Structured vs. LLM-Realized Recommender Explanations Under Shared Recommendation Evidence: A Controlled Human-Centered Comparison

Sidra Naveed[1], Abdullah Abdullah[1] and Gunnar Stevens[1]

[1]University of Siegen, Adolf-Reichwein-Straße 2, 57068 Siegen, Germany

Abstract

Large language models (LLMs) enable fluent recommender explanations, but positive evaluations can conflate the quality of recommendation evidence with the way that evidence is realized and presented. We report a controlled within-subject study comparing two user-facing explanation conditions derived from the same recommendation case: a *Structured Content-Based Recommender Explanation*, implemented as a feature-constraint checklist, and the *LLM-Realized Recommender Explanation*, implemented as grounded prose with a constraint summary. For every scenario, both explanations used the same camera, match score, feature values, and matched or unmet constraints; they differed in linguistic realization, organization, and visual presentation. In the final analytic sample (N = 142), participants compared the explanations on six user-centered constructs. The effects were small and non-uniform. Under uncorrected tests, System 2 was rated more favorably for persuasiveness, trust, reverse-coded cognitive load (indicating lower perceived effort), and decision support; after Holm-Bonferroni correction, only the persuasiveness effect remained statistically significant. Satisfaction did not differ, while System 1 showed a nonsignificant directional advantage in understandability and was favored qualitatively for rapid verification and direct feature-to-constraint checking. The overall preference split did not differ significantly from an even split. Participants nevertheless reported that presentation mattered and contrasted contextual, confidence-supporting prose with rapid checklist-based verification. We describe this sensitivity as a perception–evidence gap: user judgments can vary across explanation conditions that share the same underlying recommendation evidence. The term does not imply unfaithfulness, and this experiment does not isolate wording from layout or generic contextual interpretation. The findings support layered interfaces that preserve inspectable evidence while optionally adding grounded natural-language realization.

Keywords

Recommender Systems, Large Language Models, Recommender Explanations, User-Centered Evaluation, Perception–Evidence Gap

1. Introduction

Recommender systems (RS) help people navigate large choice spaces, but recommendation accuracy alone does not tell users whether an item fits their needs. Explanations can support transparency, trust, satisfaction, persuasion, and decision making, yet these objectives are not interchangeable [1, 2, 3, 4]. Traditional recommender explanations often rely on structured signals such as item attributes, similarities, or templates. These approaches provide consistency and control, but can be labor-intensive to design and may limit contextual flexibility, expressive richness, and linguistic naturalness [5, 6].

LLMs make it easy to produce fluent, contextualized recommendation text [7, 8, 9, 10] and are often perceived more positively [5, 11]. Existing comparisons frequently change several elements at once: an LLM may select evidence, add background knowledge, infer consequences, and realize the result in natural language [12, 13, 5]. When users prefer an LLM output, it is therefore difficult to determine whether they respond to stronger evidence, richer content, or a more compelling presentation. Three recent studies illustrate this conflation concretely. Bismay et al. [12] use LLM reasoning to jointly generate personalized recommendations and their explanations, so the explanation reflects reasoning steps that are not present in, and cannot be isolated from, the underlying recommendation logic. Lubos

IntRS'26: Joint Workshop on Interfaces and Human Decision Making for Recommender Systems, September 28, 2026, Minneapolis
sidra.naveed@uni-siegen.de (S. Naveed); abdullah.abdullah@uni-siegen.de (A. Abdullah); gunnar.stevens@uni-siegen.de (G. Stevens)
0009-0007-6291-9708 (S. Naveed); 0009-0000-8523-513X (A. Abdullah); 0000-0002-7785-5061 (G. Stevens)

et al. [13] use LLMs to generate consequence-oriented explanations describing the potential outcomes of accepting a recommendation, again introducing content beyond the item-level evidence that produced the recommendation. Lubos et al. [5] compare LLM-generated explanations against conventional explanation styles without holding the underlying recommendation evidence fixed across conditions, so observed preferences may reflect differences in evidence, content, or realization jointly rather than realization alone. In each of these studies, evidence, content, and realization change together, so it is not possible to attribute an observed user preference specifically to how an explanation is worded and presented rather than to what it claims. To our knowledge, no existing comparison holds item-level recommendation evidence strictly fixed while varying the complete user-facing explanation condition.

We address this gap with a controlled comparison in which two explanation conditions communicate exactly the same underlying recommendation – the same item, match score, feature values, and satisfied or unmet constraints – but differ in how that evidence is turned into a user-facing explanation. One condition presents the evidence as an explicit, feature-by-feature checklist; the other presents the same evidence as a short, LLM-generated narrative paragraph. We refer to these throughout as the *Structured Content-Based Recommender Explanation* and the *LLM-Realized Recommender Explanation*, respectively. This framing lets us ask the following two research questions before we introduce the fuller conceptual apparatus and contributions below:

- **RQ1.** Under shared recommendation evidence, how do the Structured Content-Based Recommender Explanation and the LLM-Realized Recommender Explanation packages differ across understandability, trust, decision support, satisfaction, cognitive load, and persuasiveness?
- **RQ2.** Which elements do participants report as driving their preference, and what do these judgments reveal about sensitivity to realization and presentation under fixed evidence?

Answering these questions requires separating what an explanation claims from how it is expressed. We separate three layers. *Recommendation evidence* comprises the scenario, recommended item, match score, item features, and matched or unmet constraints that support the recommendation. *Explanation content* comprises the explicit evidence-backed claims communicated to the user. *Realization and presentation* concern wording, tone, organization, visual cues, and layout. In practice, however, these two layers are not fully separable. Because wording and tone can shape which claims a reader infers from a passage, a change classified as realization can, in some cases, also alter the effective explanation content. We illustrate this concretely in Section 3.4: System 2 sometimes explains a matched feature by adding a generic functional rationale — for example, attributing an image stabilizer to "stable shooting" — that is not present in the underlying item data. Such additions are not required to convey the same evidence in fluent prose; they constitute a small increment in content introduced through realization. We treat this overlap as an inherent property of natural-language explanation rather than a limitation of our framework, and we account for it explicitly when interpreting the System 1–System 2 comparison (Sections 3.3, 5.3). We use *perception–evidence gap* to denote a change in user judgment across explanation conditions whose underlying item-level recommendation evidence is held fixed. This differs from faithfulness, which concerns whether claims correspond to the underlying rationale, and from trust calibration, which concerns whether user trust is appropriately aligned with system reliability [14, 15]. A perception–evidence gap can occur even when both outputs are factually grounded. Because System 2 sometimes included generic feature-function interpretations, the conditions were evidence-equivalent but not strictly content-identical.

We conducted a within-subjects online study in the digital-camera domain. We chose digital cameras because their attributes are explicit, enumerable, and directly verifiable (e.g., zoom factor, weight, price), which lets us control and check the underlying recommendation evidence exactly across conditions. This same property may favor checklist-style presentation, so we treat domain attribute structure as a scope condition rather than a neutral choice, and we frame our conclusions as applying primarily to recommendation domains with similarly explicit, checklist-verifiable attributes rather than to recommender explanations in general (Section 5.3). Each participant compared the Structured Content-Based Recommender Explanation with the LLM-Realized Recommender Explanation across six

fixed scenarios. The recommended item and item-level feature evidence were held constant across the two explanations. Importantly, the manipulation *was not wording-only*: the explanations also differed in organization and visual presentation. We therefore estimate the effect of the complete user-facing explanation condition under shared evidence, rather than a pure effect of LLM wording. Section 3.1 formalizes this design as two testable hypotheses (H1, H2).

The paper makes three contributions. We note upfront that comparing structured and LLM-generated explanations is not itself new [5, 9, 11]; our contribution is methodological and measurement-focused. First, it offers a methodological protocol for comparing evidence-equivalent recommender explanations by freezing the scenario, item, match score, feature values, and constraint status, which, to our knowledge, is the first design in this literature to hold recommendation evidence strictly fixed while manipulating the complete user-facing explanation condition. Second, it reports a nuanced empirical pattern: small construct-specific advantages for System 2 in persuasiveness, trust, lower perceived cognitive load, and decision support under uncorrected tests, of which only the persuasiveness effect survives correction for multiple comparisons (Sections 3.7, 4.2); no overall preference majority; and a verification advantage for the structured explanation in the qualitative data. Third, it clarifies the perception–evidence gap as a realization-sensitive complement to faithfulness and calibration, and derives design guidance for grounded, layered recommender explanations. To support scrutiny, the paper reproduces the matcher logic, all six scenarios, the prompt specification, the substantive questionnaire wording, and a complete paired example (Section 3.9).

2. Related Work

2.1. User-Centered Evaluation of Recommender Explanations

Recommender system explanations are intended to support several user-centered objectives beyond simply justifying why an item has been recommended. Prior research identifies transparency, scrutability, trust, persuasiveness, satisfaction, effectiveness, efficiency, and decision support as key dimensions of explanation quality, indicating that recommender explanations should be evaluated through users' perceptions and interactions rather than recommendation accuracy alone [4, 3, 16, 17]. Early work demonstrated that explanations can improve recommendation acceptance and perceived usefulness while also influencing user satisfaction [1, 18]. Importantly, Bilgic and Mooney [1] distinguish explanations designed to promote recommendations from those intended to improve user understanding and satisfaction, showing that persuasive explanations are not necessarily more transparent or informative.

Subsequent research has shown that explanation effectiveness depends on factors such as the recommendation domain, user characteristics, and explanation presentation. Feature-based and interactive explanation interfaces allow users to inspect the relationship between their preferences and recommended items, thereby improving transparency and user control [19, 20, 21]. Other studies emphasize that explanation framing and persuasive language can substantially influence users' attitudes and recommendation choices [22, 18, 23]. Recent work has also explored the use of LLMs to assist in evaluating recommendation explanations [24]. Collectively, these findings establish that explanation quality is inherently multidimensional and should be assessed through complementary user-centered measures such as understandability, trust, satisfaction, persuasiveness, and decision support. Moreover, trust should be viewed as appropriately calibrated to the available recommendation evidence rather than simply maximized [15].

2.2. LLM-Based Recommender Explanations

Natural-language explanations have traditionally been generated using templates, user preferences, item attributes, similarity relations, or manually designed rules. These approaches maintain a clear connection between recommendation evidence and explanation content but are often limited in linguistic diversity and contextual adaptability. The emergence of Large Language Models (LLMs) has considerably

expanded the capabilities of recommender systems by enabling fluent, personalized, and context-aware explanation generation. More broadly, LLMs have been explored for recommendation generation, conversational recommendation, preference elicitation, and reasoning [7, 25, 26].

Several recent studies have investigated LLMs specifically for explainable recommendation. XRec demonstrates that LLMs can produce richer and more readable recommendation justifications [8], while Silva et al. [9] and Nguyen [11] report favorable user responses, including improvements in user satisfaction and trust, for ChatGPT-generated recommendation explanations. Comparative studies further show that users generally perceive LLM-generated explanations as more natural and informative than conventional explanation methods [5]. Other approaches extend explanation generation by incorporating inferred reasoning, contextual knowledge, or consequence-oriented recommendations that communicate potential outcomes of accepting a recommendation [13, 27, 12]. These studies collectively demonstrate the potential of LLMs to improve the readability, expressiveness, and perceived usefulness of recommender explanations.

Despite these encouraging results, interpreting user evaluations remains challenging. In many existing systems, LLMs do not simply verbalize an existing recommendation rationale; they also select supporting evidence, introduce external knowledge, infer new relationships, or generate additional recommendation claims. Consequently, improved user evaluations may result from richer informational content, improved linguistic realization, or both. As highlighted by Lubos et al. [5] and Said [6], current evaluation settings differ considerably across studies, making it difficult to isolate the specific contribution of natural-language generation from changes in explanation content.

2.3. Explanation Realization, Grounding, and the Perception–Evidence Gap

Recent work has increasingly emphasized that explanation quality depends not only on what information is communicated but also on how it is communicated. Recommendation evidence provides the factual basis supporting a recommendation, explanation content represents the evidence-backed claims selected for presentation, and linguistic realization determines how that information is expressed through wording, structure, tone, and contextualization. Although these components are closely related, they are conceptually distinct. Two explanations may communicate the same recommendation evidence using different linguistic realizations, whereas LLMs may additionally enrich explanations with external knowledge, inferred reasoning, or consequence-based arguments [5, 13, 27, 12].

This distinction has motivated increasing interest in grounding and faithfulness within explainable recommendation. Recent approaches argue that generated explanations should faithfully reflect the underlying recommendation evidence rather than merely produce coherent or persuasive text. FIRE, for example, grounds generated explanations in model-derived evidence to improve explanatory faithfulness [28], while Tsai et al. [27] distinguish faithfulness from coherence and insightfulness when evaluating LLM-based recommendation reasoning. Related work further shows that fluent explanations may still lack evidential support [14, 29], and that credible recommender explanations should help users make informed decisions rather than simply increase persuasion [30]. Together, these studies suggest that user evaluations may reflect both the quality of the underlying recommendation evidence and the effectiveness of its realization and presentation. Separating these factors is therefore essential for interpreting the impact of LLM-realized recommender explanations.

3. Method

3.1. Study Design

We recruited 150 participants through Prolific [1] for a controlled within-subjects online study. Beyond the conceptual rationale given in Section 1, the camera domain also suited the practical requirements of a repeated within-subject online study: it is a familiar, moderate-involvement purchase category that Prolific participants could reasonably judge across six scenarios without specialized expertise or excessive fatigue, and it allowed System 1 to reproduce the Standard CB recommender of Naveed

[1] https://www.prolific.com/

and Ziegler [20] directly, so its matching mechanism did not need to be re-validated in an unfamiliar domain before serving as the fixed evidence source for both conditions. We revisit the resulting scope limitation in Section 5.3. For each of six camera-selection scenarios, participants inspected two outputs: **System 1**, the Structured Content-Based Recommender Explanation, and **System 2**, the LLM-Realized Recommender Explanation. The scenario, recommended camera, match score, item image, item features, and matched or unmet constraints were identical across the two outputs. Participants were not told how either output was generated. Both the scenario order and the display order of the two systems were randomized.

The scenarios were intentionally non-personalized. Every participant evaluated the same fixed recommendation cases so that variation in individual preference profiles could not change the evidence supplied to one explanation but not the other. This choice strengthens internal control but limits generalization to interactive, longitudinal, or personalized use. The explanation also differed in layout: System 1 used an explicit comparison grid and color-coded match indicators, while System 2 used a constraint summary and a prose panel. The causal contrast therefore combines linguistic realization, organization, and visual presentation under shared evidence.Several further design choices warrant explicit rationale. Each participant evaluated both explanation conditions for every scenario, rather than under a between-subjects allocation, to maximize within-subject statistical power at the target sample size and to allow direct paired comparison of the two conditions under identical evidence. Exactly two conditions were compared per scenario because the study's central manipulation is a binary contrast between the structured and LLM-realized presentation of the same evidence; extending the comparison to additional LLMs, prompts, or hybrid designs is left for future work (Section 5.3). All scenario constraints received equal weight in the match-score computation (Equation 2) because System 1 replicates the equal-weighting mechanism of our own earlier Standard Content-Based recommender [20]; introducing differential constraint weighting was outside the scope of this replication. No scenario in the final stimulus set required tie-breaking among cameras with identical maximum match scores.

This paired, evidence-controlled design lets us test the two hypotheses as stated below:

- **H1:** Under shared recommendation evidence, the LLM-Realized Recommender Explanation will receive more favorable ratings than the Structured Content-Based Recommender Explanation across the six user-centered constructs.
- **H2:** If participant judgments differ between the two conditions despite fixed underlying recommendation evidence, the pattern will be consistent with a perception-evidence gap associated with the combined realization-and-presentation manipulation.

3.2. System 1: Structured Content-Based Recommender Explanation

System 1 replicated the Standard Content-Based (CB) digital-camera recommender previously developed and evaluated by Naveed and Ziegler [20]. In that condition, users selected desired camera features and values; the system compared those requirements with each camera's attributes, ranked cameras by the resulting feature match, and explained each recommendation in a table showing the requested value, the camera value, and whether the feature matched. Here, the interactive preference-elicitation step of the Standard CB system was replaced by six fixed scenario descriptions. Each scenario therefore supplied the feature constraints, and the replicated content-based mechanism produced the recommended camera, percentage match score, and structured feature-constraint explanation shown in Figure 1.

The frozen catalog contained 60 cameras from nine brands and 16 interpretable attributes: nine binary attributes (electronic viewfinder, image stabilizer, waterproofing, WLAN, HDMI, HDR, image-sensor availability, manual focus, and touchscreen), five quantitative attributes (pixel count, weight, width, price, and zoom factor), and two categorical attributes (video resolution and storage medium). This catalog and its attribute schema were inherited unchanged from Naveed and Ziegler's earlier digital camera dataset [20].

For scenario s, let C_s denote its set of required feature constraints. For camera i, each constraint c produced a binary comparison result:

$$m_{ic} = \begin{cases} 1, & \text{if camera } i \text{ satisfies constraint } c, \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

For binary and categorical features, satisfaction required exact equality. For numerical features, satisfaction used the bound or interval stated in the scenario (e.g., price below €800, weight at most 350 g, or zoom at least 40×). Every scenario constraint had equal weight. The displayed percentage was then

$$\text{Match}(i, s) = 100 \times \frac{\sum_{c \in C_s} m_{ic}}{|C_s|}. \tag{2}$$

Thus, the score is simply the percentage of scenario constraints satisfied. For example, satisfying four of five constraints gives $100 \times 4/5 = 80\%$, as in Scenario 5. Cameras were ordered by this score, and the selected highest-scoring camera was then fixed as the recommendation for each scenario. A 100% score was labeled a complete match; when no camera reached 100%, the highest-scoring item was used as a near-match. System 1 displayed the fixed camera, the percentage score, each requested value, the camera's corresponding value, and a matched/unmatched indicator. This reproduces the published Standard CB explanation structure while making the computation explicit for the present fixed-scenario study.

3.3. System 2: LLM-Realized Recommender Explanation and Grounding Check

System 2 received the same frozen recommendation evidence used by System 1. The study materials identify the realization model as GPT-5.4 accessed through ChatGPT. The model was prompted with the scenario, user constraints, the prototype-recommended item, and its dataset features to generate a natural-language realization of the fixed recommendation. The model itself was not the object of evaluation; it was used only as a realization component, while the recommendation, supporting item-level evidence, and underlying decision process remained fixed across conditions. The prompt specification was as follows:

Prompt Specification

Input: scenario, scenario constraints, prototype-recommended item, and item features from the dataset. ***Task:*** generate one concise paragraph explaining why the fixed item fits the scenario. ***Restrictions:*** use only dataset attributes explicitly provided in the input; treat all constraints as soft constraints; do not change the item; do not mention alternatives; do not add external knowledge; do not infer unstated preferences; do not hallucinate missing specifications; and explicitly state any unmet or partially satisfied constraints in neutral language. ***Output:*** Product Name and Picture; Preference Match Score; Convincing and trustworthy explanation; Key constraints matching and non-matching.

The resulting System 2 interface combined fixed product name, picture, and score with the model-generated explanatory text and constraint summary. The instruction to produce a "Convincing and trustworthy" explanation may itself have affected rhetoric. We therefore treat the prompt wording as part of the tested System 2 condition rather than as a neutral implementation detail. Because persuasiveness and trust are precisely the constructs this instruction targets, any observed advantage for System 2 on these two constructs cannot be interpreted as a general property of LLM realization; it is at least partly a designed property of this specific prompt (Sections 4.2, 5.1).

Each final stimulus was manually reviewed for factual consistency, constraint alignment, and readability. To make that review reproducible, we operationalized the review as six checks: (1) the camera was unchanged; (2) every factual item claim mapped to a supplied item attribute; (3) no alternative item or unsupported item specification was introduced; (4) matched and unmet constraints agreed with System 1; (5) the match score was unchanged; and (6) unmet or partial constraints were stated without concealing the trade-off. Each final stimulus underwent the six-point manual grounding review described in Section 3.3, conducted by a single researcher. All stimuli deployed in the study satisfied the grounding criteria.

Table 1
The six fixed camera-selection scenarios. "Complete" means that at least one camera satisfied every stated constraint; "Near" means that the highest-matching available camera was used.

ID	Scenario	Required features	Match
S1	City trip and street photography	Budget under €800; weight $\leq$ 350 g; video resolution 3840 × 2160; WLAN = yes.	Complete
S2	Zoo/wildlife day trip on a budget	Budget under €500; zoom $\geq$ 30×; pixel count $\geq$ 20 MP; HDMI = yes.	Complete
S3	Enthusiast traveller seeking creative control	Image stabilizer = yes; image sensor = yes; manual focus = yes; highest video resolution 3840 × 2160.	Complete
S4	Gift for a travel vlogger	Price under €500; touchscreen = yes; WLAN = yes; HDMI = yes; storage medium = SDXC.	Near
S5	Lightweight safari/wildlife travel camera	Zoom $\geq$ 40×; image stabilizer = yes; electronic viewfinder = yes; WLAN = yes; manual focus = yes.	Near
S6	Family road-trip camera	WLAN = yes; HDR = yes; waterproof = yes; touchscreen = yes.	Near

Table 2
Dimension-by-dimension comparison of System 1 and System 2.

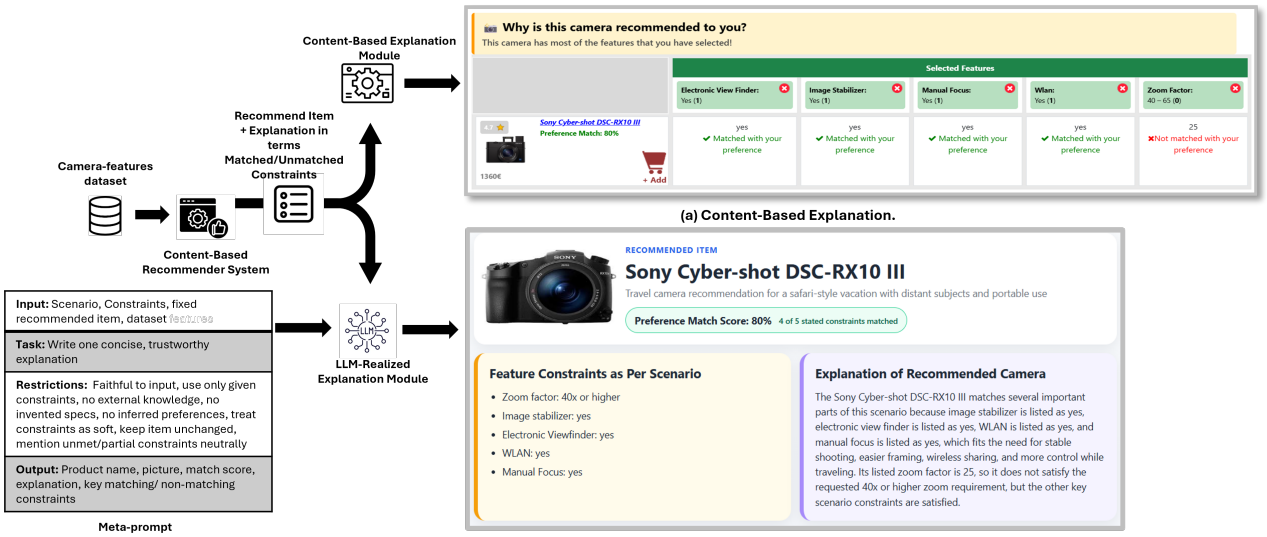
Dimension	System 1 (Structured CBRE)	System 2 (LLM-Realized Explanation)
Wording	Fixed template phrases and feature labels (e.g., "Matched with your preference")	Free-form natural-language prose generated per scenario (Section 3.3)
Amount of text	Short tabular labels and values only	One concise narrative paragraph plus a separate bulleted constraint list
Organization	Feature-by-feature comparison grid	Single narrative paragraph, separated from a structured constraint summary
Color coding	Green/red (check/cross) indicators per constraint	None; plain-text formatting only
Constraint visibility	Explicit matched/unmatched indicator for every requested feature	Constraints described narratively in prose, plus a separate constraint-summary list
Contextual interpretation	Not present; raw requested vs. actual values only	Present (e.g., generic feature-function interpretations such as "for stable shooting, easier framing"; Section 3.4)
Mismatch presentation	Marked with a red/unmatched indicator adjacent to the specific feature	Stated explicitly in neutral language within the narrative, per the prompt's restriction (Section 3.3)

3.4. Scenarios and a Complete Paired Example

Digital cameras were selected because their technical attributes are explicit enough to support evidence control and direct checking. This same property may favor a checklist over prose, so domain structure is treated as a moderator rather than a neutral choice. Table 1 reproduces all study scenarios. The six scenarios were constructed so that three cases had at least one camera satisfying every stated constraint (complete match) and three cases had no camera satisfying every constraint (near match; see the "Match" column of Table 1), so that both explanation conditions were evaluated both when communicating a fully satisfied recommendation and when communicating an unresolved trade-off; this ensures that condition preferences are not driven solely by the presence or absence of an unmet constraint.

Table 2 summarizes, at a glance, which presentational and content dimensions differ between the two conditions for every scenario, directly connecting Figure 1 to the three layers introduced in Section 1.

Figure 1 shows the workflow and the complete Scenario 5 pair. The fixed item was the Sony



(a) Content-Based Explanation.

(b) LLM-Realized Explanation.

Figure 1: Workflow for generating a) the Structured Content-Based Recommender Explanation and b) the LLM-Realized Recommender Explanation from the recommendation produced by the same content-based prototype.

Cyber-shot DSC-RX10 III. System 1 showed four matched constraints (image stabilizer, electronic viewfinder, WLAN, and manual focus) and one unmet constraint (25× zoom versus the requested 40×), yielding an 80% match. System 2 stated:

> "*The Sony Cyber-shot DSC-RX10 III matches several important parts of this scenario because image stabilizer is listed as yes, electronic view finder is listed as yes, WLAN is listed as yes, and manual focus is listed as yes, which fits the need for stable shooting, easier framing, wireless sharing, and more control while traveling. Its listed zoom factor is 25, so it does not satisfy the requested 40x or higher zoom requirement, but the other key scenario constraints are satisfied.*"

The phrases "stable shooting," "easier framing," and "more control" do not introduce new camera specifications, but they are generic feature-function interpretations that are not literally contained in the supplied attributes. Accordingly, the conditions shared item-level evidence and constraint status but were not strictly identical in all semantic content. This is another reason to interpret the manipulation as a comparison of complete explanation conditions rather than wording alone.

3.5. Procedure, Participants, and Ethics

For each randomized scenario, participants read the need description, inspected both fixed outputs, and completed comparative ratings. Participants completed the full set of comparative items (Table 3) once per scenario, immediately after viewing both explanations for that scenario, rather than once at the end of the study; construct scores were subsequently averaged first within construct and then across the six scenario-level administrations for each participant (Section 3.7). The display order of System 1 and System 2 within each scenario was randomized independently for every participant and every scenario (Section 3.1), which mitigates order-based bias in the comparative ratings. They also indicated an overall preferred explanation, selected which factors influenced that preference (explanation presentation, feature constraints, and/or recommended item), and provided open-ended comments. The open-ended items asked participants to describe their overall impression of the selected explanation and, separately, their reason for selecting it; these items imposed no minimum word or character count, so participants could respond in as much or as little detail as they chose. Recruitment applied no country or language restriction and no additional Prolific screening filters beyond the platform's standard eligibility requirements. Prolific reported an average reward of £10.93 per hour and a median completion time of 20 minutes 35 seconds for the study.

Table 3
Exact wording of the 13 substantive questionnaire items.

Construct	Comparative items
Understandability	(1) I found the explanation easier to understand. (2) The explanation was clearer and less ambiguous. (3) The explanation made the system's recommendation process transparent.
Trust	(4) The explanation increased my trust in the system's recommendation. (5) The explanation made me believe the system made the right recommendation.
Decision support	(6) The explanation helped me make a better decision about my system choice. (7) The explanation made me feel more confident in the system's recommended item.
Satisfaction	(8) I am more satisfied with the explanation. (9) The explanation met my expectations better.
Cognitive load	(10) It was more mentally demanding to understand the explanation. (12) I had to put more effort into making sense of the explanation.
Persuasiveness	(13) The explanation convinced me more about the system's recommended item. (14) The explanation influenced my opinion about the system's recommended item.

After applying the predefined data-quality checks, including the embedded attention check, eight records were excluded. The final analytic sample comprised 142 participants aged 19–72 years ($M = 32.27$, $SD = 9.94$). Participants provided informed consent, were compensated through Prolific at the rate reported above, and only anonymized responses were analyzed. The study was conducted in accordance with the ethical standards of the organization and the principles of the 1964 Helsinki Declaration and its subsequent amendments. Following a preliminary review, the organization's Ethics Council determined that no formal ethical approval was required for this study.

3.6. Measures and Operationalization

Thirteen substantive comparative items measured the six constructs – satisfaction, understandability, trust, persuasiveness, decision support, and cognitive load [8, 18] – and one additional item served as an attention-check (14 items in total). Table 3 reports the exact substantive wording. The attention-check item, numbered 11 in the survey, instructed participants to select a designated response and is omitted from the substantive table.

Responses used a five-point comparative scale from -2 (stronger preference for System 1) through 0 (about the same) to +2 (stronger preference for System 2). Cognitive-load items were reverse-coded so that positive values indicate lower perceived mental effort for System 2. Because item 6 used the phrase "system choice," participants may have interpreted it as choosing between explanation interfaces rather than evaluating the recommended camera; this wording is considered when interpreting the decision-support construct. We used a comparative rather than an absolute scale because our research questions concern relative preference between two explanations shown side by side for the same evidence, not the absolute quality of either explanation in isolation; a five-point comparative scale (-2 to +2) is a standard, easily interpreted format for paired comparative judgments and keeps the questionnaire short across six repeated scenarios. We acknowledge that a wider absolute scale (e.g., 1-7 per explanation) could provide finer-grained statistical power and would avoid any asymmetry-related bias toward positive or negative responses; we adopt this as a recommendation for future replications (Section 5.3).

3.7. Data Analysis

For each participant, item scores were averaged within construct and then across the six prespecified scenarios. Internal consistency was assessed using Cronbach's alpha. Two-sided one-sample Wilcoxon signed-rank tests compared each construct score with zero, where negative values favored System 1 and positive values favored System 2. Effect sizes are reported as r. Selections of explanation presentation, feature constraints, and the recommended item were compared using Cochran's Q and

pairwise McNemar tests. Chi-square tests examined whether these selected factors differed between participants preferring System 1 and those preferring System 2. Because six constructs were tested against zero within the same participant sample, we additionally applied a Holm–Bonferroni correction across the six tests reported in Table 4 to control the family-wise Type I error rate; we report both the uncorrected and the Holm-corrected results (Section 4.2).

3.8. Qualitative Analysis

Each record included comparative comments about the two explanation styles ("comparative comments" refers to the free-text field in which participants could note similarities or differences they noticed between the two explanations for a scenario, distinct from the separate "overall impression" and "reason for selection" fields), an overall impression, the reason for the final selected explanation, and the selected system (66 System 1; 76 System 2). Participants' rationales were grouped by the selected system. The other comparative comments were read to resolve ambiguous rationales and to select representative quotations, but they were not included in theme-prevalence counts. We developed an inductive, non-exclusive codebook, applied it to every rationale, and audited all assignments in a second pass. The codebook was developed inductively from an initial reading of the free-text rationales (open coding), then consolidated into the seven non-exclusive themes reported in Table 5 through iterative comparison and merging of overlapping codes, following standard practice for inductive thematic coding of short open-ended survey responses.Theme frequencies are descriptive rather than inferential.

3.9. Reproducibility and Data Availability

To support scrutiny and reuse, the following materials are available from the authors upon reasonable request: the frozen 60-camera catalog and its attribute schema (Section 3.2); the six scenario specifications (Table 1); the System 1 matcher implementation (Equations 1–2); the System 2 prompt specification and all six generated stimuli (Section 3.3); the grounding-check protocol and its outcomes; the exact questionnaire items (Table 3); the randomization procedure; the data-quality exclusion rules; and the analysis scripts underlying Sections 3.7–4.

4. Results

4.1. Reliability Check

Internal consistency was high for all constructs: understandability $\alpha = .91$, trust $\alpha = .90$, decision support $\alpha = .90$, satisfaction $\alpha = .90$, cognitive load $\alpha = .91$, and persuasiveness $\alpha = .88$. These values support participant-level construct aggregation for the fixed scenario set.

4.2. Results with Respect to H1

H1 proposed that, under shared recommendation evidence, the LLM-Realized Recommender Explanation would receive more favorable ratings than the Structured Content-Based Recommender Explanation. The results provide partial support for H1. As shown in Table 4, all effects were small. System 2 was rated more favorably for persuasiveness ($M = 0.23$, $p = .006$, $r = .24$), trust ($M = 0.20$, $p = .016$, $r = .21$), reverse-coded cognitive load ($M = 0.20$, $p = .019$, $r = .21$), and decision support ($M = 0.19$, $p = .031$, $r = .19$); all four differences were statistically significant at $\alpha = .05$. Satisfaction did not differ ($M = 0.10$, $p = .264$, $r = .10$). Understandability was the only construct with a negative mean ($M = -0.17$, $p = .063$, $r = .16$), indicating a small, nonsignificant directional advantage for System 1. Taken together, System 2 showed small advantages across four of the six constructs, but the pattern was not uniform across all outcomes.

Because six tests were conducted on the same participant sample, we applied a Holm–Bonferroni correction across these six comparisons to control the family-wise Type I error rate (Section 3.7). Ordering the uncorrected p-values from smallest to largest (.006, .016, .019, .031, .063, .264) and comparing

Table 4
Structured Content-Based Recommender Explanation (System 1) vs. LLM-Realized Recommender Explanation (System 2) (N = 142)

Construct	M	SD	Z	p	r	Sig. after Holm correction
Persuasiveness	0.23**	0.94	2.76	.006	0.24	Yes
Cognitive Load (rev.)	0.20*	1.10	2.35	.019	0.21	No
Trust	0.20*	1.02	2.40	.016	0.21	No
Decision Support	0.19*	1.02	2.15	.031	0.19	No
Satisfaction	0.10	1.11	1.12	.264	0.10	No
Understandability	-0.17†	1.11	-1.86	.063	0.16	No

Note. Scores ranged from −2 (preference for System 1) to +2 (preference for System 2). Values are participant-level means aggregated across the six scenarios. Cognitive load was reverse-coded such that higher values indicate lower perceived mental effort for System 2. Two-sided one-sample Wilcoxon signed-rank tests against zero were conducted. $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$.

each to its step-down threshold (α/6 = .0083, α/5 = .01, α/4 = .0125, α/3 = .0167, α/2 = .025, α/1 = .05), only the persuasiveness effect (p = .006 < .0083) survives correction. The next-smallest p-value, .016 for trust, already exceeds its threshold of .01, so the Holm procedure stops there: the trust, cognitive-load, and decision-support effects (uncorrected p = .016–.031) no longer meet the corrected significance threshold and should be interpreted as descriptive and hypothesis-generating rather than confirmatory. This substantially narrows the statistical support for H1 to a single robust, construct-specific effect — a persuasiveness advantage for System 2 — and directly addresses the multiple-comparisons concern raised in review. Table 4 reports both the uncorrected and Holm-corrected results for full transparency.

The qualitative responses help clarify this pattern. Participants who favored the LLM-Realized Recommender Explanation often described it as *more detailed*, *convincing*, *informative*, and *trustworthy*, and some explicitly linked it to greater confidence in the recommendation. Others described it as more human-like; for example, one participant wrote that *"it feels like a real recommendation from a person"*. At the same time, the Structured Content-Based Recommender Explanation retained a practical advantage for fast inspection and verification, being described as *clearer*, *shorter*, and *easier to scan*. This was especially evident in comments referring to the explicit checklist and red/green cues, such as *"you can quickly see which requirements are met."* Overall, the qualitative data mirror the quantitative pattern: System 2 was associated more with contextual interpretation and persuasion, whereas System 1 was associated with rapid verification and immediate feature-to-constraint checking.

4.3. Results with Respect to H2

H2 concerned whether participant judgments differed across explanation conditions despite fixed underlying recommendation evidence. Seventy-six participants (53.5%) selected System 2 and 66 (46.5%) selected System 1. The exact binomial test was not significant ($p = .450$), so the overall preference split did not support a majority advantage for either system. Evidence relevant to H2 instead comes from the construct-specific ratings, factor-selection results, and qualitative rationales.

Participants most frequently identified explanation presentation as important ($n = 106$, 74.6%), followed by feature constraints ($n = 86$, 60.6%), whereas the recommended item was selected least often ($n = 44$, 31.0%). A Cochran's Q test indicated that these selection frequencies differed significantly, $Q(2) = 50.92$, $p < .001$. Pairwise McNemar tests showed that explanation presentation was selected more often than feature constraints ($p = .042$) and more often than the recommended item ($p < .001$); feature constraints were also selected more often than the recommended item ($p < .001$).

Exploratory group comparisons further indicated that participants preferring System 2 selected explanation presentation as important more often than those preferring System 1 (92.1% vs. 54.5%), $\chi^2(1, N = 142) = 26.33$, $p < .001$. Conversely, participants preferring System 1 selected feature constraints more often than those preferring System 2 (69.7% vs. 52.6%), $\chi^2(1, N = 142) = 4.31$, $p = .038$. Because these group comparisons were secondary and exploratory, the latter result should be interpreted

cautiously.

4.4. Qualitative Evidence for H1 and H2

Among the 76 participants who preferred System 2, the most prevalent themes were detail or informativeness (29/76, 38.2%), clarity or accessibility (29/76, 38.2%), contextual reasoning (23/76, 30.3%) (Table 5). One participant valued that System 2 explained "why those features matter for my specific scenario" and described its handling of unmet constraints as something that "built much more trust." Another summarized the benefit as helping users understand "how each feature relates to the user's intended use." These responses support the natural-language and contextualization mechanism proposed in H1.

Among the 66 participants who preferred System 1, the dominant themes were brevity or low effort (45/66, 68.2%), clarity or accessibility (35/66, 53.0%), and direct feature verification (22/66, 33.3%). One participant preferred the "clear, side-by-side comparison" because it allowed technical details to be checked "instantly without having to read through long paragraphs." Another noted that green check marks and red X marks made critical mismatches "impossible to miss." Thus, the qualitative evidence also explains why H1 was only partially supported: System 2 facilitated contextual interpretation and confidence, whereas System 1 reduced search effort and supported rapid verification.

Table 5
Themes in reasons for selecting the final explanation condition. Codes are non-exclusive; percentages use the corresponding preference group as the denominator. Counts refer to the final analytic sample ($N = 142$).

Theme	System 1 ($n = 66$)	System 2 ($n = 76$)
Brevity / low effort	45 (68.2%)	3 (3.9%)
Clarity / accessibility	35 (53.0%)	29 (38.2%)
Direct feature verification	22 (33.3%)	4 (5.3%)
Detail / informativeness	1 (1.5%)	29 (38.2%)
Contextual reasoning	0 (0.0%)	23 (30.3%)
Trust / persuasion / personalization	1 (1.5%)	19 (25.0%)
Presentation / visual appeal	16 (24.2%)	21 (27.6%)

A useful qualitative distinction is that both groups invoked clarity but meant different things. For System 1, clarity generally meant low search cost, visible constraint status, and quick comparison. For System 2, clarity generally meant understanding why a specification mattered in the scenario. The same evidence could therefore be experienced as clear through either compression or contextual elaboration.

The qualitative responses also converge with H2. Presentation, organization, or visual appeal was explicitly cited in 16/66 (24.2%) System 1 rationales and 21/76 (27.6%) System 2 rationales, indicating that the user-facing form mattered in both groups even though it supported different goals. One participant observed, "They are pretty much the same, just the layout changes," explicitly recognizing a shared informational basis while reacting to its presentation. Another summarized the functional trade-off: "System 1 helps users make quick and accurate decisions, while System 2 helps users understand how each feature relates to the user's intended use." Even some System 2-preferring participants requested clearer mismatch signaling; one wrote that it was visually preferable but "needs improvement to indicate when something does not match requirements."

Taken together, the construct-specific quantitative differences, factor-selection results, and qualitative rationales provide qualified support for H2: judgments differed across the complete user-facing explanation conditions under shared item-level evidence. This support is bounded by the small, non-uniform effects and by the fact that realization, layout, and generic contextual interpretation changed together. It does not depend on a statistically significant majority preference for System 2.

5. Discussion

5.1. H1: Partial Support for a Modest Perceptual Advantage of LLM Realization

After Holm–Bonferroni correction for the six simultaneous tests (Sections 3.7, 4.2), only the persuasiveness effect remains statistically significant; the trust, cognitive-load, and decision-support differences should be read as suggestive rather than confirmatory, which narrows this claim considerably relative to the uncorrected pattern. With recommendation evidence held constant, the LLM-Realized Recommender Explanations showed at most a small perceptual advantage, robustly for*persuasiveness* and, on an exploratory basis, for *trust*, with a parallel reduction in perceived cognitive load. Moreover, because the System 2 prompt explicitly instructed the model to produce a "convincing and trustworthy" explanation (Section 3.3), the persuasiveness and trust results are at least partly a designed property of this specific prompt rather than a general property of LLM-realized explanations; we caution against generalizing these effect sizes to LLM realizations produced with different, less persuasion-oriented prompts. This pattern is consistent with prior work showing that LLM-generated recommender explanations are often experienced as more natural, accessible, or trust-supporting [9, 5, 6]. Satisfaction did not differ, and understandability showed a nonsignificant directional tendency in favor of the Structured Content-Based Recommender Explanation. The findings therefore indicate a modest, construct-specific, and only partially robust, advantage rather than a uniform superiority of LLM realization.

The qualitative findings strengthen this interpretation without changing the inferential conclusion. Respondents who selected System 2 repeatedly cited the mechanisms specified in H1: contextual explanation, richer detail, scenario relevance, and confidence. Respondents who selected System 1 instead valued brevity, side-by-side comparison, and visible match status. These themes explain the quantitative divergence between persuasiveness and understandability. More language was useful when users needed interpretation, but less language and stronger visual structure were useful when they needed verification. Explanation quality is therefore not a single continuum from "less" to "more" detail.

5.2. H2: Qualified Evidence for a Perception–Evidence Gap

The nonsignificant 53.5% versus 46.5% preference split does not establish a majority preference for either system. The quantitative evidence relevant to H2 instead lies in the construct-specific differences—most pronounced for persuasiveness, the one effect that survives correction for multiple comparisons (Section 4.2)—together with the reported importance of presentation and the distinct factor-selection patterns of the two preference groups. The qualitative comments further attribute differences to layout, contextualization, and visibility.

Because the scenario, recommended camera, match score, feature values, and matched or unmet constraints were held constant, these converging results provide qualified support for H2 and are consistent with a perception-evidence gap: recent work argues that recommender explanations can be persuasive without being equally credible or faithful to the underlying rationale [30, 6], and persuasive-explanation research shows that wording and framing can shape judgments independently of utility [23, 22].

The support is strongest for persuasiveness and for reported presentation-related drivers, and weaker for constructs such as satisfaction and understandability. The gap should therefore be interpreted as modest and non-uniform, not as a general proof that LLM explanations are better. This distinction matters concretely for trust calibration in our own data: System 2 received a significant rating advantage for persuasiveness (Table 4) despite communicating exactly the same match score and constraint status as System 1 (Sections 3.2, 3.3). Recall that recommendation evidence denotes the scenario, item, match score, and constraint outcomes (Section 1), and that the LLM-realized package denotes the complete System 2 output — prose realization plus this same underlying evidence (Section 3.3). Because trust and persuasion increased without any corresponding change in evidence, our own results directly illustrate the risk we flag here: designers who observe higher trust or persuasiveness ratings for an LLM-realized

explanation should not treat that increase as confirmation that the underlying recommendation is better supported, since in our data the increase is attributable to realization rather than to evidence. At the same time, our qualitative results (Section 4.4, Table 5) show that participants who preferred System 2 valued it for contextual reasoning and clarity about why specific features mattered, not for unsupported claims; this suggests that a perception-evidence gap is not itself evidence of manipulation or unfaithfulness, and that contextual language can legitimately aid interpretation of fixed evidence. The resulting design implication, grounded in these two observations rather than asserted independently of them, is that layered interfaces should let users inspect the underlying constraint-level evidence (as in System 1) while optionally offering grounded natural-language realization (as in System 2) for users who want contextual interpretation — so that fluent rhetoric adds accessibility without becoming the sole basis for trust. The theoretical claim should remain proportional to the data: we treat the perception-evidence gap as a descriptive lens and a controlled comparison strategy, not as a large universal effect or a validated standalone scale.

5.3. Presentation, Content, Measurement, and Domain Limitations

The experiment controls item-level recommendation evidence but not all aspects of the user-facing explanation. System 1 combines a grid, color coded indicators, and terse feature values; System 2 combines a constraint list with prose. The qualitative data confirm that participants noticed the complete interface: visual organization was cited in both preference groups, and one response described the systems as "pretty much the same, just the layout changes." The directional understandability advantage for System 1 may therefore reflect the checklist, color cues, linguistic brevity, or their combination. Future factorial work should cross language form (structured vs. prose), layout (grid vs. paragraph), and visual cues.

The example also shows that System 2 added generic feature-function interpretations, despite the prompt's restriction against external knowledge and inferred preferences. These phrases did not add camera specifications, but they may have increased contextual content. Future studies should either prohibit such interpretations through stricter generation and auditing or manipulate them as a separate factor. In addition, the phrase "convincing and trustworthy" in the prompt may have contributed to the persuasiveness result.

Digital cameras offer explicit, enumerable attributes. This makes one-to-one constraint checking unusually effective and may increase the verification advantage of the Structured Content-Based Recommender Explanation. In domains with subjective attributes, latent preferences, or experiential qualities—such as books, music, restaurants, or education—natural language may add more useful interpretation, but grounding becomes harder. Domain attribute structure is therefore a plausible moderator. The same reasoning extends to high-stakes domains such as financial services or housing, where the cost of a poor recommendation is higher and users may weigh persuasive language more cautiously; the checklist-versus-prose trade-off documented here may shift considerably in such settings, and we do not extend our conclusions to them. Two further limitations deserve explicit emphasis, raised directly by the reviews. First, the six scenarios were fixed rather than randomly sampled from a larger population of possible scenarios, and scenario was not modeled as a random effect in the analysis (Section 3.7); the reported effects may therefore partly reflect the specific wording of these six LLM-generated stimuli rather than LLM realization in general. A mixed-effects model with scenario as a random factor, applied to a larger and randomized scenario set, would provide a stronger basis for generalization. Second, both the qualitative coding (Section 3.8) and the grounding review of the System 2 stimuli (Section 3.3) were each conducted by a single researcher without a second coder or a reported inter-rater reliability statistic, which may introduce coder-specific bias; independent double-coding in future replications would strengthen these checks.

Other limitations include the non-personalized scenario design, reliance on self-reported perceptions rather than observed choice behavior, the −2 to +2 comparative rating scale rather than a wider absolute scale (Section 3.6), use of one LLM realization setup and one prompt, and single-coder qualitative analysis. The decision-support item referring to "system choice" may also have been interpreted as

a choice between explanation interfaces. Finally, future reporting should record the exact GPT-5.4 interface variant and access date, as well as all prespecified exclusion criteria, to improve reproducibility (Sections 3.3, 3.9).

6. Conclusion and Outlook

This paper presented a controlled within-subjects comparison of the Structured Content-Based Recommender Explanation and the LLM-Realized Recommender Explanation under shared recommendation evidence, within a single, attribute-explicit product domain (digital cameras) and a single LLM configuration; we do not claim these findings generalize to recommendation domains with more subjective or experiential attributes, or to other models and prompts (Section 5.3). Under uncorrected tests, System 2 showed small advantages for persuasiveness, trust, lower perceived cognitive load, and decision support; after Holm–Bonferroni correction for multiple comparisons, only the persuasiveness advantage remained statistically significant, and we report the others as descriptive rather than confirmatory (Sections 3.7, 4.2). Satisfaction did not differ, while System 1 showed a nonsignificant directional understandability advantage and was preferred qualitatively for verification-oriented use. The overall preference split did not differ significantly from an even split. More broadly, the results are consistent with a modest perception–evidence gap: user judgments varied across complete explanation conditions despite fixed item-level evidence. This interpretation is bounded by the combined changes in wording, layout, and generic contextual interpretation, as well as by the single product domain, Prolific-based scenario evaluation, reliance on self-report, the fixed six-scenario stimulus set analyzed without scenario as a random effect, single-coder qualitative and grounding review, and use of one LLM setup. Future recommender interfaces should therefore move beyond one-size-fits-all explanation design and adopt layered, evidence- and context-grounded approaches that preserve inspectable constraints while optionally providing natural-language interpretation for users who need it.

Declaration on Generative AI During the preparation of this revised manuscript, the authors used ChatGPT (OpenAI) for grammar and spelling checking, consistency review, and limited paraphrasing. The authors reviewed and edited the suggested changes and take full responsibility for the manuscript's content.

References

[1] M. Bilgic, R. J. Mooney, Explaining recommendations: Satisfaction vs. promotion, in: Beyond personalization workshop, IUI, volume 5, 2005, p. 153.

[2] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: Proceedings of the fifth ACM conference on Recommender systems, 2011, pp. 157–164.

[3] N. Tintarev, J. Masthoff, Effective explanations of recommendations: user-centered design, in: Proceedings of the 2007 ACM conference on Recommender systems, 2007, pp. 153–156.

[4] N. Tintarev, Explanations of recommendations, in: Proceedings of the 2007 ACM conference on Recommender systems, 2007, pp. 203–206.

[5] S. Lubos, T. N. T. Tran, A. Felfernig, S. Polat Erdeniz, V.-M. Le, Llm-generated explanations for recommender systems, in: Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, 2024, pp. 276–285.

[6] A. Said, On explaining recommendations with large language models: a review, Frontiers in big Data 7 (2025) 1505284.

[7] L. Li, Y. Zhang, D. Liu, L. Chen, Large language models for generative recommendation: A survey and visionary discussions, in: Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024), 2024, pp. 10146–10159.

[8] Q. Ma, X. Ren, C. Huang, Xrec: Large language models for explainable recommendation, in: Findings of the Association for Computational Linguistics: EMNLP 2024, 2024, pp. 391–402.
[9] Í. Silva, L. Marinho, A. Said, M. C. Willemsen, Leveraging chatgpt for automated human-centered explanations in recommender systems, in: Proceedings of the 29th International Conference on Intelligent User Interfaces, 2024, pp. 597–608.
[10] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, et al., A survey on large language models for recommendation, World Wide Web 27 (2024) 60.
[11] L. V. Nguyen, User satisfaction and trust on recommendation systems with llms-generated explanations, in: International Conference on Networks, Communications and Intelligent Computing, Springer, 2024, pp. 1177–1188.
[12] M. Bismay, X. Dong, J. Caverlee, Reasoningrec: Bridging personalized recommendations and human-interpretable explanations through llm reasoning, in: Findings of the Association for Computational Linguistics: NAACL 2025, 2025, pp. 8132–8148.
[13] S. Lubos, M. Gartner, A. Felfernig, R. Willfort, Leveraging llms to explain the consequences of recommendations, in: Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, 2025, pp. 318–322.
[14] A. Madsen, S. Chandar, S. Reddy, Are self-explanations from large language models faithful?, 2024, URL https://arxiv. org/abs/2401.07927 (2024).
[15] M. Wischnewski, N. Krämer, E. Müller, Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions, in: Proceedings of the 2023 CHI conference on human factors in computing systems, 2023, pp. 1–16.
[16] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).
[17] S. Naveed, G. Stevens, D. Robin-Kern, An overview of the empirical evaluation of explainable ai (xai): A comprehensive guideline for user-centered evaluation in xai, Applied Sciences 14 (2024) 11288.
[18] S. Gkika, G. Lekakos, The persuasive role of explanations in recommender systems., in: Bcss@ Persuasive, 2014, pp. 59–68.
[19] M. Millecamp, S. Naveed, K. Verbert, J. Ziegler, To explain or not to explain: the effects of personal characteristics when explaining feature-based recommendations in different domains, in: IntRS@RecSys, 2019. URL: https://api.semanticscholar.org/CorpusID:203415984.
[20] S. Naveed, J. Ziegler, Featuristic: An interactive hybrid system for generating explainable recommendations-beyond system accuracy., in: IntRS@ RecSys, 2020, pp. 14–25.
[21] S. Naveed, An Interactive Hybrid Approach to Generate Explainable and Controllable Recommendations, Ph.D. thesis, Duisburg, Essen, 2021, 2021.
[22] H. Alizadeh Noughabi, B. Behkamal, F. Zarrinkalam, M. Kahani, Persuasive explanations for path reasoning recommendations, Journal of Intelligent Information Systems 63 (2025) 413–439.
[23] S. T. Rahman, D. Siemon, T. Ruotsalo, Persuasive explanations for recommender systems: how explanations can influence users' choices?, International Journal of Human-Computer Studies (2025) 103720.
[24] X. Zhang, Y. Li, J. Wang, B. Sun, W. Ma, P. Sun, M. Zhang, Large language models as evaluators for recommendation explanations, in: Proceedings of the 18th ACM Conference on Recommender Systems, 2024, pp. 33–42.
[25] S. Sanner, K. Balog, F. Radlinski, B. Wedin, L. Dixon, Large language models are competitive near cold-start recommenders for language-and item-based preferences, in: Proceedings of the 17th ACM conference on recommender systems, 2023, pp. 890–896.
[26] H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, R. Chen, C. Leung, J. Tang, J. Luo, Llm-rec: Personalized recommendation via prompting large language models, in: Findings of the Association for Computational Linguistics: NAACL 2024, 2024, pp. 583–612.
[27] A. Tsai, A. Kraft, L. Jin, C. Cai, A. Hosseini, T. Xu, Z. Zhang, L. Hong, E. H. Chi, X. Yi, Leveraging llm reasoning enhances personalized recommender systems, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 13176–13188.

[28] S. Sani, A. Meskin, M. Amanlou, H. R. Rabiee, Fire: Faithful interpretable recommendation explanations, arXiv preprint arXiv:2508.05225 (2025).

[29] B. Kabongo, V. Guigue, On the factual consistency of text-based explainable recommendation models, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2026, pp. 16–28.

[30] P. Qin, C. Huang, Y. Deng, W. Lei, T.-S. Chua, Beyond persuasion: Towards conversational recommender system with credible explanations, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2024, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 4264–4282. URL: https://aclanthology.org/2024.findings-emnlp.247/. doi:10.18653/v1/2024.findings-emnlp.247.