

A Survey on Natural Language User Profiles for Recommendation: Methods, Datasets, and Metrics

Mariam Arustashvili Krisztian Balog

University of Stavanger, Norway

{mariam.arustashvili,krisztian.balog}@uis.no

Abstract

Natural language (NL) user interest profiles—concise, human-readable descriptions of a user’s preferences—are an emerging alternative to opaque embeddings in recommender systems, offering transparency and user control alongside predictive utility. Recent advances in large language models have made generating and consuming such profiles increasingly practical, spurring a growing but fragmented body of research. In this survey, we provide the first systematic review of NL user profiles for recommendation, organizing the literature around two coupled tasks—profile generation and profile-based recommendation—and surveying the methods, datasets, and evaluation practices currently in use. We further propose the Reference-Anchored Evaluation Taxonomy (RAET), which organizes intrinsic and extrinsic metrics around explicit (output, reference signal) pairs, and reveals where efforts have concentrated and which axes remain unexplored. Finally, we identify structural open challenges—such as the tension between transparency and task-driven optimization—establishing a clear roadmap for future research in this space.

1 Introduction

The advent of large language models (LLMs) has driven a surge of work at the intersection of NLP and recommender systems, with a prominent strand representing users through natural language. Rather than relying solely on opaque, dense embeddings or sparse interaction matrices, researchers are increasingly exploring the use of natural language to capture, reason about, and act upon user preferences. This shift reflects a broader movement toward user-centric recommender systems, where transparency, interpretability, and user agency are increasingly recognized as first-class design goals alongside predictive accuracy (Ekstrand and Willemsen, 2016; Ekstrand et al., 2025; Wu et al., 2026; Vineis and Tolomei, 2026).

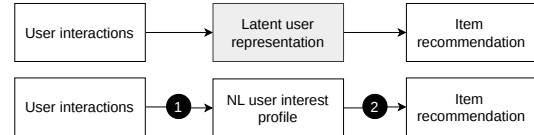


Figure 1: Comparison of traditional latent user representation (top) and the NL profile-based architecture (bottom). The proposed pipeline explicitly decouples into two tasks: (1) NL profile generation and (2) NL profile-based recommendation.

As illustrated in Figure 1, this paradigm fundamentally splits the recommendation process into two coupled tasks: (1) generating the NL profile and (2) recommending items based on that profile. To date, the vast majority of literature has addressed these tasks independently. Closing the loop to address them jointly remains one of the field’s most pressing open questions, which in turn introduces significant complexities in measurement and metrics—a challenge our proposed evaluation taxonomy aims to resolve.

As this research area rapidly expands across the recommender systems, NLP, and information retrieval communities, the terminology has begun to fracture and findings remain siloed. Most notably, the distinction between traditional *text-based user profiles* (which often consist of raw, concatenated textual metadata such as past reviews or search logs) and *natural language-based* (NL-based) user interest profiles (which are coherent, synthesized textual descriptions of a user’s preferences) remains ambiguous in the literature. Expressing preferences as free-form text also reopens a decades-long debate on the boundary between search and recommendation (Belkin and Croft, 1992; Li et al., 2024): a sufficiently detailed NL profile functions much like an expressive search query.

To provide clarity and structure to this emerging subfield, we define an *NL-based user interest profile* (*NL user profile* for short) as a fluent, natural language description of a user’s interests. This def-

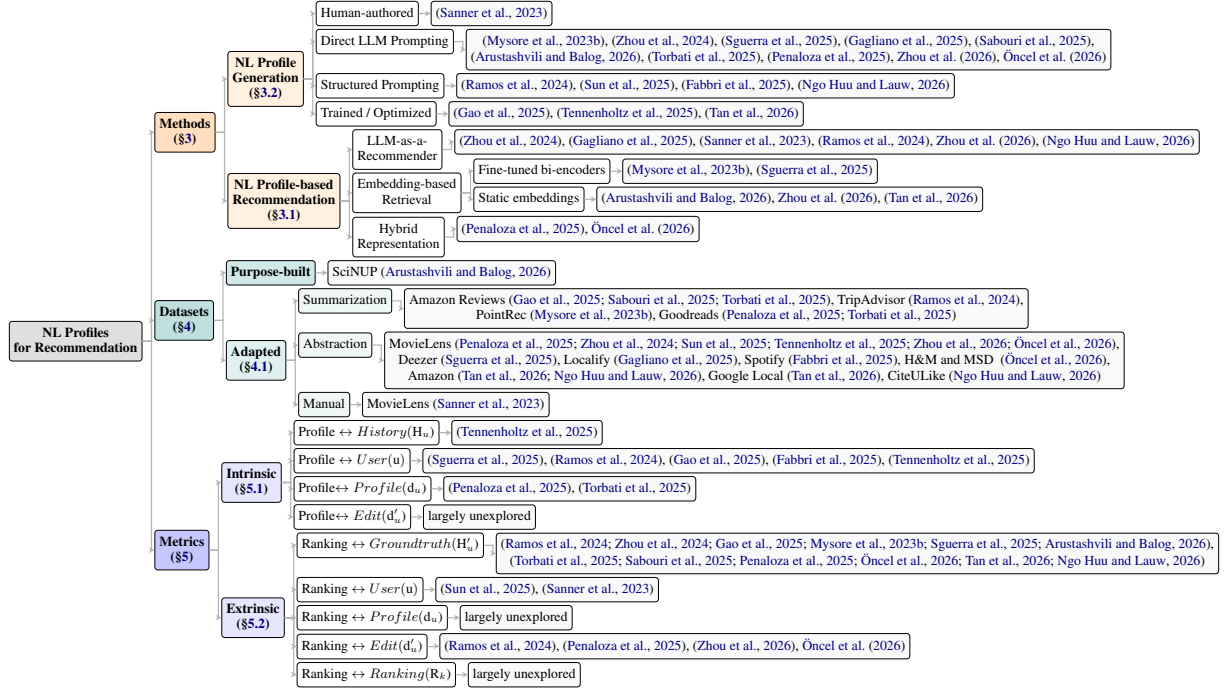


Figure 2: Taxonomy of the NL profile-based recommendation literature. The field is organized across three primary pillars: *Methods* (Section 3), *Datasets* (Section 4), and *Metrics* (Section 5). Within *Methods*, approaches are split by the two core pipeline tasks (generation and recommendation). *Datasets* are categorized by adaptation strategy, with SciNUP separated as the sole purpose-built dataset. The *Metrics* branch illustrates the Reference-Anchored Evaluation Taxonomy (RAET), mapping intrinsic and extrinsic evaluation axes by their reference signals. Leaf nodes indicate the specific papers populating each category.

initiation encompasses both general (*standing*) preferences and short-term (*situational*) interests, which naturally dissipate once a specific information need is satisfied. In this survey, we focus specifically on the task of providing recommendations based on such NL user profiles, framing it fundamentally as a problem of either rating prediction or top- k ranking from an item catalog.

By reviewing the literature (with methodology detailed in Appendix A), we find that (1) the vast majority of existing work focuses on the downstream recommendation problem, while profile generation and refinement receive less attention; (2) specialized datasets tailored to NL profiles are scarce; and (3) evaluation is fragmented—extrinsic ranking accuracy dominates, while user-facing properties such as transparency and controllability are measured only in a handful of studies, using study-specific evaluation protocols that hinder cross-study comparison. Because this rapid, multi-disciplinary expansion has reached a critical inflection point, synthesizing this fragmented landscape is now crucial to establish shared foundations and enable cumulative progress.

To address these gaps and synthesize the current state of the field, our main contributions are:

- (1) **Formalization and Scoping:** We define the emerging NL user profile paradigm, distinguishing human-readable, synthesized preference narratives from opaque embeddings and raw metadata.
- (2) **Literature Synthesis:** We systematically categorize the fragmented body of existing work across two coupled tasks—profile generation and profile-based recommendation—highlighting the heavy reliance on repurposed data over specialized benchmarks.
- (3) **Evaluation Taxonomy:** We propose the Reference-Anchored Evaluation Taxonomy (RAET), pairing intrinsic and extrinsic metrics with explicit reference signals, and expose the critical gap in measuring user-facing benefits like interpretability and control.
- (4) **Open challenges:** We identify the structural bottlenecks and methodological tensions currently limiting the field, providing a synthesized roadmap for future research.

Figure 2 provides a visual summary of the literature taxonomy explored throughout this survey.

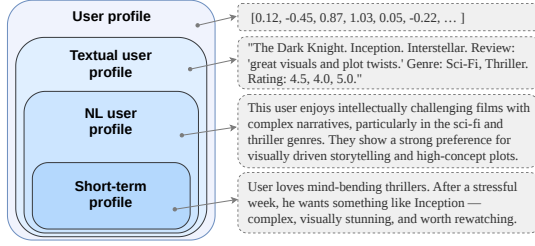


Figure 3: Forms of user profiles in recommender systems. A user profile can take the form of a latent representation or a textual profile; textual profiles range from raw or lightly structured text to synthesized, human-readable NL profiles. Short-term profiles are a situational special case of NL profiles.

2 Natural Language User Profiles for Recommendation

We identify two main tasks in the NL profile-based recommendation process: profile generation and profile-based recommendation. Before defining these tasks, we first clarify what constitutes an NL user interest profile and how it differs from related text-based representations.

2.1 Natural Language User Interest Profile

Definition 2.1 (NL user interest profile). We define the *NL user interest profile* (or *NL profile* for short) of a user u as a short document d_u that provides a user-facing summary of their preferences and interests. Expressed in well-formed natural language, d_u consists of coherent sentences or paragraphs a human can readily interpret, inspect, and edit.

This definition aligns with how NL profiles are implicitly treated in prior work (Balog et al., 2019; Radlinski et al., 2022; Sanner et al., 2023; Arustashvili and Balog, 2026), albeit none of those studies provide a formal definition. NL profiles are also referred to in the literature as user summaries (Penaloza et al., 2025), scrutable summaries, descriptive history representations (Tennenholtz et al., 2025), and narratives (Mysore et al., 2023b).

To clarify our scope, we explicitly exclude text-based representations that lack narrative coherence (as illustrated in Figure 3). These include keyword sets (Torbaty et al., 2025), ranked tag/topic lists (Mysore et al., 2023a), and raw review aggregations (Salemi et al., 2024). While useful as inputs for profile generation, they do not offer the human readability that characterizes a true NL profile. Similarly outside our scope are memory-augmented and agentic recommenders that maintain textual user memories as internal system com-

ponents consumed by downstream LLMs rather than as user-facing, editable artifacts (Chen et al., 2026; Shi et al., 2025).

Within these boundaries, a primary distinction among NL profiles is their temporal scope, capturing a spectrum from *standing* preferences to *situational* intents. Standing preferences reflect stable, long-term interests that persist across sessions (Radlinski et al., 2022; Sanner et al., 2023). In contrast, situational intents reflect immediate goals or context-bound needs—such as in narrative-driven recommendation (Mysore et al., 2023b)—which naturally dissipate once the specific information need is satisfied.

Ultimately, formalizing these profiles serves a dual purpose. From a system perspective, they serve as rich user representations, replacing or augmenting traditional collaborative filtering signals. From a user perspective, they provide *transparency* and *agency*, functioning as a legible, reflective tool that users can inspect and correct. Such profiles need not be purely descriptive: they may also capture aspirational interests that a user’s interaction history has yet to reflect. Notably, Sun et al. (2025) observe that a profile subtly misaligned with a user’s actual interests can actually be beneficial: the friction increases user engagement by inviting them to actively participate in refining their profile.

2.2 NL Profile Generation

Definition 2.2 (NL profile generation). Given a user u and their interaction history H_u , *NL profile generation* is the task of producing a NL user interest profile d_u that summarizes u ’s preferences. Formally, this can be expressed as a generation function $d_u = \mathcal{G}(H_u)$.

While d_u is typically generated automatically by a language model acting as $\mathcal{G}$, the profile may also be authored directly by the user (Sanner et al., 2023) or co-authored collaboratively by the system and the user (Sun et al., 2025).

Because user interests naturally evolve over time, an essential extension of this task is *profile updating*. When a user engages in new interactions, denoted as ΔH_u , the system must adapt the profile to reflect these recent signals. This updating process can generally be approached through three distinct paradigms:

- **Regeneration (stateless):** The system completely regenerates the profile from scratch by applying the generation function over the

newly expanded history, producing an updated profile $d'_u = \mathcal{G}(H_u \cup \Delta H_u)$.

- **Incremental Update (stateful):** To reduce computational overhead and maintain continuity in the text, the system takes the existing profile d_u and the recent interactions ΔH_u as inputs to an update function $\mathcal{U}$, modifying only the necessary parts of the profile: $d'_u = \mathcal{U}(d_u, \Delta H_u)$.
- **Manual Revision (user-driven):** The user directly inspects and edits the text of d_u , bypassing automated generation to explicitly correct the system’s understanding or to inject new, aspirational preferences.

2.3 NL Profile-Based Recommendation

Definition 2.3 (NL profile-based recommendation). Given an NL profile d_u for user u and a pool of candidate items C , the downstream recommendation task is to assess the relevance of items in C against the user interests articulated in d_u .

This task is typically operationalized through a scoring function $\mathcal{S}(d_u, x)$ that evaluates each candidate item $x \in C$ based on the NL profile. This generally supports one of two objectives:

- **Top- k ranking:** $\mathcal{S}$ acts as a relevance scorer, and the system returns the top- k items from C as a ranked list R_k .
- **Rating prediction:** $\mathcal{S}$ acts as a regression or classification function estimating the explicit feedback (e.g., a star rating) the user would assign to x .

While NL profiles can support both formulations, we focus primarily on the top- k ranking task. Ranking is widely argued to be the more realistic, user-facing objective in the broader recommender systems literature (Cremonesi et al., 2010; Steck, 2013), and it is how the recommendation task is predominantly formulated in the NL profile research surveyed here.

For evaluation, we let H'_u denote the held-out portion of the user’s interaction history that has not been used in the generation of d_u , which serves as the ground truth against which recommendations are compared.

3 Methods

Although an end-to-end NL-profile-based recommender system necessarily involves both pro-

file generation and profile-based recommendation, most studies specialize in one and treat the other as a fixed pipeline component. A few works close the loop via integrated pipelines (Ramos et al., 2024; Sabouri et al., 2025) or joint optimization (Gao et al., 2025; Tennenholtz et al., 2025; Penaloza et al., 2025; Öncel et al., 2026; Tan et al., 2026).

Below, we survey NL profile-based recommendation (Section 3.1) before NL profile generation (Section 3.2) because downstream recommendation performance is one of the primary evaluation signals for profile generation methods.

3.1 NL Profile-based Recommendation

We organize approaches based on how NL profiles are consumed: as direct input to a generative model, as a source of dense representations for retrieval, or as one signal fused with other learned representations.

3.1.1 LLM-as-a-Recommender

The simplest family treats the LLM itself as the recommendation engine. The model receives an NL profile and item metadata to directly predict a rating or ranking. This splits into two paradigms: The *prompting* variant passes the profile–item input to an instruction-tuned LLM in a zero- or few-shot manner, without any task-specific training (Zhou et al., 2024; Gagliano et al., 2025; Sanner et al., 2023; Zhou et al., 2026; Ngo Huu and Lauw, 2026). The *fine-tuning* variant trains the LLM on labeled (profile, item, rating) tuples; for example, Ramos et al. (2024) fine-tune GPT-2 to perform regression over profile–item pairs to predict a rating.

3.1.2 Embedding-Based Retrieval

A second family encodes profiles and items into a shared dense vector space, ranking items by similarity. We distinguish two sub-variants: *Fine-tuned bi-encoders* optimize dedicated models against the recommendation objective (Mysore et al., 2023b; Sguerra et al., 2025). *Static embedding retrieval*, by contrast, relies on frozen pre-trained text encoders to retrieve items via nearest-neighbour search (Arustashvili and Balog, 2026; Zhou et al., 2026). A contrastive learning objective can also be introduced to refine this representation space (Gao et al., 2025).

3.1.3 Hybrid Representation

Hybrid approaches combine the interpretability of NL profiles with the predictive power of latent collaborative filtering embeddings. Penaloza et al.

(2025) propose TEARS, aligning a fine-tuned text encoder with a black-box latent encoder via optimal transport. The final user representation is a convex combination of the NL profile embedding and the latent embedding. CCBP (Öncel et al., 2026) extends this design to multimodal content-based settings, aligning an LLM-generated profile to a frozen CF backbone.

3.2 NL Profile Generation

Approaches to NL profile generation range from human-authored text to fully automated, task-optimized pipelines. We present them from least to most coupled with the downstream task, reflecting their varying optimization goals: user-facing interpretability versus downstream performance. Table 1 summarizes the trade-offs across these four approaches.

3.2.1 Human-authored Profiles

Collecting profiles directly from users captures preferences without summarization artifacts, serving as a quality ceiling and a source of training data. While scaling this is expensive, crowdsourced elicitation has proven three to four times faster than equivalent item-based preference elicitation (Saner et al., 2023).

3.2.2 Direct LLM Prompting

The most common approach prompts an LLM with a user’s raw interaction history to produce a free-text preference summary. Its simplicity has driven adoption across diverse domains, including movies (Zhou et al., 2024, 2026), music (Gagliano et al., 2025; Sguerra et al., 2025), scientific literature (Arustashvili and Balog, 2026), and points of interest (Mysore et al., 2023b). While various prompting strategies have been explored—such as prompt engineering, output length tuning, and temporal decomposition (detailed in Appendix D)—a recurring concern is faithfulness. Systematic biases occur (e.g., genre biases in music summaries (Sguerra et al., 2025)), and user-perceived representativeness correlates only weakly with downstream recommendation performance.

3.2.3 Structured Prompting

A natural extension introduces a preprocessing step to structure the raw interaction history before prompting the LLM. By injecting structured context rather than raw logs, these methods aim to reduce noise and surface latent preference signals. Implementations range from extracting ex-

Table 1: Trade-offs of NL profile generation approaches.

Approach	Pros	Cons
Human-authored Profiles	• highest authenticity • reference for NL profile evaluation	• does not scale • users may struggle to articulate preferences
Direct LLM Prompting	• simple • no domain or task-specific training	• weak faithfulness • systematic biases
Structured Prompting	• better grounding	• preprocessing overhead
Trained and Optimized Generation	• aligns with downstream task • strong ranking gains	• risk to drift from readable form • needs labels/rewards • training costs

plicit preference features to applying topic modeling, utilizing aggregated metadata or combining textual descriptions of collaborative neighbors (see Appendix D for specific mechanisms) (Ramos et al., 2024; Sun et al., 2025; Fabbri et al., 2025; Ngo Huu and Lauw, 2026).

3.2.4 Trained and Optimized Generation

Motivated by the observation that zero-shot prompting produces readable but unoptimized profiles, recent work trains the profile-generating LLM directly in service of the downstream objective. Approaches include end-to-end fine-tuning via reinforcement learning against ranking metrics (Gao et al., 2025) or framing generation as learned compression under partial observability, where the LLM is trained to produce sufficient statistics for a downstream decision agent (Tennenholtz et al., 2025). BLUE (Tan et al., 2026) instead trains a profiler LLM via GRPO against an embedding-space alignment reward, with a KL constraint to preserve fluency. These methods close the loop between generation and recommendation, but introduce the tension between task alignment and transparency, which we discuss in Section 6.

4 Datasets

NL profile research draws on a relatively small set of benchmark datasets, almost all of which were originally designed for collaborative filtering or review-based recommendation and have been repurposed for profile generation. The exception is SciNUP (Arustashvili and Balog, 2026), which was purpose-built for NL profile-based recommendation and is the only publicly available dataset that pairs profiles with ground-truth relevance labels (Section 4.2). As Table 2 shows, the majority of

Table 2: Datasets used for NL user profile construction. *Original Purpose* identifies the initial target task. *Input* lists the user data supplied to the generator. *Method* classifies the adaptation strategy: Summarization (condensing existing text), Abstraction (inferring from non-textual signals), or Manual (human-authored). Shaded rows indicate purpose-built and publicly available datasets for NL profiles.

Dataset	Domain	Original Purpose	NL Profile Generation		Papers
			Input	Method	
Amazon Re-views	Books, Movies, E-comm.	Rating prediction	Reviews and Ratings	Summarization	(Gao et al., 2025) (Books, Movies), (Sabouri et al., 2025) (Movies, Games), (Torbati et al., 2025) (Books)
			Ratings and Metadata	Abstraction	(Tennenholtz et al., 2025) (E-comm.), (Tan et al., 2026) (Clothing, Books, Electronics, Sports) (Ngo Huu and Lauw, 2026) (Toys, Movies)
H&M	Fashion	Ranking	Item images	Abstraction	(Öncel et al., 2026)
MovieLens	Movies	Rating prediction	Ratings	Abstraction	(Penaloza et al., 2025) (ML1M), (Zhou et al., 2024) (Tag Genome), (Sun et al., 2025), (Tennenholtz et al., 2025) (ML-25M), (Zhou et al., 2026) (ML-25M), (Öncel et al., 2026) (ML-20M)
				Manual	(Sanner et al., 2023) (ML-25M)
Deezer	Music	Ranking	Tracks	Abstraction	(Sguerra et al., 2025)
Localify	Music	Ranking	Artists	Abstraction	(Gagliano et al., 2025)
MSD	Music	Ranking	Audio	Abstraction	(Öncel et al., 2026)
Spotify	Podcasts	Ranking	History	Abstraction	(Fabbri et al., 2025)
TripAdvisor	Travel	Rating prediction	Reviews and Ratings	Summarization	(Ramos et al., 2024)
PointRec	Travel/POI	Ranking	Reviews	Summarization	(Mysore et al., 2023b)
Google Local	Travel/POI	Ranking	Ratings	Abstraction	(Tan et al., 2026) (NY State)
SciNUP	Sci. Lit.	NL profile-based rec.	Papers	Summarization	(Arustashvili and Balog, 2026)
CiteULike	Sci. Lit.	Ranking	Interactions	Abstraction	(Ngo Huu and Lauw, 2026)
Goodreads / Goodbooks	Books	Rating pred. / Ranking	Reviews	Summarization	(Penaloza et al., 2025), (Torbati et al., 2025)

work focuses on movies and music, with sparser coverage of books, travel, scientific literature, and e-commerce—in sharp contrast to a complete absence of datasets in critical domains such as health, education, or news where the user-facing benefits of transparent profiles arguably matter most. Per-dataset details are given in Appendix E. Here we focus on the broader pattern: how each dataset is converted into a collection of NL profiles, and what consequences that conversion has for downstream evaluation. Per-paper implementation details—including evaluation splits, candidate pools, and the specific LLMs used for profile generation—are summarized in Appendix F.

4.1 Adaptation Strategies and Trade-offs

As detailed in Table 2, converting existing benchmarks into NL profile collections relies on three primary adaptation strategies. *Summarization* distills

user-generated text (e.g., reviews or publications) into coherent profiles. This yields well-grounded narratives anchored in authentic user language, though researchers must make heuristic decisions about what evidence to aggregate. *Abstraction* infers latent preferences purely from numerical or behavioral signals (e.g., ratings or listening histories) using item metadata as a semantic bridge. While this vastly broadens the pool of usable datasets, it elevates the risk of hallucination: when metadata is sparse, the LLM may fabricate preferences or default to generic, uninformative descriptions. Finally, *Manual authoring* bypasses automated reconstruction entirely, yielding uniquely authentic profiles that are difficult to scale but invaluable as ground-truth references. Crucially, the choice of strategy dictates what downstream evaluation can actually measure. Both summarization and abstraction rest on the unvalidated assumption that pref-

Table 3: The Reference-Anchored Evaluation Taxonomy (RAET). Each row represents a reference signal; cells indicate properties measured against the output being evaluated (d_u or R_k), along with the papers reporting them.

Ref.	Intrinsic (d_u)		Extrinsic (R_k)	
H_u	faithfulness	(Tennenholtz et al., 2025)	N/A	N/A
	groundedness			
H'_u	N/A	N/A	accuracy	Nearly all surveyed papers
u	representativeness	(Ramos et al., 2024; Sguerra et al., 2025; Gao et al., 2025; Fabbri et al., 2025; Tennenholtz et al., 2025)	satisfaction	(Sun et al., 2025; Sanner et al., 2023)
	self-identification		perceived utility	
	perceived relevance		engagement	
d_u	fluency	(Penaloza et al., 2025; Torbati et al., 2025)	profile-rec.	align-
	coherence		ment	—
	conciseness			
	lexical diversity			
d'_u	edit conciseness	—	scrutability	(Ramos et al., 2024; Penaloza et al., 2025; Zhou et al., 2026; Öncel et al., 2026; Tan et al., 2026)
	edit coherence		controllability	
	sustained readability			
R_k	N/A	N/A	diversity	—
			novelty	
			fairness	

ferences can be faithfully reconstructed from past interactions, whereas manual authoring assumes users can accurately articulate their own needs. Consequently, summarized profiles can be intrinsically checked for groundedness against their source text, whereas abstracted profiles are much harder to verify, forcing researchers to rely on extrinsic recommendation accuracy as an indirect (and conflated) proxy for profile quality.

4.2 Scale and Public Availability

The scale of profile collections varies considerably, from tens of profiles in small user studies (Sguerra et al., 2025; Sanner et al., 2023) to tens of thousands in larger experimental setups (Gao et al., 2025; Ramos et al., 2024; Sabouri et al., 2025). Only a small number of works release their generated profiles publicly (Ramos et al., 2024; Arustashvili and Balog, 2026; Sabouri et al., 2025), and even among these, the coverage is incomplete: Torbati et al. (2025), for example, release only keyword and phrase variants, omitting their LLM-generated narratives. The resulting opacity makes cross-study comparison difficult: two papers may report results on the same nominal benchmark while operating on entirely different profile collections, generated by different LLMs with different prompts. As noted above, SciNUP (Arustashvili and Balog, 2026) is the only publicly available dataset purpose-built for NL profile-based recommendation. The Localify (Gagliano et al., 2025) and the Deezer (Sguerra et al., 2025) datasets are

proprietary streaming-platform datasets and are not publicly available.

The scarcity of publicly available profile collections remains a structural bottleneck for the field. Without shared, reusable profiles, progress is difficult to build upon cumulatively. More purpose-built, publicly released datasets—spanning a broader range of high-stakes domains—are urgently needed to move the field toward reproducible, standardized evaluation.

5 Metrics

Evaluation of NL profile-based recommendation operates at two levels: *intrinsic* evaluation of the profile d_u as a textual artifact, and *extrinsic* evaluation of the downstream ranking R_k . In practice, studies heavily favor extrinsic evaluation, leaving intrinsic properties underexplored. This imbalance persists partly because the field lacks a shared vocabulary to organize these metrics as complementary axes. The adjacent RAG community has addressed similar fragmentation using frameworks that explicitly pair outputs with ground-truth references (Yu et al., 2024). We adopt an analogous approach, introducing the Reference-Anchored Evaluation Taxonomy (RAET), specialized for NL profile-based recommendation with task-specific reference signals. RAET associates each metric with the specific reference signal against which the evaluable output—either the profile d_u (intrinsic) or the ranking R_k (extrinsic)—is compared. Summarized in Table 3, each half of the taxon-

omy is organized around six reference signals, revealing where the field concentrates and where structural gaps remain. Related surveys on controllable learning (Shen et al., 2026), LLM personalization (Zhang et al., 2025), LLMs for recommendation (Wu et al., 2024), and LLM-RL integration (Gao et al., 2026) address adjacent methodological questions but do not organize evaluation around the NL profile itself. Notably, the choice of reference signal is orthogonal to the type of evaluator (human, automatic, or LLM-as-a-Judge). Per-paper coverage of each axis is detailed in Appendix G.

5.1 Intrinsic Axes

Interaction history (H_u). This axis assesses whether the profile accurately reflects behaviors in H_u without hallucinating claims, expressed as *faithfulness* and *groundedness*. Explicit checks of this kind are nearly absent from the surveyed work. The only example is Tennenholtz et al. (2025), who ask human raters and an LLM judge to rate profile $\leftrightarrow$ history consistency on a five-point scale.

User (u). This axis assesses whether users recognize themselves in the profile, measured by *representativeness*, *self-identification*, and *perceived relevance*. As the only intrinsic axis directly capturing the user-facing motivation for NL profiles, it is the best-populated in practice. Evaluations range from human ratings (Ramos et al., 2024; Sguerra et al., 2025) to LLM-as-a-Judge estimates (Fabbri et al., 2025). Gao et al. (2025) introduce a behavioral variant, using the time taken by a judge to predict the next item from a profile as a proxy for cognitive load. Generally, studies find that while generative fluency is reliably high, representativeness varies significantly across LLMs and user types (Sguerra et al., 2025; Fabbri et al., 2025).

Profile alone (d_u). This axis covers reference-free linguistic properties like *fluency*, *coherence*, *conciseness*, and *lexical diversity*. These are computed automatically: Penaloza et al. (2025) use BLEU, pairwise edit distance, profile length, and conciseness to measure profile uniqueness and diversity across users, while Torbati et al. (2025) analyze POS-tag and IDF-weight distributions between generated and source-text distributions as proxies for writing style and faithfulness.

Edit (d'_u). This axis captures whether the edited profile is more *concise* or *coherent* than the origi-

nal, whether editing preserves *user-readability*, and whether a dynamically updated profile remains legible over time? In the surveyed literature, edits are simulated or researcher-controlled rather than performed by real users, leaving this side essentially unexplored; filling the gap will require user studies of authentic edits and continuously updated profiles.

5.2 Extrinsic Axes

Ground truth (H'_u). This axis assesses *accuracy* against held-out interactions and is the workhorse of the field. Depending on task framing, papers report rating prediction errors (RMSE, MAE) (Ramos et al., 2024; Zhou et al., 2024) or top- k ranking metrics (NDCG, Recall, MRR, and MAP) (Gao et al., 2025; Mysore et al., 2023b; Sguerra et al., 2025; Arustashvili and Balog, 2026; Torbati et al., 2025; Sabouri et al., 2025; Penaloza et al., 2025). (Appendix C summarizes standard baselines.) While essential for establishing utility, this axis conflates profile quality with the recommender’s capabilities, obscuring the profile’s isolated contribution.

User (u). This axis captures *satisfaction*, *perceived utility*, and *engagement* via user feedback on recommendations. Few studies go beyond offline metrics here. Sun et al. (2025) conduct an eight-week longitudinal field experiment—an expensive but uniquely informative method that detects behavioral shifts missed by offline benchmarks. Behavioral variants use preference-elicitation speed as an indirect utility signal (Sanner et al., 2023).

NL profile (d_u). This axis assesses if the recommended items actually reflect the preferences the profile claims to encode, that is, *profile $\leftrightarrow$ recommendation alignment*. It mirrors intrinsic scrutability, as high alignment ensures that editing the profile reliably reshapes recommendations. Though essentially absent from surveyed literature, the taxonomy highlights it as a critical blind spot.

Edit (d'_u). This axis measures whether deliberate profile edits yield predictable recommendation changes—*scrutability* and *controllability*. This is a key advantage of NL profiles relative to latent representations. Ramos et al. (2024) validate scrutability by measuring how recommendations change after simulated profile edits, while Penaloza et al. (2025), SteerEval (Zhou et al., 2026), and CCBP (Öncel et al., 2026) evaluate controllability systematically

across a range of interventions—interest flipping, item boosting, instruction-based steering, and template or language-level rewrites, and attribute-level removal or addition. However, protocols remain unstandardized, making cross-system comparisons difficult.

Ranking alone (R_k). This axis covers reference-free properties of the ranked set itself, such as *intra-list diversity* and *novelty*—the extrinsic counterpart to the intrinsic *profile alone* axis. While standard in mainstream recommendation evaluation, these beyond-accuracy metrics remain largely unadopted in NL profile research.

6 Discussion and Open Challenges

The Transparency vs. Task-Optimization Tension. A core motivation for NL profiles is user inspectability. Yet, to improve downstream recommendation accuracy, recent methods optimize the profile generator directly against ranking objectives using reinforcement learning or learned compression (Gao et al., 2025; Tennenholtz et al., 2025; Tan et al., 2026). This creates a fundamental tension: task-optimized profiles tend to drift toward an internal model language that maximizes utility but sacrifices genuine human readability and editability. The field currently lacks a principled mechanism for quantifying or controlling this trade-off. Resolving this tension requires more user studies and evaluation protocols centered on real user needs.

The Intrinsic–Extrinsic Evaluation Disconnect

The above optimization tension creates an evaluation bottleneck: intrinsic profile quality correlates only weakly with extrinsic accuracy (Sguerra et al., 2025; Fabbri et al., 2025), and recent task-optimized work often skips intrinsic evaluation entirely (Öncel et al., 2026; Tan et al., 2026). Furthermore, study-specific protocols and varying LLM-as-a-Judge prompts make cross-study intrinsic comparisons difficult. Our evaluation taxonomy (RAET, in Section 5) clarifies why this disconnect persists: existing work concentrates heavily on three isolated reference signals (extrinsic ground truth, intrinsic user, and intrinsic profile). Crucially, axes that bridge generation and recommendation are underpopulated: extrinsic user satisfaction is covered by only a handful of studies, while the extrinsic profile–ranking alignment axis is essentially blank. Going forward, we argue that studies should evaluate more holistically, jointly reporting extrin-

sic and intrinsic measures. Because optimizing one isolated signal often degrades another, unified evaluation must do more than report both sides independently; it must measure how well transparency is preserved under optimization, as seen in partial-transparency systems (Penaloza et al., 2025) where only specific layers remain legible.

Reconciling Collaborative Signals with Faithfulness.

The drive to improve extrinsic accuracy naturally points toward incorporating collaborative filtering (CF) signals (“users like you also liked”). However, injecting CF signals into *profile* generation severely complicates intrinsic evaluation. If a profile surfaces an interest a user has not yet explicitly expressed, it may be highly predictive (extrinsic success) but factually ungrounded in their history (intrinsic failure). Fusing CF signals with natural language while keeping the resulting profile faithful, interpretable, and user-editable remains an open modeling challenge, though hybrid alignment (Penaloza et al., 2025; Öncel et al., 2026), CF-as-text approaches (Ngo Huu and Lauw, 2026), and collaborative memory graphs (Chen et al., 2026) are early attempts in this direction.

Agentic Paradigms as a Synthesizing Solution.

To break these deadlocks, the field must move beyond static, single-pass profile generation. An agentic paradigm—where an LLM iteratively constructs, critiques, and revises a profile through tool use and self-correction—offers a structural solution to the interdependencies above. Drawing on frameworks like Reflexion (Shinn et al., 2023), PersonaX (Shi et al., 2025), or MemRec (Chen et al., 2026), an agent could query CF signals as external tools, safely integrating collaborative insights, while explicitly using self-critique prompts to ensure the final narrative remains perfectly aligned with human-readable transparency constraints.

7 Conclusion

This survey reviewed the emerging body of work on natural language user interest profiles for recommendation. We proposed a unifying definition, distinguished NL profiles from related text-based representations, and organized the literature across *methods*, *datasets*, and *metrics*. Finally, we introduced the Reference-Anchored Evaluation Taxonomy (RAET) to bridge intrinsic and extrinsic evaluation, and identified the structural open challenges shaping the future of the field.

Limitations

The literature search, while conducted carefully, cannot guarantee complete coverage—given the terminological fragmentation in this area, some relevant work may have been missed, as discussed in Appendix A. Furthermore, our synthesis is inherently limited by the structural immaturity of the field. The heavy reliance on repurposed, proprietary datasets and the absence of standardized evaluation protocols preclude any rigorous quantitative meta-analysis of cross-study performance. Finally, the surveyed literature exhibits a strong English-centric bias; whether the generative quality, interpretability, and user acceptance of NL profiles hold across multilingual and cross-cultural contexts remains a significant blind spot.

Ethical considerations

NL profiles externalize sensitive inferences about users in a human-readable form. They might contain inferences users would not have made themselves, can carry stereotyped or biased framings inherited from the underlying LLM, and can persist beyond the user’s awareness. A particular concern is that interaction histories used to generate profiles—reviews, ratings, listening logs—were often not collected with this purpose in mind, raising questions about the scope of the original consent. Furthermore, a profile distilled from browsing or viewing history may inadvertently surface sensitive attributes such as health concerns, political views, or religious interests that the user never explicitly disclosed. Even more concerning is the risk of generative hallucination: an LLM might infer and record deeply stigmatizing or inappropriate preferences based on noisy data, before the user has the chance to correct it.

Conversely, the same explicitness that creates these risks also creates an opportunity for genuine user agency: the right to inspect, edit, contest, and delete one’s profile. Realizing this opportunity requires that systems treat the NL profile not merely as an internal representation but as a user-facing artifact with clear ownership and control mechanisms. We see this as one of the most important design obligations for future work in this area.

References

- Jafar Afzali, Aleksander Mark Drzewiecki, and Krisztian Balog. 2021. POINTREC: A Test Collection for Narrative-driven Point of Interest Recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2478–2484.
- Mariam Arustashvili and Krisztian Balog. 2026. Scinup: Natural language user interest profiles for scientific literature recommendation. In *European Conference on Information Retrieval, ECIR ’26*, pages 432–446.
- Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, scrutable and explainable user models for personalized recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’19*, pages 265–274.
- Nicholas J. Belkin and W. Bruce Croft. 1992. Information filtering and information retrieval: two sides of the same coin? *Commun. ACM*, 35(12):29–38.
- Thierry Bertin-Mahieux, Daniel Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset. *ISMIR ’11*, pages 591–596.
- Weixin Chen, Yuhao Zhao, Jingyuan Huang, Zhihe Ye, Mingxuan Ju, Tong Zhao, Neil Shah, Li Chen, and Yongfeng Zhang. 2026. MemRec: Collaborative memory-augmented agentic recommender system. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL ’26, pages 44515–44544.
- Paolo Cremonesi, Yehuda Koren, and Roberto Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys ’10*, page 39–46.
- Michael D. Ekstrand, Afsaneh Razi, Aleksandra Sarcevic, Maria Soledad Pera, Robin Burke, and Katherine Landau Wright. 2025. Recommending with, not for: Co-designing recommender systems for social good. *ACM Trans. Recomm. Syst.*
- Michael D. Ekstrand and Martijn C. Willemsen. 2016. Behaviorism is Not Enough: Better Recommendations through Listening to Users. In *Proceedings of the 10th ACM Conference on Recommender Systems, RecSys ’16*, pages 221–224.
- Francesco Fabbri, Gustavo Penha, Edoardo D’Amico, Alice Wang, Marco De Nadai, Jackie Doremus, Paul Giglioli, Andreas Damianou, Oskar Stål, and Mounia Lalmas. 2025. Evaluating podcast recommendations with profile-aware llm-as-a-judge. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys ’25*, page 1181–1186.
- Paul Gagliano, Griffin Homan, Douglas Turnbull, and Venkata S Govindarajan. 2025. Using language

- models for music recommendations with natural-language profiles. In *Proceedings of the 3rd Music Recommender Systems Workshop, MuRS '25*.
- Mengyao Gao, Chongming Gao, Jiakai Tang, Jingsen Zhang, Xinpeng Zhao, Bohao Wang, Jiawei Chen, Haoran He, Ling Pan, Xu Chen, and 1 others. 2026. [Integrating large language models with reinforcement learning: A survey of LLM-RL synergistic recommendation](#). *TechRxiv*.
- Zhaolin Gao, Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2025. Langptune: Optimizing language-based user profiles for recommendation. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, page 707–717.
- Carlos García Ling, ElizabethHMGroup, FridaRim, inversion, Jaime Ferrando, Maggie, neuraloverflow, and xlsrln. 2022. H&M personalized fashion recommendations. <https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations>. Kaggle Competition.
- Shijie Geng, Shuchang Liu, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2022. Recommendation as language processing (rlp): A unified pretrain, personalized prompt & predict paradigm (p5). In *Proceedings of the 16th ACM conference on recommender systems, RecSys '22*, pages 299–315.
- F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19.
- Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, pages 507–517.
- Yupeng Hou, Jiacheng Li, Xiangjun Fu, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2026. [Bridging language and items for retrieval and recommendation: Benchmarking LLMs as semantic encoders](#). *Preprint*, arXiv:2403.03952.
- Amelia Kovacs, Jerry Chee, and Sarah Dean. 2025. Datasets for navigating sensitive topics in recommendation systems. In *Companion Proceedings of the ACM on Web Conference 2025, WWW '25*, pages 741–744.
- Jiacheng Li, Jingbo Shang, and Julian McAuley. 2022. Uctopic: Unsupervised contrastive learning for phrase representations and topic mining. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '22, pages 6159–6169.
- Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL '21, pages 4947–4957.
- Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26.
- Yongqi Li, Xinyu Lin, Wenjie Wang, Fuli Feng, Liang Pang, Wenjie Li, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2024. [A survey of generative search and recommendation in the era of large language models](#). *Preprint*, arXiv:2404.16924.
- Sheshera Mysore, Mahmood Jasim, Andrew McCallum, and Hamed Zamani. 2023a. Editable user profiles for controllable text recommendations. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23*, page 993–1003.
- Sheshera Mysore, Andrew McCallum, and Hamed Zamani. 2023b. Large language model augmented narrative driven recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems, RecSys '23*, page 777–783.
- Manh-Khanh Ngo Huu and Hady W. Lauw. 2026. Verbalizing lightgcn: Direct learning of textual representations from user-item interaction graph via llms. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '26*, page 1393–1403.
- Fırat Öncel, Jihoon Jeong, Emiliano Penaloza, Mirco Ravanelli, Laurent Charlin, and Cem Subakan. 2026. [Controllable and content-based recommendations](#). *Preprint*, arXiv:2607.20938.
- Emiliano Penaloza, Olivier Gouvert, Haolun Wu, and Laurent Charlin. 2025. Tears: Textual representations for scrutable recommendations. In *Proceedings of the ACM on Web Conference 2025*, pages 4949–4968.
- Filip Radlinski, Krisztian Balog, Fernando Diaz, Lucas Dixon, and Ben Wedin. 2022. On natural language user profiles for transparent and scrutable recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22*, pages 2863–2874.
- Jerome Ramos, Hossein A. Rahmani, Xi Wang, Xiao Fu, and Aldo Lipani. 2024. Transparent and Scrutable Recommendations Using Natural Language User Profiles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '24, pages 13971–13984.
- Milad Sabouri, Masoud Mansoury, Kun Lin, and Bamshad Mobasher. 2025. Towards explainable temporal user profiling with llms. In *Adjunct Proceedings of the 33rd ACM Conference on User Modeling, Adaptation and Personalization, UMAP Adjunct '25*, page 219–227.

- Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL '24, pages 7370–7392.
- Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. 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*, RecSys '23, pages 890–896.
- Bruno Sguerra, Elena V Epure, Harin Lee, and Manuel Moussallam. 2025. Biases in llm-generated musical taste profiles for recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25, pages 527–532.
- Chenglei Shen, Xiao Zhang, Teng Shi, Changshuo Zhang, Guofu Xie, Jun Xu, Ming He, and Jianping Fan. 2026. A survey of controllable learning: Methods and applications in information retrieval. *Frontiers of Computer Science*, 20(10):2010619.
- Yunxiao Shi, Wujiang Xu, Zhang Zeqi, Xing Zi, Qiang Wu, and Min Xu. 2025. Personax: A recommendation agent-oriented user modeling framework for long behavior sequence. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5764–5787.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652.
- Harald Steck. 2013. Evaluation of recommendations: rating-prediction and ranking. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 213–220.
- Ruixuan Sun, Junyuan Wang, Sanjali Roy, and Joseph A Konstan. 2025. Co-authoring the self: A human-ai interface for interest reflection in recommenders.
- Zhaoxuan Tan, Xiang Zhai, Yan Zhu, Meng Jiang, and Mohamed Hammad. 2026. Bridging textual profiles and latent user embeddings for personalization. *Preprint*, arXiv:2605.06981.
- Guy Tennenholtz, Jihwan Jeong, Chih-Wei Hsu, Yinlam Chow, and Craig Boutilier. 2025. Descriptive history representations: Learning representations by answering questions.
- Ghazaleh H Torbati, Anna Tigunova, Gerhard Weikum, and Andrew Yates. 2025. Cup: A framework for resource-efficient review-based recommenders. In *European Conference on Information Retrieval*, ECIR '25, pages 360–375.
- Vittoria Vineis and Gabriele Tolomei. 2026. From engagement to empowerment: A capability-theoretic rethinking of recommender systems. ECIR '26, page 352–368.
- Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.
- Mengke Wu, Weizi Liu, Yanyun Wang, Weiyu Ding, and Mike Yao. 2026. Rethinking user empowerment in ai recommender system: Innovating transparent and controllable interfaces. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26.
- Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. 2024. Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data*, pages 102–120. Springer.
- Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dermoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2025. Personalization of large language models: A survey. *Transactions on Machine Learning Research*.
- Joyce Zhou, Yijia Dai, and Thorsten Joachims. 2024. Language-based user profiles for recommendation. *Preprint*, arXiv:2402.15623.
- Joyce Zhou, Weijie Zhou, Doug Turnbull, and Thorsten Joachims. 2026. Steereval: A framework for evaluating steerability with natural language profiles for recommendation. *Preprint*, arXiv:2601.21105.

A Literature Collection Methodology

Our literature search began with (Radlinski et al., 2022), which introduces the concept of scrutable, natural language-based user representations as an alternative to black-box recommendation models, and served as a natural seed paper for this survey. From there, we followed two complementary trails: forward citation search (papers that cite (Radlinski et al., 2022)) and backward citation search (papers cited within each newly identified work). This snowballing procedure accounted for the majority of the papers included in the survey.

To supplement this, we conducted a series of keyword searches on Google Scholar using terms such as “*natural language user profile*,” “*user interest profile recommendation*,” “*LLM user summary recommendation*,” and “*scrutable user model*.” These searches surfaced a small number of additional papers not reachable through citation trails alone. A known limitation of this step is that the terminology used across the literature is highly inconsistent: the same concept is variously described as a user summary, a narrative profile, a descriptive history representation, or a preference description, making exhaustive keyword coverage difficult to guarantee. Regardless of this terminological fragmentation, we believe the core body of work is represented here.

B NL Profile Examples

Table 4 provides concrete examples of NL or textual user profiles extracted or adapted from the reviewed approaches.

C Baselines for Recommendation Evaluation

We summarize the details of the collaborative filtering, retrieval-based, and LLM-based baseline methods used in the surveyed works.

Collaborative filtering (CF) methods form the most common baseline family, operating on user-item interaction matrices without explicit item content. Matrix factorization variants—including MF, NMF, WRMF, BPR-SLIM, EASE, and VAE-based models—appear across multiple works (Gagliano et al., 2025; Zhou et al., 2024; Sanner et al., 2023; Penaloza et al., 2025; Torbati et al., 2025; Sabouri et al., 2025), as do neighbourhood-based methods such as UserKNN and Item-kNN (Ramos et al., 2024). The prevalence of CF baselines is largely a consequence of dataset choice: most works

repurpose standard recommendation benchmarks (MovieLens, Amazon, Netflix) that were originally designed for the CF setting, making interaction-based methods the natural comparison point.

Retrieval-based methods appear as baselines primarily in works that use datasets or tasks closer to information retrieval. Sparse methods (BM25, RM3) and dense methods (SciBERT-KNN, BGE, Contriever, BERT-MSM) are used in scientific literature and POI recommendation settings (Arustashvili and Balog, 2026; Mysore et al., 2023b), where the task of matching a user profile to a candidate document pool is more naturally framed as retrieval than as interaction-based ranking.

LLM-based methods can broadly be considered a form of content-based filtering, as they operate on textual representations of items and users rather than interaction matrices. Unlike classical content-based methods, however, they leverage world knowledge encoded during pre-training, going beyond item metadata alone. Direct LLM prompting without any profile construction step is used as a baseline to isolate the contribution of the profile generation component (Gagliano et al., 2025; Zhou et al., 2024), while prompting strategy comparisons (zero-shot, few-shot, completion) assess sensitivity to prompt design (Sanner et al., 2023). LLMRank, which prompts an LLM to rank candidates from raw interaction history without profile construction, and P5 (Geng et al., 2022), a T5-based generative recommendation model, represent stronger LLM-based reference points (Torbati et al., 2025). Neural explainable recommendation models PETER+ (Li et al., 2021) and PEPLER-MLP (Li et al., 2023) occupy a borderline position between LLM-based and CF methods, as they combine transformer-based language modeling with interaction-derived supervision (Ramos et al., 2024).

D Prompting Implementation Details

Within the direct prompting paradigm, several approaches have been explored. Prompt-engineering variations—such as contrastive examples, few-shot prompting, and first-person formatting—have generally been found to yield limited gains in downstream recommendation accuracy (Gagliano et al., 2025). Adjusting the target profile length (e.g., from 50 to 200 words) similarly produces inconsistent improvements (Zhou et al., 2024). To capture evolving preferences, temporal structure can be in-

troduced by prompting for separate short-term and long-term interest profiles; these are later fused at the model level via attention mechanisms, surfacing the relative influence of recent versus historical activities (Sabouri et al., 2025).

In the structured prompting paradigm, specific preprocessing mechanisms vary based on the available data. When rich textual feedback is present, explicit preference features can be extracted from user reviews and ratings before the LLM synthesizes them into a coherent profile (Ramos et al., 2024). For interaction-only data, topic modeling and contrastive filtering can be applied to explicitly separate a user’s positive and negative interests prior to prompting (Sun et al., 2025). Raw interaction logs can be replaced by aggregated structural metadata, such as compiled listening statistics (Fabbri et al., 2025), providing the LLM with cleaner semantic anchors. When collaborative signals are available, the LLM can be treated as a graph message-passing operator that iteratively refines user and item profiles from aggregated textual descriptions of collaborative neighbors (Ngo Huu and Lauw, 2026).

E Datasets

This section gives a per-domain walkthrough of the datasets summarized in Table 2.

Movies. *MovieLens* (Harper and Konstan, 2015)¹ is the most widely used benchmark, appearing in multiple variants. ML-1M and ML-25M provide user-movie rating matrices, while the Tag Genome extension² augments items with fine-grained semantic relevance scores for a vocabulary of descriptive tags. *Netflix*³ is a large-scale movie rating dataset originally released for the Netflix Prize. Both are used purely for numerical ratings, which are aggregated and converted into natural language preference summaries (Penaloza et al., 2025; Sun et al., 2025; Zhou et al., 2024; Sanner et al., 2023; Tennenholtz et al., 2025; Zhou et al., 2026). CCBR (Öncel et al., 2026) uses ML-20M with video trailers as item content, generating profiles from video-derived per-item descriptions rather than from ratings. SteerEval (Zhou et al., 2026) additionally uses *NavigatingSensitivity 2025* (Kovacs

et al., 2025), a set of 128 movie trigger-warning tags to evaluate steerability on niche topics. *Amazon Movies & TV* (He and McAuley, 2016; Hou et al., 2026) supplements ratings with user-written reviews and rich item metadata (title, description, category, price), enabling more content-grounded profile generation (Gao et al., 2025; Ramos et al., 2024).

Books. *Amazon Books* mirrors the Movies & TV collection in structure and is used analogously, with user reviews and product metadata feeding the profile generation process (Gao et al., 2025). *Goodbooks*⁴ offers ratings and metadata without review text and is used similarly to MovieLens and Netflix by TEARS, where titles, ratings, and genres are prompted to generate user preference summaries (Penaloza et al., 2025).

Music. *Spotify* and *Localify* (Gagliano et al., 2025) are used for music and podcast recommendation. Localify is a small dataset of users linked to their Spotify accounts, providing listening histories from which artist-level profiles are constructed (Gagliano et al., 2025; Fabbri et al., 2025). *Deezer* is a music streaming dataset containing user listening behavior at the track level, used by Sguerra et al. (2025) to generate short natural language taste profiles. The *Million Song Dataset* (MSD) (Bertin-Mahieux et al., 2011) provides audio clips and is used by CCBR (Öncel et al., 2026) for content-based recommendation, with profiles generated from audio-derived per-item descriptions. Localify and Deezer are proprietary streaming-platform datasets and are not publicly available.

Travel and POI. *TripAdvisor*⁵ provides user reviews of hotels and attractions and is used alongside Amazon data by Ramos et al. (2024) to construct preference profiles from review text. *PointRec* (Afzali et al., 2021) is a point-of-interest recommendation dataset specifically designed for narrative-driven recommendation; user reviews are used to generate natural language queries reflecting individual travel preferences. *Google Local Reviews* (Li et al., 2022) extends the domain to a broader set of local businesses (restaurants, shops). BLUE (Tan et al., 2026) uses the New York state subset, treating interaction sequences as generator input.

¹<https://grouplens.org/datasets/movielens>

²<https://grouplens.org/datasets/movielens/tag-genome-2021/>

³<https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data>

⁴<https://github.com/zygmuntz/goodbooks-10k>

⁵<https://www.tripadvisor.com>

E-commerce. *Amazon reviews* are used by [Tenenholtz et al. \(2025\)](#), who generate NL summaries of a user’s shopping history via a trained LLM, and by BLUE ([Tan et al., 2026](#)), which trains a profiler LLM against an embedding-alignment reward across four Amazon Reviews 2023 subsets (Clothing, Books, Electronics, Sports). *H&M fashion dataset* ([García Ling et al., 2022](#)) focuses on the fashion retail domain, providing product images (1,943 items, 70,201 users) used by CCBR ([Öncel et al., 2026](#)) for image-based content recommendation; profiles are generated from vision-language captions of items in the user’s interaction history.

Scientific literature. *SciNUP* ([Arustashvili and Balog, 2026](#)) is the only dataset purpose-built for NL profile-based recommendation. It provides researcher interest profiles summarized from each researcher’s published papers, a candidate paper pool, and ground-truth relevance labels, making it directly applicable to the evaluation of NL profile-driven scientific paper recommendation. Verba-LightGCN ([Ngo Huu and Lauw, 2026](#)) uses a small user-item subset from the *CiteULike* bibliography bookmarking platform to construct textual profiles from the CF interaction graph.

F Implementation Details Across Surveyed Papers

Table 5 summarizes three reproducibility-relevant implementation choices for every surveyed paper that reports an empirical evaluation: how the user’s interaction history is split between profile generation and evaluation, what candidate pool the recommendation metrics are computed over, and which LLM produces the natural language profiles.

G Evaluation

Tables 6 and 7 map the surveyed papers onto the RAET axes proposed in Section 5.

[Fabbri et al. \(2025\)](#) collect pointwise user ratings of recommended podcast episodes and employ an LLM-as-a-Judge that consumes the NL profile as context to predict whether a recommended item aligns with user taste. While this constitutes an extrinsic evaluation of NL profiles, it is excluded from Table 7 because recommendations are not generated from the NL profiles—the profiles serve only as context for the judge, not as the basis for retrieval or ranking.

Table 4: Examples of NL user profiles reproduced from the cited published sources.

Paper	Example NL Profile
(Sun et al., 2025)	“AI-Generated - Movies that feature the unique animation style of Aardman and showcase performances by Alan Rickman are appreciated. - Films that include Michael Peña and explore the adventures of Nick and Nora Charles, including their series, are favored. - Movies that are adaptations of books, have won Oscars for original scores, and are based on true stories are enjoyed.”
(Penaloza et al., 2025)	“The user has a strong preference for crime dramas, often enjoying those that incorporate elements of thrillers and mysteries. They are particularly drawn to complex narratives that involve intricate plots, moral ambiguity, and character-driven stories. The user appreciates the tension and intellectual engagement that comes with unraveling a mystery or following the maneuvers of a criminal mastermind. They also show a high regard for films that blend crime with dramatic depth, exploring the human condition and the consequences of criminal activity. On the other hand, the user does not enjoy crime movies that incorporate excessive comedy or romance, suggesting a preference for more serious and gritty narratives over those that might dilute tension with humor or love stories. While other viewers may find the lighter moments in action-comedy crime films entertaining, this user tends to steer clear of those plot points, favoring a more straightforward and intense viewing experience.”
(Torbaty et al., 2025)	“I’m a history buff with a passion for non - fiction books, particularly those that delve into the lives of ordinary people during extraordinary times. I enjoy reading about the experiences of women, immigrants, and marginalized groups”
(Gagliano et al., 2025)	“My musical taste weaves velvety jazz-infused neo-soul with driving, synth-heavy electronic grooves that effortlessly blend warmth and futurism. I gravitate toward smoky, upright-bass-led lounge numbers that evoke intimate club corners, alongside pulsating house tracks that ignite late-night dancefloors. I refresh my sets with experimental ambient textures and lo-fi hip-hop beats that layer nostalgic vinyl crackle over head-nodding rhythms, while occasional avant-garde free-jazz injections add daring dissonance. This balance of cozy soulfulness and boundary-pushing sonic exploration speaks to my love of music that soothes and stimulates in equal measure. The result is a listening profile rooted in sophistication and spontaneity, comfort and curiosity, offering a journey that feels both familiar and thrilling.”
(Arustashvili and Balog, 2026)	“My research lies at the intersection of information theory, probability, and convex geometry, with a particular focus on entropy inequalities, information-theoretic limits, and log-concave distributions. I investigate fundamental properties of entropy in both discrete and continuous settings, develop analogs of additive combinatorics in an information-theoretic framework, and explore applications to problems such as compound Poisson approximation, polar coding, and combinatorial geometry. My work often uncovers deep connections between probabilistic methods, functional inequalities, and structural aspects of high-dimensional spaces.”
(Santer et al., 2023)	“I like comedy genre movies, while watching comedy movies I will feel very happy and relaxed. Comedy films are designed to make the audience laugh. It has different kinds of categories in comedy genres such as horror comedy, romantic comedy, comedy thriller, musical-comedy. I dislike action genre movies because watching fights gives me a headache and bored me. These kinds of movies mainly concentrate on violence and physical feats.”
(Mysore et al., 2023b)	“Looking for restaurants and food in Rotterdam, NL Hi all - my father and I will be visiting your fair city next week, and we’re looking for a place to have dinner on a Sunday night. Are there any places that you would recommend? We like tasty food, good beer, and a non-stuffy atmosphere. Considerations: Open on Sunday (I gather this is somewhat limiting?) Our hotel is right near the Leuvehaven metro stop, and walking distance would be nice ideally not Asian(-inspired) food - we’re from southeast Asia and would rather try something different. Totally up for everything else though! Thank you!”
(Zhou et al., 2024)	“You enjoy movies with high ratings, such as The Imitation Game, Princess Bride, The Matrix, and Lord of the Rings. You dislike movies with low ratings, such as Star Wars: Episode II - Attack of the Clones, Pirates of the Caribbean: At World’s End, and Suicide Squad. You also have mixed opinions on movies like Inglourious Basterds, The Incredibles, and Wonder Woman.”
(Ramos et al., 2024)	“I enjoy watching movies and TV shows, particularly those with strong performances and well-crafted stories. I have a preference for classic films and directors such as Bergman and Fellini, and I appreciate a good adaptation of a play. I also enjoy watching actors take on complex roles and explore their craft. While I have mixed opinions on some actresses, I generally enjoy watching a talented performance. Recently, I have been drawn to films about Elizabeth I, and I appreciate a good historical drama. Additionally, I find it refreshing when a director can successfully weave together multiple storylines and timelines into a cohesive and engaging film.”

Paper	Example NL Profile
(Tennenholtz et al., 2025)	“This user demonstrates varied shopping interests, with a clear preference for practical and comfortable footwear and accessories for themselves and potentially others. A recurring theme is a fondness for items emphasizing ease and warmth, evidenced by the 5-star ratings for snow boots (Totes Zelus & Eric), walking shoes (New Balance), and a stretch bracelet (Simulated Pink Pearl). While appreciating affordability (as revealed by purchases from brands like totes and M&c), the user isn’t averse to investing in quality when it aligns with their needs, indicated by the 4-star rating for Frye boots. Style leans towards sporty-casual and practical, with a tolerance for some embellishments (beaded bracelet, lacrosse necklace – likely a gift). Negative ratings suggest an unwillingness to purchase ill-fitting (Dickies pants) or poorly-performing (Vikorous top) items. An unclear gender preference outside of sale of women’s apparel.”
(Gao et al., 2025)	“Storytelling: The user enjoys complex storylines with twists and surprises. They appreciate movies that explore themes of adventure, discovery, and suspense. The user particularly liked the unexpected connections and surprises in “Life of Pi” and the twists in “10 Cloverfield Lane”. Suspense and Thrills: The user enjoys suspenseful movies that keep them guessing until the end. They highly rated “10 Cloverfield Lane” for its ability to maintain tension and uncertainty throughout the film. Visuals: The user appreciates visually stunning movies with groundbreaking effects. They praised the visuals in “Life of Pi” and might have appreciated the action sequences and spy-thriller elements in “Kingsman: The Secret Service”.”
(Zhou et al., 2026)	“This user has a strong affinity for timeless, character-driven stories that often involve themes of redemption, love, and the human condition, frequently set against a backdrop of adventure, drama, and wit.”
(Tan et al., 2026)	“... Furthermore, the user displays a clear interest in children’s apparel, particularly with the Disney Girls’ Doc McStuffins and Gerber Baby items, suggesting a desire to provide comfort and security for their children.”
(Ngo Huu and Lauw, 2026)	“The user enjoys emotionally charged and character-driven stories, visually dynamic and action-packed content, morally complex storylines, and critically acclaimed, award-winning content. They appreciate complex characters, morally challenging narratives, and realistic portrayals of real-life events, particularly in military and war themes...”
(Sguerra et al., 2025), (Fabbri et al., 2025), (Öncel et al., 2026)	<i>Example not available</i>

Table 5: Evaluation protocols across surveyed papers. *Split* describes how each user’s interaction history is partitioned between the portion that feeds d_u and the portion that serves as ground truth H'_u . *Candidate pool* describes the items ranked at evaluation time. *LLM* lists the model(s) used to generate the NL profiles.

Paper	Split	Candidate pool	LLM
Arustashvili and Balog (2026)	Per-user temporal: earlier half of authored papers feeds d_u , later half generates ground truth via references	1,000 items: ground truth + random arXiv negatives weighted by user’s categories	Llama-4 Maverick; GPT-4o
Zhou et al. (2024)	Per-user random subset of 150-rating history; 3 held-out items per sample for evaluation	N/A (rating prediction & pairwise tasks)	Llama 2 7B/13B; Sakura-SOLAR 10.7B
Fabbri et al. (2025)	90-day listening window feeds d_u ; live user study (47 participants)	N/A (pointwise user ratings)	GPT-4.1
Gagliano et al. (2025)	Per-user random split of seed artists into profile half and candidate half	Candidate half + same-sized distractor set from other users	GPT-4o-mini; Gemma-3-4B-it
Gao et al. (2025)	Session-based: 4-item history feeds d_u , next item is the target	Full catalogue (10,533 Movie-TV; 4,819 Books)	Gemma-2B-it; Llama-3-8B-it
Mysore et al. (2023b)	No internal split; synthetic Yelp queries train retrieval, Pointrec (112 queries) for evaluation	Items filtered by query city and request category	text-davinci-003 (175B Instruct-GPT)
Penaloza et al. (2025)	Per-user chronological: oldest ≤ 50 ratings feed d_u , remaining ratings evaluate; user-level holdout (500/2,000 users)	Full catalogue (Recall@ k , NDCG@ k at $k \in \{20, 50\}$)	GPT-4-turbo; LLaMA 3.1-405b
Ramos et al. (2024)	8:1:1 random split with warm-start constraint (every user/item in every split; ≥ 5 reviews per user in training)	Condensed lists: only items rated in the test set	Llama2-7B-Instruct; Mistral-7B-Instruct
Sabouri et al. (2025)	Per-user chronological 60/20/20 (training/validation/test)	Held-out items + 5 negatives per positive (training); evaluation pool size unspecified	GPT-4o-mini
Sanner et al. (2023)	No offline split; 153-rater user study with personalized judged pools	40 items per rater (10 RandPop + 10 RandMidPop + 10 EASE + 10 BM25-Fusion)	PaLM 62B
Sguerra et al. (2025)	Time-windowed: top-15 artist-track pairs in a 30/90/180/365-day window feed d_u ; 10 items not used for profile generation held out per user	10 positives + 1,000 negatives per user, sampled from non-interacted tracks	Llama 3.2; DeepSeek-R1; Gemini 2.0 Flash
Sun et al. (2025)	No held-out split; 8-week MovieLens field deployment (1,775 active users); rating-value/recency partitioning feeds d_u	N/A (recommender pipeline unchanged)	GPT-4o-mini
Tennenholtz et al. (2025)	Per-user chronological: history prefix feeds the encoder, later items held out for training and evaluation	Held-out item set per user	Gemini 1.5 Flash; Gemma V3 (4B/12B)
Torbati et al. (2025)	60/20/20 random split by interaction	1:100 positive-to-negative ratio per test point	GPT-3.5-turbo , Meta-Llama-3.1-8B-Instruct
Zhou et al. (2026)	Per-user chronological: Profile uses the first 50 movies; 940 evaluation scenarios per steering action	items filtered by target genre or tags	Llama-3.1-8B-Instruct
Öncel et al. (2026)	Strong-generalization user-disjoint (1000–5000 held-out users per dataset)	Full item catalogue per dataset (1,943 / 11,355 / 31,046)	Aggregator: Qwen3-30B-A3B-Instruct-2507; item captioners: Qwen3-VL-32B, VideoLLaMA-3, MusicFlamingo
Tan et al. (2026)	Per-user leave-last-out (target: most recent interaction within a 1-hour window of the final timestamp; users filtered to 20–1,000 interactions)	100 items (1 positive + 99 random negatives)	Profiler: Gemma3-1B-it; embedding: Qwen3-Embedding-0.6B (frozen)
Ngo Huu and Lauw (2026)	The last two interacted item per user for testing	1 positive + 19 negatives per test point	Qwen2.5-7B-Instruct (LLM-as-graph-operator); Gemini 2.5 Flash & DeepSeek-V3 as judges

Table 6: Coverage of intrinsic evaluation axes. The evaluable output is the profile d_u .

Reference	Properties measured	Papers in the survey
History (H_u)	faithfulness, groundedness	(Tennenholtz et al., 2025): Profile–History Consistency rated 1–5 by human raters and an AI judge
User (u)	representativeness, self-identification, perceived relevance	(Ramos et al., 2024): Binary yes/no ratings of fluency, relevance, informativeness, conciseness; (Sguerra et al., 2025): self-identification on Likert scale; (Gao et al., 2025): behavioral interpretability via two-candidate next-item prediction, with decision duration as cognitive-load proxy; (Fabbri et al., 2025): LLM-as-a-Judge calibrated against human judgments (Tennenholtz et al., 2025): human raters predict user preference given only the profile (Prediction Fidelity)
Profile (d_u)	fluency, coherence, conciseness, lexical diversity	(Penaloza et al., 2025): BLEU, pairwise edit distance for cross-user diversity; profile length and conciseness as secondary descriptors; (Torbaty et al., 2025): POS-tag distribution, idf-weight mass distribution, KL divergence between generated and source-text distributions as proxies for writing style and faithfulness
Edited profile (d'_u)	edit conciseness, edit coherence, sustained readability	—

Table 7: Coverage of extrinsic evaluation axes. The evaluable output is the ranking prediction R_k .

Reference	Properties measured	Papers in the survey
Ground truth (H'_u)	accuracy: NDCG, Recall, MRR, MAP, RMSE, MAE	The dominant axis: used by essentially every surveyed paper that reports offline evaluation
User (u)	satisfaction, perceived utility, engagement	(Sun et al., 2025): eight-week longitudinal field experiment measuring engagement, exploration, reflection, satisfaction); (Sanner et al., 2023): preference-elicitation speed as an indirect utility signal, 3–4× faster than item-based elicitation
Profile (d_u)	profile–recommendation alignment	— (largely unexplored as a stand-alone evaluation, despite being a natural sanity check on profile-conditioned ranking)
Edited profile (d'_u)	scrutability, controllability	(Ramos et al., 2024): recommendation changes after simulated profile edits; (Penaloza et al., 2025): rank delta under item boosting; genre-wise NDCG under interest flipping; (Zhou et al., 2026): AUC change on target items under four intervention types; (Öncel et al., 2026): per-attribute recall change under attribute removal and attribute addition interventions
Recommendation (R_k)	diversity, novelty, coverage, fairness	— (rarely reported in NL profile-based work; widely used in mainstream recommender-systems evaluation)