (0.075)	0.023 (0.083)	-0.010 (0.032)
Trust	0.153 (0.086)	0.091 (0.083)	0.128 (0.083)	0.455*** (0.088)	0.251*** (0.042)	0.243*** (0.072)	0.087 (0.074)	0.338*** (0.081)	0.332*** (0.089)	0.282*** (0.038)
Effectiveness	-0.017 (0.097)	-0.012 (0.100)	0.162 (0.098)	0.132 (0.087)	0.005 (0.046)	0.150 (0.082)	0.070 (0.089)	0.080 (0.096)	0.016 (0.089)	0.039 (0.042)
Efficiency	0.076 (0.073)	0.104 (0.094)	-0.031 (0.093)	-0.052 (0.103)	-0.004 (0.045)	0.101 (0.061)	0.014 (0.083)	0.053 (0.091)	0.139 (0.105)	0.059 (0.041)
Persuasiveness	0.228** (0.077)	0.197* (0.078)	0.256** (0.087)	0.295*** (0.071)	0.306*** (0.038)	0.071 (0.065)	0.150* (0.069)	0.176* (0.086)	0.218** (0.073)	0.189*** (0.034)
Satisfaction	0.089 (0.069)	0.003 (0.079)	0.041 (0.074)	-0.138 (0.093)	0.020 (0.039)	0.079 (0.058)	0.207** (0.070)	0.049 (0.073)	0.022 (0.094)	0.101** (0.036)
R²	0.502	0.393	0.476	0.596	0.479	0.570	0.525	0.495	0.536	0.520

Note: Standard errors are reported in parentheses. Rel. Bus. = Relevant Business explanation; Combined

= all explanation conditions pooled. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.