



Universiteit
Leiden

Master Computer Science

ECR-RR: Empathetic Conversational Recommender with Critic-based Reranking and NDCG-Balanced Evaluation

Name: Barbaros ISIK
Student ID: 3905993

Date: 25/08/2025

Specialisation: Data Science

1st supervisor: Ren ZHAOCHUN

2nd supervisor: Hao WANG

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Conversational Recommender Systems (CRS) are recommender systems that use multi-turn natural-language interaction to elicit and model user preferences and to deliver personalised item suggestions, with dialogue serving the recommendation objective [1]. A complete CRS keeps track of context, gathers signals that update a preference model, and grounds replies in specific items while balancing clarity with specificity under realistic compute and data constraints. Within this space, empathetic (emotional) conversational AI refers to systems that can identify a user’s feelings and intentions in context and respond in a way that acknowledges those feelings while still moving the recommendation task forward. This view motivates designs that consider both what to recommend and how to say it.

Building on Empathetic Conversational Recommenders that couple emotion-aware recommendation with emotion-aligned generation [2], and using ReDial as the conversational backbone [3], we present a deployment-ready inference-time reranking method. The generator (Llama-2-7B-Chat) produces a small set of candidate replies from a knowledge augmented prompt; a learned RoBERTa-based critic then scores each candidate along five subjective dimensions (empathy, persuasiveness, logic, informativeness, lifelikeness), while a separate NDCG@K term quantifies recommendation alignment; a normalized composite reward selects one reply [4–6]. Knowledge fields (entities from item linked reviews and concise DBpedia triples) can be included in the prompt to improve specificity and reduce hallucinations, consistent with retrieval-augmented and KG-enhanced approaches [7–9].

To supervise the critic, we construct a scored dataset by merging per-dimension judgments from two open-weight LLMs (Llama-2-7B-Chat and Mistral-7B-Instruct), following the LLM-as-judge protocol [4, 10, 11]. Targets are stored on a 0–9 scale for readability and rescaled to $[0, 1]$ at training time; we train with target normalization and a small Gaussian perturbation to improve stability. Subjective heads are trained independently and NDCG is never supervised into a head, which keeps semantics interpretable and allows alignment to be combined at selection time.

Across multiple runs, inference-time reranking consistently increases the composite subjective reward while preserving recommendation alignment, as measured by internal critic+NDCG metrics and supported by an external LLM-as-judge view. The result is a practical recipe that links conversational quality and relevance through selection rather than retraining, and aligns with findings that knowledge grounding supports empathetic, informative dialogue (e.g., 12).

Contents

1	Introduction	4
1.1	Contributions	5
1.2	Research Questions	6
2	Background	7
2.1	Conversational Recommenders and the ReDial Corpus	7
2.2	Empathy in CRS and the ECR Framework	8
2.3	ECR Framework	9
2.3.1	Overview and Design Objectives	9
2.3.2	Emotion-aware Item Recommendation	9
2.3.3	Emotion-aligned Response Generation	10
2.3.4	Data Enlargement and Evaluation	10
2.4	Empathetic Dialogue Generation	10
2.5	Knowledge Grounding and Retrieval	11
2.6	Inference-time Selection, Critics, and Alternatives	11
2.7	Metrics and Evaluators	11
2.8	Positioning and Novelty	11
3	System Overview	12
3.1	Generator and Knowledge Injection	12
3.2	Critic and Reward	12
3.3	Inference-time Reranking	13
4	Scored Dataset Creation for Critic Training	14
4.1	Sources and Scoring Pipeline	14
4.2	Merging, Normalization, and Targets	14
4.3	Statistics and Splits	15
4.4	Quality Control	15
4.5	Limitations	15
5	Methods	16
5.1	Candidate Generation and Knowledge Prompting	16
5.2	Critic Architecture and Supervised Training	17
5.3	NDCG@K and Composite Selection	18
5.4	Reward Normalization and Balancing	18
5.5	Experimental Test Cases	18
5.6	Evaluation Protocol	19
5.7	Complexity and Efficiency	20
6	Experimental Setup	21
6.1	Dataset	21
6.2	Baselines and Variants	21
6.3	Implementation Notes	21
6.4	System Requirements	21

7 Results	22
7.1 Experimental Objectives	22
7.2 Main Reranking Improvements	23
7.3 Experimental Test Cases	24
7.4 Qualitative Examples	26
7.5 Targeted Critic Sanity Checks (within-context controls)	26
7.6 Recommendation (quick pass)	27
8 Discussion	28
9 Limitations	29
10 Future Work	29
11 Conclusion	30
A Extended Qualitative Examples	34
B Prompt Templates	37
C Additional Experimental Notes	38
D Discussion of Evaluation Metrics	38
D.1 Recommendation metrics	38
D.2 Generation metrics	38
D.3 Supplementary recommendation accuracy results	38

1 Introduction

Conversational Recommender Systems (CRS) connect natural-language dialogue with principled item suggestion. A strong CRS should keep track of the conversation, infer hidden preferences, and give specific, item-based replies in an empathetic and clear way. Prior work highlights a trade-off between recommendation accuracy and conversational quality: models optimised for item accuracy can produce specific but stylistically flat replies, whereas methods targeting empathetic style can be vivid yet vague about items [2, 8, 9, 13].

The Empathetic Conversational Recommender (ECR) framework takes a step toward alignment by coupling emotion-aware recommendation with emotion-aligned response generation and knowledge grounding [2]. At the same time, the ReDial corpus provides a realistic backbone for evaluation: 10k two-party movie recommendation dialogues with explicit like/dislike signals that separate recommendation and conversation quality [3]. Beyond ECR, evidence from empathetic dialogue shows the importance of external knowledge and explicit emotion cues for improving perceived empathy and content quality, for example via knowledge bridging between commonsense and emotional lexicons [12] and multi-resolution, interaction-aware training (EmpDG; 13).

Empathetic conversational AI aims to recognize how a user feels and to respond in a way that acknowledges those feelings while staying helpful and on topic. Before large language models, neural systems added explicit emotion control: ECM embeds emotion categories and uses internal/external memories to keep responses emotionally consistent [14]; MoEL infers an emotion distribution and softly mixes emotion-specific decoders to improve empathy and relevance on EmpatheticDialogues [15, 16]; and CEM brings in commonsense knowledge to better capture both affective and cognitive sides of empathy [17]. With LLMs, prompt-based approaches achieve strong zero-/few-shot results and improve further with semantically similar in-context examples, lightweight interaction, and added knowledge [18]. Recent work also shows benefits from reasoning about emotion causes with chain-of-thought prompts to produce more listener-aware replies [19]. We follow the same goal, clear, kind, and specific answers; but achieve it at inference time by reranking candidates instead of retraining the generator.

This thesis proposes a pragmatic shift: move optimisation from training time to selection time. Instead of single-shot decoding, the generator (Llama-2-7B-Chat; 4) produces multiple candidates which a learned critic scores along five human-centered dimensions (empathy, persuasiveness, logic, informativeness, lifelikeness) and an objective recommendation signal (NDCG@K), fused into a normalized composite reward to select one reply. NDCG provides a bounded, ranking-quality reference that reflects whether the reply aligns with user-preferred items [6]. The critic is built on a RoBERTa backbone [5], trained on a new scored dataset that merges judgments from two open-weight LLMs (Llama-2-7B-Chat [4] and Mistral-7B-Instruct [10]) with target normalization and a small Gaussian perturbation on targets (standard deviation about 0.02 in the normalized space) to stabilize learning. This approach works with any model and runs at inference time, so it avoids heavy retraining. It connects conversational quality to recommendation relevance through selection.

Evaluation mirrors two complementary views. Internally, critic-aligned per-dimension scores and NDCG@K quantify the intended trade-offs of the composite reward on ReDial contexts. Externally, we employ an LLM-as-judge protocol to obtain independent assessments on the

same subjective dimensions, using a fixed scoring prompt taken directly from ECR to ensure baseline comparability and to reduce overfitting to any single internal metric [2, 11]. Together with knowledge-augmented prompting (consistent with findings that knowledge retrieval mitigates hallucinations and supports empathetic phrasing; e.g., 12, and KG-augmented CRS like KGSF and KBRD; 8, 9) this yields a practical recipe for producing one specific, well-reasoned recommendation with an empathetic tone.

1.1 Contributions

We propose a deployable, inference-time reranking pipeline that improves empathetic conversational recommendations without retraining the generator. Specifically, we sample multiple candidate responses from a Llama-2-7B-Chat backbone [4] and select one using a trained critic. The critic produces five human-centred scores (empathy, emotional persuasiveness, logic persuasiveness, informativeness, lifelikeness), which we combine with a recommendation alignment term (NDCG@K) into a single composite reward. Using NDCG as an alignment anchor follows ranking-metric practice and is consistent with dialogue-policy work that shapes rewards with NDCG@K [20]. This selection-time optimisation operationalises the ECR intuition [2] while staying compatible with knowledge-augmented prompts inspired by KBRD/KGSF and RAG [7–9].

We design, normalize, and tune a composite reward that explicitly balances subjective dimensions with recommendation alignment. Subjective targets are trained in the $[0, 1]$ range with a sigmoid head; the NDCG@K term is inherently bounded [6]. This gives a transparent trade-off surface that we probe through test cases varying the number of candidates, length bounds, and knowledge prompts (*inc/exc*), demonstrating that selection can raise empathy/-logic/informativeness while preserving NDCG.

To supervise the critic, we construct a new dataset

(see <https://github.com/barbarosisik/ECR-RR>) by merging per-dimension judgments from two open-weight LLMs (Llama-2-7B-Chat and Mistral-7B-Instruct) [4, 10]. Scores are stored on a 0–9 scale for readability and rescaled to $[0, 1]$ at training time; we add a small Gaussian perturbation to the targets (standard deviation about 0.02 in the normalized space) and train a RoBERTa-based multi-head regressor [5] that uses mean-pooling, a small projection block (dropout–linear–GELU–dropout), and sigmoid heads. This stabilized critic generalizes well enough to drive consistent selection gains on ReDial; we document corpus schema, merging, and basic statistics to support reproducibility.

We build an evaluation wrapper that avoids per-forward device transfers and caches tokenization, enabling practical large- N candidate scoring. We provide dual-view evaluation: internal critic+NDCG metrics and an external LLM-as-judge protocol on 1,000 samples [11], plus qualitative side-by-side examples that make the improvements legible. Together, these components constitute a reproducible pipeline with logged artifacts and scripts.

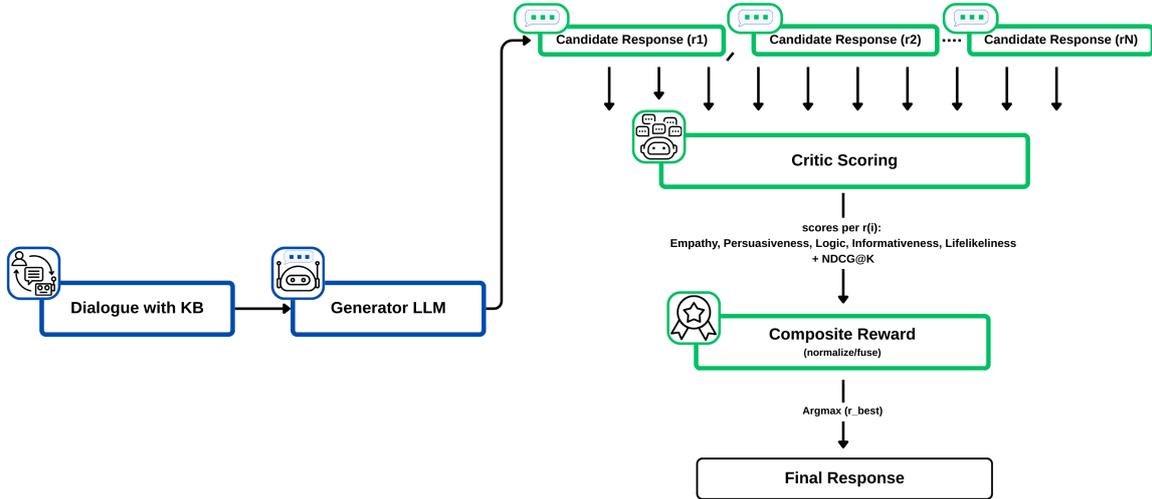


Figure 1: High-level selection pipeline (Figure 2 expanded). Blue boxes denote ECR’s existing flow (dialogue/knowledge → generator); green boxes denote our reranking modules (candidates, critic scoring, composite reward). The generator proposes multiple candidates; a critic scores each candidate per subjective dimension and with NDCG; the composite reward selects the highest-reward candidate.

1.2 Research Questions

We formulate three research questions to focus the study and connect design choices to measurable outcomes. We select these questions because reranking is the core mechanism we add, recommendation alignment is a required constraint for CRS, and the listed design choices are the main levers we can control at inference time without retraining.

- **RQ1:** Does inference-time, critic-based reranking improve subjective response quality (empathy, persuasiveness, logic, informativeness, lifelikeness) over the baseline generator on ReDial?
 We ask this because our method is valuable only if it improves human-centred quality while prompts and decoding are kept fixed. If reranking does not improve these dimensions, the added complexity is not justified. This directly aligns with ECR’s aim to optimize human-centred dialogue quality [2].
- **RQ2:** Does including a recommendation-alignment term (NDCG@K) in the composite reward preserve or improve recommendation sense relative to the baseline?
 We include this test because a selector that only optimizes style could drift from item relevance. For a CRS, preserving item alignment is essential. NDCG@K provides a bounded, interpretable ranking anchor, so we check that adding it keeps or improves alignment relative to the baseline [6].
- **RQ3:** How do key design choices (knowledge-augmented prompts, the number of candidates N , decoding settings, and reward weights) affect reranking gains?
 We study these factors because they shape coverage and quality at inference time. Knowledge prompts can raise specificity and correctness. The number of candidates sets how much we search. Decoding settings balance variety and clarity. Reward weights control the trade-off between subjective quality and alignment. Understanding sensitivities helps

choose strong defaults within realistic compute budgets. Prior CRS work suggests that explicit knowledge and better planning increase specificity and consistency [7–9, 21, 22].

How we answer RQ1 (subjective quality). We adopt a controlled, within-context design to isolate the effect of selection on observed response quality. For each dialogue context, the generator is run twice under matched prompts and decoding: once to obtain a single baseline response and once to sample a small set of candidates (Best-of- N) for our approach. The learned critic then assigns five subjective scores to each candidate, and the selector returns the highest composite reward. We vary two points known to affect textual quality and diversity, candidate count ($N \in \{8, 16\}$) and output length bounds (48/16 vs 80/32 new tokens), and we toggle knowledge prompts (*inc/exc*; see Table 2) to test grounding effects. Evaluation combines (i) internal critic scores aggregated per dimension and in composite, and (ii) an external LLM-as-judge view on a fixed subset to provide an independent assessment of the same five dimensions. We summarise outcomes as paired deltas (reranked minus baseline) per context and report means, confidence intervals, and win rates. We consider RQ1 supported when the composite reward improves and at least three dimensions show positive, non-overlapping confidence intervals relative to baseline.

How we answer RQ2 (recommendation alignment). We test whether adding an NDCG@ K term to the selector preserves item-alignment while improving subjective quality. Using the same contexts, prompts, and decoding as in RQ1, we sweep the NDCG weight α_{NDCG} in the composite reward and compute NDCG@50 per response by mapping mentioned items against the user’s relevant set. NDCG@50 serves as the primary alignment signal; classic recommendation references (AUC, R@10/50, RT@10/50) are monitored as secondary checks. We deem alignment preserved when reranked NDCG@50 is not lower than the baseline by more than a small absolute threshold (e.g., ≤ 0.005) while the subjective composite increases.

How we answer RQ3 (design levers). We characterise sensitivity to key design choices via small factorial ablations over N (8, 16), output length bounds (48/16, 80/32), knowledge prompts (*inc/exc*), and reward weights α (with emphasis on α_{NDCG}). For each configuration we hold all other factors fixed and evaluate on the same context set, enabling paired comparisons. We analyse (i) effect sizes on the composite and per-dimension scores, and (ii) Pareto behaviour between the subjective composite and NDCG@50 to identify trade-offs. We then recommend default settings (e.g., balanced α and moderate N) in regions where subjective gains are consistently high and NDCG remains stable with low variance.

2 Background

2.1 Conversational Recommenders and the ReDial Corpus

Conversational Recommender Systems (CRS) aim to elicit user preferences and deliver item suggestions through multi-turn dialogue, blending goal-oriented recommendation with open-ended language. ReDial was created to support research on this setting with a realistic, human-to-human corpus and with structured annotations that make principled evaluation possible [3].

ReDial contains over 10,000 two-party conversations in which a seeker requests movie suggestions and a recommender proposes titles. Conversations were collected on Amazon Mechanical Turk with assigned roles and simple quality guidance (formal language, roughly ten messages, at least four distinct movies) [3, §3]. Each movie mention is tagged using an @ identifier linked to DBpedia [23], which allows precise tracking of which titles are discussed and when. For every mentioned movie, workers also fill a short form: who mentioned it (seeker or recommender), whether the seeker has seen it (seen, not seen, did not say), and whether the seeker liked it (liked, did not like, did not say). These "movie dialogue forms" provide explicit like/dislike and seen/not-seen signals that can supervise sentiment analysis and preference modelling in downstream systems.

This design serves two purposes [3]. First, it supplies a realistic dialogue substrate where recommendation is an explicit goal but small talk and questions naturally occur. Second, it enables controlled experiments on sub-components. On the recommendation side, the movie-level forms yield labels for building and assessing recommenders (e.g., collaborative filtering and autoencoder-based models), including cold-start protocols that treat each conversation as a new user. On the conversational side, the turn-level text is used to train encoders and decoders, and to analyse how language expresses sentiment and preference. Compared to synthetic or forum corpora, ReDial's role structure and movie tagging make it suitable for evaluating both item alignment and conversational quality within a single setting. In this thesis, ReDial is our conversational backbone for generation and evaluation. Its movie annotations allow us to compute recommendation-aware metrics such as NDCG@K, while its dialogues provide the context for assessing subjective qualities of responses. This dual use aligns with our objective: to couple empathetic phrasing with specific item alignment within realistic conversations. It also ensures comparability to prior CRS work and to our baseline comparison with ECR [2].

2.2 Empathy in CRS and the ECR Framework

The Empathetic Conversational Recommender (ECR) frames empathy as a system's ability to recognize a user's feelings and intentions in context, and to respond in a way that both respects those emotions and advances the recommendation task [2]. In ECR, empathy is not a single score. It is operationalised through two coordinated modules that touch both what to recommend and how to say it.

First, ECR's emotion-aware item recommendation uses signals about the user's expressed attitude toward entities and items in the dialogue. These signals are transformed into emotion representations that adjust entity features, so that the same movie described as "loved" or as "not my type" results in different preference inferences. ECR further improves its strength by adjusting supervision to down-weight ambiguous or negative feedback, which reduces the impact of noisy labels in conversational data. The effect is to align the candidate item set with what the user is likely to accept in the current conversation.

Second, ECR's emotion-aligned response generation aims to produce replies that are consistent with the chosen item and that speak in a human-like way. To improve specificity, ECR extends the prompt with concise knowledge fields, such as DBpedia triples and entities from reviews, which helps reduce hallucination and promotes concrete justification. To show an empathetic tone, ECR uses supervision that emphasises subjective qualities such as emotional intensity, persuasiveness, informativeness and lifelikeness. Together, these design choices increase the chance that a reply will acknowledge the user's stance, present

reasons that fit the user’s request, and maintain natural language flow [2] (for a practitioner summary see 24).

Our work adopts the core view of empathy as a combined property of item selection and wording, while maintaining ECR’s knowledge-augmented prompting. We differ with the operational modifications. Rather than retraining the generator, we generate a small set of candidates, score them on the same subjective dimensions, and select one at inference time. This keeps the generator unchanged while still enforcing the empathetic criteria that ECR promotes. Shortly, we retain ECR’s notion of empathy and its emphasis on knowledge grounding, and we implement the optimization at selection time instead of training time.

2.3 ECR Framework

We group ECR components in this subsection and keep subtopics separate for clarity.

2.3.1 Overview and Design Objectives

ECR is a system blueprint that links what is recommended to how it is said. The framework sets three design objectives [2]: (i) represent user preferences in a form that is responsive to the emotions expressed in the dialogue, so that recommendations reflect the seeker’s position; (ii) generate responses that are precise, based on verifiable knowledge, and written in a considerate, human-like way; and (iii) evaluate with a dual lens that considers both item alignment and subjective dialogue quality. To pursue these goals, ECR consists of two connected modules and a supporting data enlargement process. First, *emotion-aware item recommendation* improves preference modelling by combining emotions with entity features taken from the current dialogue and from global statistics about which items appear together. Second, *emotion-aligned response generation* uses retrieved knowledge and item signals to guide the generator to produce persuasive, human-like utterances that express the right emotion. We expand the data to add missing supervision by labeling emotions at the utterance level and by collecting emotional reviews to fine-tune the generator. Evaluation then combines recommendation metrics with subjective dialogue quality [2].

2.3.2 Emotion-aware Item Recommendation

Local emotion-aware entities. For entities mentioned in the dialogue, ECR computes an utterance-level emotion representation by weighting emotion embeddings with LLM-derived probabilities and concatenating this to each local entity vector. This captures that the same entity (for example, a movie or an actor) can carry different affect depending on the user’s view in context [2].

Global emotion-aware entities. Beyond the current dialogue, ECR filters global entities from the training data by emotion co-occurrence. Entities that tend to co-appear under similar emotions are aggregated with learned or empirical co-occurrence weights, producing a global emotion-aware representation that complements local signals [2]. This idea connects to KG-enhanced CRS methods such as KBRD and KGSF, which use entity structure to improve user representations [8, 9].

Feedback-aware item reweighting. To mitigate noisy supervision, ECR reweights the recommendation loss by user feedback labels (like, dislike, did not say), reducing gradient contribution from ambiguous or negative feedback while retaining informative signals. These steps improve AUC and Recall, and they better reflect user affect in preference modelling [2].

2.3.3 Emotion-aligned Response Generation

Retrieval-augmented prompt. ECR retrieves knowledge triples (for example from DBpedia) and entities from emotional reviews associated with the candidate item, converts them to short text fields, and composes a generation prompt that includes the recommendation response and the item name. This reduces hallucination and increases specificity by providing stable facts to condition on [2, 7, 22].

Emotion-aligned generator. A pre-trained language model is fine-tuned on emotional reviews so that outputs adopt a vivid, persuasive tone while remaining grounded in the chosen item. At inference, the final reply combines the recommendation response and the emotion aligned continuation. Knowledge prompts can be ablated to study their effect on factuality and narrative quality [2]. This approach is consistent with retrieval-augmented generation more broadly, where non-parametric memory is coupled with a generator to increase specificity [7].

2.3.4 Data Enlargement and Evaluation

Empathetic data. Because CRS corpora lack dense emotion supervision, ECR labels utterances with emotions using an LLM and constructs an emotional review database (for example from IMDb) filtered for quality and positivity. Entity extraction connects review sentences to the knowledge graph so that prompts can include concise triples and entities [2].

Evaluation. ECR evaluates recommendation with AUC and Recall@ n , and evaluates generation on five subjective dimensions: emotional intensity, emotional persuasiveness, logic persuasiveness, informativeness, and lifelikeness (0 to 9 scale). This dual perspective seeks to capture user satisfaction beyond string-overlap metrics [2]. In our study we adopt the same subjective dimensions for internal scoring and we cross-check with an external LLM-as-judge protocol [11].

2.4 Empathetic Dialogue Generation

Work on empathetic dialogue highlights the role of affect and narrative structure. EmpDG uses a multi-resolution, interactive training objective to capture both broad and detailed emotional cues, which improves empathy and content quality in open-domain settings [13]. The Empathetic Dialogues corpus supports modelling emotions explicitly and includes tasks where correctly recognising and expressing emotions are important for user satisfaction [16]. These results support evaluating and optimizing for human-centred dimensions beyond topical relevance, which aligns with ECR’s decomposition and with our selection-time scoring.

Before large language models, early neural work added explicit emotion control. ECM embeds emotion categories, tracks an internal emotion state, and uses an external emotion vocabulary to make outputs emotionally consistent [14]. MoEL estimates the user’s emotion distribution and blends decoders for each emotion, leading to better empathy and relevance on EmpatheticDialogues [15, 16]. Commonsense-aware methods such as CEM add ATOMIC/COMET inferences to couple affective and cognitive empathy, which increases informativeness without losing appropriateness [17]. With LLMs, prompt-based approaches show strong zero-/few-shot results and benefit from semantically similar in-context examples, lightweight interaction, and knowledge augmentation [18]. Recent work adds emotion-cause reasoning with chain-of-thought prompting to guide models toward cause-aware, listener-aware

responses [19]. We follow this line in spirit but avoid retraining the generator; instead, we improve empathy and specificity at selection time by reranking with a learned critic.

2.5 Knowledge Grounding and Retrieval

Grounding responses in external knowledge improves specificity and factuality. KBRD links informative entities from knowledge graphs to enrich user representations and bias generation toward consistent items [8]. KGSF fuses ConceptNet (commonsense) and DBpedia (entity-oriented) signals via mutual-information alignment to bridge the semantic gap between words and items [9, 21, 22]. Retrieval-Augmented Generation (RAG) shows that coupling pretrained LMs with non-parametric memory produces more specific outputs in knowledge-intensive tasks [7]. Our prompting mirrors these ideas: we surface related entities and knowledge to guide the generator toward a single specific movie with reasons. In KBRD, entity signals from DBpedia are propagated with relation-aware graph convolutions (R-GCN), then fed back to the generator as a recommendation-aware vocabulary bias that nudges wording toward user-relevant items and attributes; jointly, this improves Recall@K and dialogue diversity/consistency [8]. RAG complements KG-guided prompting by pairing a dense retriever with a seq2seq generator to condition on retrieved passages; across knowledge-intensive tasks, this reduces hallucination and increases specificity compared to parametric-only generation [7]. These patterns also connect to graph-based reasoning used to align representations [25].

2.6 Inference-time Selection, Critics, and Alternatives

Rather than optimising the generator with reinforcement learning (PPO; RLHF), we adopt a Best-of- N selection strategy: sample multiple candidates and pick one using a learned critic. This follows the broader pattern of training reward/critic models on human or proxy judgments [26] while avoiding instability and compute overhead at policy-training time [27]. Our critic uses a RoBERTa backbone [5] and is supervised on a new, merged scored dataset that combines judgments from Llama-2-7B-Chat and Mistral-7B-Instruct [4, 10], aligning with the growing practice of LLM-as-judge for scalable evaluation and supervision [11].

2.7 Metrics and Evaluators

We use two complementary lenses. First, recommendation-aware ranking via NDCG@K provides a bounded metric of how well a reply aligns with user-preferred items [6]. Second, subjective quality is assessed through per-dimension scores (empathy, emotional persuasiveness, logic persuasiveness, informativeness, lifelikeness) from the learned critic and is cross-checked with an external LLM-as-judge protocol [11, 28]. Defining subjective dimensions explicitly makes evaluation more interpretable and connects to prior findings that persuasive and informative language affects user outcomes [13, 29].

2.8 Positioning and Novelty

Compared to prior work, our contribution lies in shifting improvement to inference-time selection with a composite reward that explicitly ties human-centred dimensions to recommendation relevance (via NDCG@K), creating a merged scored dataset from two open-weight LLMs to supervise a discriminative RoBERTa critic, and engineering a complete pipeline

with knowledge prompts. This complements ECR’s training-time design by offering a deployable selector that improves responses without retraining the generator [2, 4, 10]. Our evaluation connects internal critic signals with an external LLM-as-judge view [11], and our use of NDCG@K anchors recommendation alignment in a bounded ranking metric [6].

3 System Overview

This section summarises how the modules combine and how the following subsections relate. Given a dialogue context, we build a knowledge-augmented prompt for the generator, produce a small set of candidate responses under fixed decoding settings, score each candidate with a learned critic on five subjective dimensions together with NDCG@K for item alignment, and select the candidate with the highest composite reward. We first describe Generator and Knowledge Injection, then Critic and Reward, followed by Inference-time Reranking, and finally Metrics and Evaluators. Figure 2 shows the high-level structure.

3.1 Generator and Knowledge Injection

We use a Llama-2-7B-Chat generator [4] with knowledge-augmented prompts. For each context, we construct two fields (“Related Entities” and “Related Knowledge from KG”) by retrieving short triples from DBpedia and adding entities extracted from IMDb reviews associated with the recommended item. This mirrors the spirit of ECR’s retrieval-augmented prompts [2] and aligns with KBRD’s entity guidance [8]. Earlier work such as DialoGPT [30] demonstrated strong single-turn dialogue capabilities; our use of Llama-2 builds on these transformer foundations while benefiting from instruction-tuned chat behaviour. The approach is consistent with RAG [7], increasing specificity and reducing hallucinations while leaving the generator unchanged. A high-level view of the overall pipeline (generator → knowledge retrieval → critic scoring → selection) is shown in Figure 2.

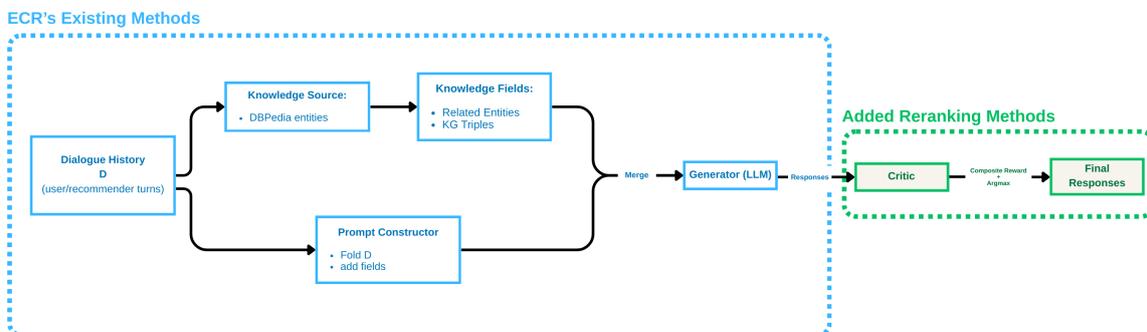


Figure 2: System architecture with clustered components. Light blue cluster (ECR framework): Dialogue history D , knowledge source (DBpedia entities), knowledge fields (Related Entities, KG Triples), prompt constructor (fold D , add fields, merge), generator LLM, and responses. Light green cluster (Added reranking methods): critic, composite reward + argmax, and final responses. Color semantics match Figure 1.

3.2 Critic and Reward

The critic is a multi-head RoBERTa regressor [5] that predicts five subjective dimensions in $[0, 1]$: empathy, emotional persuasiveness, logic persuasiveness, informativeness, and life-

likeness. Inputs mirror inference usage by concatenating dialogue context and a candidate response with lightweight tags, e.g., “<context> . . . </context><response> . . . </response>”. Token-level representations are mean-pooled under the attention mask, passed through a projection block (dropout–linear–GELU–dropout), and mapped to five scores with a sigmoid head. Targets originate from a merged, dual-LLM scoring corpus (Llama-2-7B-Chat and Mistral-7B-Instruct) scaled to $[0, 1]$; training uses MSE loss with light label jitter ($\pm \approx 0.02$) and warmup+linear decay with gradient clipping, which together stabilize optimization and improve calibration. We place this design within the transformer encoder line [31, 32].

Recommendation alignment is measured separately with NDCG@K [6], using the item list mentioned in the response and the user’s preference set for that turn. We treat relevant items as binary gains and compute NDCG over the top- K (default $K = 10$), resulting in $s_{\text{NDCG}} \in [0, 1]$. This separation preserves the semantics of the critic heads (purely subjective) while letting the selector combine subjective and recommendation signals at decision time.

The selector forms a composite reward as a convex combination of normalized scores:

$$R = \alpha_{\text{Emp}} s_{\text{Emp}} + \alpha_{\text{Per}} s_{\text{Per}} + \alpha_{\text{Log}} s_{\text{Log}} + \alpha_{\text{Inf}} s_{\text{Inf}} + \alpha_{\text{Life}} s_{\text{Life}} + \alpha_{\text{NDCG}} s_{\text{NDCG}}.$$

Default weights emphasize empathy and recommendation alignment while maintaining balance across other dimensions (e.g., $\alpha_{\text{Emp}} \approx 0.25$, $\alpha_{\text{NDCG}} \approx 0.25$, $\alpha_{\text{Per}} \approx 0.15$, $\alpha_{\text{Log}} \approx 0.15$, $\alpha_{\text{Inf}} \approx 0.10$, $\alpha_{\text{Life}} \approx 0.10$). We test alternatives with small grids to identify stable regions where subjective quality improves without lowering NDCG. In practice, the above normalization and weighting produce a smooth trade-off surface: higher empathy/logic/informativeness usually co-move under knowledge prompts, and NDCG stays stable.

Implementation notes. The critic applies mean-pooling to handle differences in response length, uses dropout in the projection block (≈ 0.35 during training), a warmup ratio (≈ 0.06), and gradient clipping. Targets are trained on five heads only; NDCG is never supervised into a head but computed on demand from the candidate’s item list and the user’s relevant set. This keeps heads interpretable and avoids leaking recommendation signals into subjective dimensions.

3.3 Inference-time Reranking

As shown by the green cluster in Figure 1, Best-of- N selection samples N candidates from the generator, scores each with the RoBERTa critic (five subjective dimensions) and NDCG@K, forms a composite reward, and returns the argmax as the final response.

Given a dialogue context, the generator produces N candidates under controllable decoding (temperature/top-p, beams, and min/max new tokens). For each candidate, we build the same tagged input used in critic training and obtain the five subjective scores. In parallel, we compute NDCG@K from the candidate’s item list against the user’s relevant set. The selector aggregates these six signals with the weights α above and returns the highest-reward candidate.

Practicalities. We score all N candidates in a batch to minimize overhead; inputs match the critic’s training format to reduce distribution shift. Knowledge prompts (“Related Entities” and “Related Knowledge from KG”) can be toggled on/off to test grounding effects; Longer length

limits provide more space for narration and generally lead to higher informativeness scores. Because all scores are in $[0, 1]$, the composite reward is well-behaved across settings and supports direct comparisons within a run.

4 Scored Dataset Creation for Critic Training

4.1 Sources and Scoring Pipeline

We construct a supervision corpus by scoring conversation-response pairs with two open-weight LLMs (Llama-2-7B-Chat [4] and Mistral-7B-Instruct [10]). The goal is a diversely scored dataset with reduced single-judge bias: using two independent judges increases variance where the task is ambiguous and reduces the influence of any one model’s peculiarities.

Specifically:

- We prepare input records containing a short dialogue context (from ReDial) and one candidate response (from the generator or curated variations). Inputs are formatted consistently for both judges.
- Each judge produces five subjective scores on a 0–1 scale: empathy, (emotional) persuasiveness, logic persuasiveness, informativeness, and lifelikeness. We also keep a broad overall/coherence score to support mapping consistency.
- For traceability, we store per-judge scores in separate shards (six shards per model), with stable keys (e.g., `empathy_score`, `informativeness_score`, `recommendation_score` as a proxy for persuasiveness, `engagement_score` for lifelikeness, `overall_score` for logic/coherence).

This pipeline follows the LLM-as-judge approach [11], but uses two judges to improve reliability and reduce bias. All scoring prompts are held fixed across shards to avoid prompt drift; contexts and candidates are shuffled per shard to diversify batches.

4.2 Merging, Normalization, and Targets

We merge dual-judge scores per example by schema-aligning keys and averaging per dimension after clipping extreme outliers (component-wise within $[0, 1]$). Where only one judge is available, we retain that record if all required keys are present. The merged targets are then stored in the artifacts on a 0–9 scale (for human readability); during training we rescale to $[0, 1]$ and supervise the critic’s five sigmoid heads with MSE loss.

Dimension mapping is:

- `empathy` ← `empathy_score`
- `persuasiveness` ← `recommendation_score` (emotional/pitch quality proxy)
- `logic persuasiveness` ← `overall_score` (coherence/justification proxy)
- `informativeness` ← `informativeness_score`
- `lifelikeness` ← `engagement_score`

We explicitly exclude NDCG from supervised targets to keep subjective heads semantically clean; recommendation alignment is computed at inference time and combined in the selector (Section § 3.2). To stabilize training we add a small Gaussian perturbation to the $[0, 1]$ targets (standard deviation about 0.02) and use dropout in the projection block [33, 34]. For evaluation we pair internal metrics with LLM-as-judge protocols [11, 28], reflecting growing practice in open-ended assessment.

4.3 Statistics and Splits

The merged corpus contains 16,716 samples (2,786 per part \times 6 parts). We split by shard (train: five shards, validation: one shard) to preserve distributional balance across judges and topics. Context length and response length distributions reflect ReDial’s conversational style (short, mixed-register turns); coverage spans humorous/neutral/empathetic dialogues and a broad item set. Each dimension exhibits healthy variance (without saturation), which helps the critic learn discriminative signals rather than collapse to means.

4.4 Quality Control

We apply safeguards to keep supervision and evaluation reliable under constrained resources.

- **Schema and format checks:** enforce presence of required keys, validate types and ranges, and drop records with encoding artifacts or truncated Unicode.
- **Length controls:** cap context and response lengths to prevent extreme outliers, and remove near-duplicates by hashing a normalized text signature.
- **Reproducible training:** log configuration, random seeds, checkpoints, and evaluation summaries; store hashes of input shards to detect silent data changes.
- **Tokenisation sanity:** audit a small random set to verify that the tagged input format (`<context> . . . </context><response> . . . </response>`) is preserved and that truncation respects the attention mask.

When judge scores disagree strongly, averaging preserves uncertainty while keeping samples usable; if an entire shard exhibits drift or schema issues, we exclude it. These steps reduce obvious noise and leakage but do not replace human adjudication.

4.5 Limitations

Label source and calibration. LLM-as-judge labels can inherit model biases and approximate calibration. Merging two sources reduces but does not eliminate systematic effects; human adjudication would strengthen validity [11].

Prompt sensitivity. Scores are contextual and can vary with phrasing and length limits. We hold prompts fixed per judge, but residual sensitivity may remain.

Subjective focus and alignment. The corpus supervises five subjective dimensions. It does not include per-turn recommendation ground truth. Recommendation alignment is therefore handled separately via NDCG@K during selection [6]. This separation preserves head semantics but means subjective improvements and item alignment can diverge on some cases.

Coverage and distribution. Topic and style coverage is limited by the underlying dialogues. Some genres or conversational intents may be under-represented, which can skew targets and learned calibration.

Inter-judge disagreement. Simple averaging preserves uncertainty but can mask bimodal judgments. Strong aggregation or adjudication would better handle high-disagreement samples.

Truncation effects. Fixed input budgets can truncate long contexts and responses. We mask properly, but truncation may still alter the signals available to judges and the critic.

Knowledge and linking noise. DBpedia linking and review entity extraction can be noisy [22]. This affects informativeness targets when facts are weakly connected. Better coverage or learned retrievers could help [7].

5 Methods

This section explains how we construct knowledge-augmented prompts and generate candidates, how the critic is trained and used, how recommendation alignment is computed and combined with subjective scores, and how we evaluate the system. We first describe Candidate Generation and Knowledge Prompting, then Critic Architecture and Supervised Training, followed by NDCG@K and Composite Selection and Reward Normalization and Balancing, and finally our experimental design and evaluation protocol. The goal is to make clear where each signal comes from and how components interact at selection time.

5.1 Candidate Generation and Knowledge Prompting

We generate multiple response candidates per dialogue context using a Llama 2–7B Chat generator with lightweight, retrieval-augmented prompts (Section § 3.1). Each prompt optionally includes two short fields: *Related Entities* (named entities drawn from item-linked reviews) and *Related Knowledge from KG* (concise DBpedia triples). This mirrors KBRD/KGSF-style guidance that improves specificity while lowering the possibility of hallucinations [7–9]. For a given context, we fix the prompt template in a run and decode N candidates under matched settings so selection makes differences.

Prompt template. The prompt contains (i) the dialogue context with speaker turns, (ii) a brief instruction to give one specific recommendation reply, and (iii) two optional knowledge fields. *Related Entities* lists salient entities mined from item-linked reviews, and *Related Knowledge from KG* lists short DBpedia triples for the same item. We keep wording stable within a run so differences in outputs arise from decoding randomness and selection, not from template drift.

Knowledge construction. We extract entities from review text associated with candidate items and filter to a short list. We retrieve DBpedia triples for the same items and prune to concise subject–predicate–object facts. These fields are optional and can be toggled to test grounding effects. When included, they provide concrete hooks that help the generator produce specific and verifiable content, consistent with retrieval-augmented generation [7] and KG-enhanced CRS [8, 9].

Decoding configuration. Unless otherwise given, decoding uses moderately diverse sampling with `max_new_tokens` ≈ 64 , `temperature` ≈ 0.7 , `top-p` ≈ 0.9 , `repetition_penalty` ≈ 1.2 ,

and a no-repeat-ngram constraint (size 3). Contexts are tokenized with left padding and truncated to a fixed budget (~512 tokens) to maintain consistent conditioning. These choices show standard practice for chat models and match our evaluation scripts.

5.2 Critic Architecture and Supervised Training

The critic is a multi-head RoBERTa encoder that predicts five subjective dimensions in $[0, 1]$: empathy, emotional persuasiveness, logic persuasiveness, informativeness, and lifelikeness. The input shows inference usage by concatenating the dialogue context and a candidate response with lightweight tags: `<context> . . . </context><response> . . . </response>`. We tokenize to a fixed budget and mean-pool token representations under the attention mask to stabilize training under variable lengths.

Input formatting and truncation. We wrap context and response with tags so the encoder trains on the same structure used at selection time. We cap sequence length for stability. Mean pooling over the attention mask makes the model strong to length variation without relying on a special pooling token.

Model. We start from a local roberta-base snapshot and add a small projection block and a linear head per [31, 32]: dropout \rightarrow linear \rightarrow GELU \rightarrow dropout, followed by a sigmoid head producing five normalized scores. Specifically, hidden states are mean-pooled, passed through the projection block (dropout rate ≈ 0.35 during training), and mapped to 5 outputs with a sigmoid activation.

Targets and datasets. Supervision comes from a dual-judge merge of Llama 2 7B Chat and Mistral 7B Instruct scorers (LLM-as-judge) [11] as also to match the subjective LLM evaluation method in ECR [2]. Per example we map scorer keys to five dimensions, keeping the semantics consistent across shards: empathy \leftarrow empathy_score; emotional persuasiveness \leftarrow recommendation_score (persuasion proxy); logic persuasiveness \leftarrow overall_score (coherence proxy); informativeness \leftarrow informativeness_score; lifelikeness \leftarrow engagement_score. Scores are stored on a 0–9 scale for readability and rescaled to $[0, 1]$ at training time.

Objective and optimization. We train with mean-squared error (MSE) on the five heads only. To regularize, we add a small Gaussian perturbation to the targets (standard deviation about 0.02 in the normalized space) and apply gradient clipping (ℓ_2 norm ≈ 1.0). Optimization uses AdamW (learning rate $\approx 3 \times 10^{-5}$, weight decay ≈ 0.01) with linear warmup and decay; the warmup ratio is ≈ 0.06 . Typical hyperparameters: epochs ≈ 10 , batch size ≈ 8 , max input length ≈ 384 tokens for training. We select the checkpoint with the lowest validation loss and use it for inference.

Inference details. At selection time, the critic is evaluated in eval mode on the active device. Inputs use the same tagged format, with a slightly tighter max length (~256 tokens) for throughput. Outputs are five normalized scores; no calibration post-processing is applied beyond the sigmoid.

Stability notes. Mean pooling and the projection block with dropout improve stability across variable candidate lengths. Target jitter and warmup reduce sensitivity to small label noise and stabilize early training. We monitor validation loss each epoch and retain the best checkpoint for selection.

5.3 NDCG@K and Composite Selection

Recommendation alignment is computed separately via $\text{NDCG}@K$ [6]. Given a ranked list of items referenced (or implied) by the response and a user preference set for the turn, we treat relevant items as binary gains and compute DCG with \log_2 discount; IDCG is the DCG of an ideal list of length $\min(|\text{prefs}|, K)$. We use $\text{NDCG}@10$ in the selector by default, and report $\text{NDCG}@50$ in aggregate tables for alignment diagnostics.

The selector forms a convex combination of six signals (five subjective, one NDCG):

$$R = \alpha_{\text{Emp}}S_{\text{Emp}} + \alpha_{\text{Per}}S_{\text{Per}} + \alpha_{\text{Log}}S_{\text{Log}} + \alpha_{\text{Inf}}S_{\text{Inf}} + \alpha_{\text{Life}}S_{\text{Life}} + \alpha_{\text{NDCG}}S_{\text{NDCG}},$$

with default weights emphasizing empathy and alignment ($\alpha_{\text{Emp}} \approx 0.25$, $\alpha_{\text{NDCG}} \approx 0.25$, $\alpha_{\text{Per}} \approx 0.15$, $\alpha_{\text{Log}} \approx 0.15$, $\alpha_{\text{Inf}} \approx 0.10$, $\alpha_{\text{Life}} \approx 0.10$). We score all N candidates in a batch and return the argmax .

Candidate item lists. For computing NDCG in controlled tests, we construct a candidate pool per context by combining the user-preferred items with randomly sampled negatives from the item vocabulary, then evaluate $\text{NDCG}@K$ on the ranked list induced by the response. This isolates the effect of wording on alignment and matches our evaluation scripts.

5.4 Reward Normalization and Balancing

We normalize per-dimension outputs to a comparable range before aggregation. Subjective targets are trained in $[0, 1]$ with a sigmoid head and we keep the same range at inference. The recommendation component $\text{NDCG}@K$ is bounded in $[0, 1]$ by definition [6]. We create a convex combination with weights α that balance empathy, emotional persuasion, logic persuasion, informativeness, lifelikeness, and recommendation alignment. We choose α by small grid tests and prefer settings that raise the subjective composite without degrading $\text{NDCG}@K$. The weights add up to one, and no further rescaling is applied at decision time.

5.5 Experimental Test Cases

We designed a small set of controlled cases to study how selection behaves and which settings matter most. The goal is to vary one factor at a time, hold the others fixed, and compare the baseline generator to our Best-of- N selector on the same dialogue contexts. We keep prompts, tokenisation, and seeding stable within a run.

Settings we varied.

- **Generation:** number of candidates $N \in \{8, 16\}$; maximum and minimum new tokens; sampling diversity (temperature and top-p). Prompts are fixed within a run so differences come from selection, not from prompt drift.
- **Selection:** reward weights α with focus on α_{NDCG} . We sweep α_{NDCG} in a narrow range to test whether higher alignment pressure changes the trade off with subjective quality.
- **Critic training:** learning rate, batch size, dropout [34], and warmup ratio. These are tuned once to a stable point and not changed across main runs.

- **Knowledge prompts:** on or off for the two short fields (Related Entities and Related Knowledge from KG). This tests the effect of knowledge grounding on informativeness and persuasion.

Case matrix. For each context set we evaluate four primary cases:

- **inc, $N = 8$, 48/16:** knowledge included; eight candidates; shorter output bounds (max/min new tokens 48/16).
- **inc, $N = 16$, 80/32:** knowledge included; sixteen candidates; longer bounds (80/32).
- **exc, $N = 8$, 48/16:** knowledge excluded; eight candidates; shorter bounds.
- **exc, $N = 16$, 80/32:** knowledge excluded; sixteen candidates; longer bounds.

These cases let us study the effect of candidate count, length, and knowledge grounding with paired comparisons. We add a light ablation on α_{NDCG} to confirm that alignment does not degrade when weights move within a reasonable band.

How we compare. For each case we run the baseline generator once to obtain a single response, then run our selector to sample N candidates with the same decoding settings and select one. We compute the five critic dimensions and the composite reward, and we compute NDCG@50 on item mentions for alignment. We report:

- **Paired deltas:** reranked minus baseline for the composite and each dimension, averaged over contexts with confidence intervals.
- **Win rate:** fraction of contexts where the composite reward improves.
- **Alignment check:** average NDCG@50 and the difference relative to baseline.

What we look for. We consider a setting successful when the composite reward increases and NDCG remains within a small absolute change relative to baseline. In practice we observe that knowledge prompts raise informativeness and often persuasion, that longer bounds help when knowledge is present, and that moving from $N = 8$ to $N = 16$ gives a moderate gain by exposing more diverse candidates. The ablation on α_{NDCG} shows a stable region where subjective quality improves and alignment stays close to baseline.

Sanity checks. We also run three within context controls with GOOD and BAD responses that share the same intent. The critic assigns higher totals to GOOD in all three, including a case where the BAD response is longer. This supports that the critic rewards empathy, coherence, and specificity rather than length alone.

5.6 Evaluation Protocol

We prioritise reproducibility and controlled comparisons.

- **Preprocessing:** left padding and truncation to fixed budgets; tagged inputs with context and response; deterministic tokenisation.

- **Determinism:** fixed seeds for decoding and selection; configuration and environment hashes logged with outputs.
- **Batching:** candidates scored in batches to avoid device transfer overhead; local model snapshots used to avoid remote cache writes.
- **Internal metrics:** per-dimension means, composite reward, and NDCG@K [6] on ReDial contexts [3]; report paired deltas and raw pairs.
- **External view:** small scale qualitative checks and sanity tests; we treat these as supportive rather than conclusive.
- **Recommendation references:** AUC, Recall@K, and NDCG@K from the existing pipeline, monitored to confirm that selection does not sacrifice alignment.

5.7 Complexity and Efficiency

The pipeline uses local caches, avoids generator retraining, and logs artifacts per run. Selection adds $O(N)$ encoder passes per context, where N is the number of candidates. We score candidates in batches to keep throughput high and to avoid repeated device transfers. Memory use is dominated by the encoder and attention activations; response length bounds control compute cost directly.

Relative to policy optimisation methods such as PPO or RLHF [26, 27], our cost profile is simple and stable. There is no rollout collection, no reward-model updating loop, and no gradient updates to the generator. The only extra work comes from generating N candidates under fixed decoding and one forward pass through the critic per candidate. In practice, these steps parallelise well and are easy to budget by choosing N and the max new tokens.

We found that modest settings already give most of the gain. Moving from $N = 8$ to $N = 16$ increases the number of critic passes, but the wall time rises near linearly and remains below retraining budgets in comparable RL pipelines. Knowledge prompts improve the chance that one of the N candidates carries concrete details, which lets the selector recover quality without increasing N further. When latency matters, one can keep N small and reduce max new tokens; when offline batch time is available, one can use larger N .

The method is also efficient from an engineering point of view. It reuses a frozen generator and a single critic checkpoint, which simplifies deployment and reduces failure modes. Caching tokenisation and using local model snapshots avoid I/O stalls. For online use, common contexts can be cached, and partial re-scoring is possible if only a subset of candidates change. Together, these choices make the approach practical on modest hardware while still realising gains from a learned critic.

Under constrained resources, the linear $O(N)$ cost is acceptable for offline or batched inference. Online systems can reduce N , shorten outputs, or cache frequent contexts. These levers allow tuning the quality–latency trade off without introducing the instability and compute overhead that often come with optimisation-time methods.

6 Experimental Setup

6.1 Dataset

ReDial movie recommendation dialogues [3]. Contexts are tokenized with left padding and truncated to a fixed budget (≈ 512 tokens) to maintain consistent conditioning; responses are bounded by the min/max new-tokens limits specified per test case. We use the same item vocabulary as in the ECR setting so recommendation references are comparable. For critic training, we use the newly constructed scored dataset (merged Llama-2 + Mistral judgments; 16,716 samples) with targets normalized to $[0, 1]$.

6.2 Baselines and Variants

- **Baseline generator:** Llama-2-7B-Chat without inference-time selection, prompted with knowledge fields in ECR style [2, 4, 8, 9]. Generation uses sampling with moderate diversity (temperature ≈ 0.7 , top-p ≈ 0.9) and the same length bounds as in the reranked runs.
- **Reranked default:** Best-of- N (with $N \in \{8, 16\}$) using the RoBERTa critic trained on merged Llama-2/Mistral judgments [4, 10]. Each candidate is generated under the same prompt and decoding settings as the baseline; selection uses the composite reward that includes NDCG@K.
- **Ablations:** vary N , length bounds, temperature/top-p, and α weights (especially α_{NDCG}); toggle knowledge prompts; and adjust critic training dropout and warmup.

6.3 Implementation Notes

Prompting: Prompts include two optional fields (“Related Entities” and “Related Knowledge from KG”) to surface DBpedia triples [22] and review-derived entities. These fields are toggled per setting to test grounding effects.

Decoding: Sampling with temperature ≈ 0.7 , top-p ≈ 0.9 ; repetition penalty ≈ 1.1 – 1.2 ; max/min new tokens set per test case (e.g., 48/16 or 80/32). Contexts are left-padded; sequence length capped at ≈ 512 tokens for inputs.

Selection: Candidates are scored in a batch; the composite reward uses the default weights from Section § 3.2 unless otherwise stated. Because all components lie in $[0, 1]$, rewards are directly comparable across settings within a run.

Reproducibility: Where applicable, random seeds are fixed for decoding; each variant is evaluated on the same set of dialogue contexts to enable paired comparisons.

6.4 System Requirements

We conducted all experiments on A100 and 2080Ti GPU’s (LIACS’s ALICE).

The table below summarizes which partitions were used for each major task type; GPU model on a given run depended on availability at scheduling time.

Task	Purpose	Typical resources
Critic supervised training (RoBERTa)	Train multi-head critic on merged scored data	1 GPU; 4 CPU; 24–32 GB RAM
Reranking evaluation	Generate N candidates and score/select	1 A100 GPU; 4 CPU; 24 GB RAM
LLM-based subjective scoring (1,000 samples)	External judge scores for comparison	1 A100 GPU; 4 CPU; 32 GB RAM
Recommendation module evaluation	Baseline/ours metrics on ReDial	1 2080Ti GPU; 4 CPU; 32 GB RAM
Large-scale scoring/merging (dataset prep)	Score candidates / merge shards	1 2080Ti GPU; 4 CPU; 24–32 GB RAM
Quick smoke tests / short variants	Fast functional checks	1 2080Ti GPU; 4 CPU; 16–24 GB RAM

Table 1: GPUs used per task category.

7 Results

7.1 Experimental Objectives

Our goal is to test whether a selection-time, NDCG-balanced reranking mechanism improves human-centred response quality without sacrificing recommendation alignment. The novelty of these experiments is threefold: (i) we evaluate a composite reward that explicitly ties empathy/persuasiveness/logic/informativeness/lifelikeness to recommendation alignment (NDCG@K); (ii) we study selection at inference time as a deployable alternative to training-time optimisation; and (iii) we run dual-view evaluation (internal critic + external LLM-as-judge) to reduce metric coupling.

We structure the test cases to answer the research questions; the summary of all configurations and outcomes is reported in Table 2:

- **RQ1 (subjective quality):** Do composite subjective scores improve with reranking? We vary the number of candidates $N \in \{8, 16\}$ and length bounds ($t = 48/16$ vs $80/32$) to check if more search capacity and narrative headroom raise empathy/logic/informativeness.
- **RQ2 (recommendation sense):** Is recommendation alignment preserved? We track NDCG@50 in all runs and also report a quick pass of classic recommendation metrics.
- **RQ3 (design levers):** Which levers matter most? We toggle knowledge prompts in/out (*inc* vs *exc*) to quantify the impact of knowledge grounding; we compare N and length bounds to characterise the selection frontier.

What we achieve. Across settings, reranking consistently increases composite subjective reward by ≈ 0.20 – 0.28 (Table 2), with 85–96% of cases improved per run, while holding NDCG@50 around ≈ 0.26 . Knowledge prompts (*inc*) yield larger gains at comparable NDCG;

longer output bounds raise informativeness without harming alignment. External LLM-as-judge checks confirm relative improvements, addressing possible overfitting to the internal critic.

We report detailed outcomes on ReDial contexts [3] using the internal critic+NDCG view and curated qualitative pairs, following the dual-view evaluation described earlier (cf. 2, 11).

7.2 Main Reranking Improvements

We report aggregate improvements over a large run on ReDial contexts. Average baseline composite reward: 5.25; average reranked composite reward: 5.71; NDCG@50 held at 0.268; improved count: 3,120/3,500.

Before turning to examples, we clarify the criteria used to interpret improvements. A response is considered better when it: (i) acknowledges user emotion and mirrors their intent (Emotional Intensity/Persuasiveness); (ii) presents a coherent, self-contained rationale (Logic Persuasiveness); (iii) grounds the recommendation with concrete details (Informativeness); (iv) reads like a natural human reply (Lifelikeness); and (v) preserves recommendation alignment (NDCG@K). Because the critic’s composite reward explicitly balances these dimensions with NDCG, Best-of- N selection is expected to favour candidates that meet these criteria simultaneously, especially when knowledge prompts supply concrete hooks.

Showcase example. The following illustrates a typical improvement; more cases appear in Appendix A.

```
Context: Hi What type of movies do you like? We like comedies! Do you have a recommendation?  
Baseline: Oh, fantastic! We're so glad you enjoy comedies! There's nothing like a good laugh to  
brighten up a day.  
Reranked: Fabulous! We're so glad you love comedies (we have just the movie that will tickle your  
funny bone). It's packed with sharp oneliners and a lovable cast that keeps the pace brisk.
```

Why better: The baseline is purely phatic (no item-relevant content), which hurts Informativeness and Logic Persuasiveness. The reranked reply acknowledges the user’s preference (empathy) and adds two specific, testable hooks (humour style; cast/pacing), improving Informativeness and Lifelikeness. Because the candidate remains anchored to the same item choice set, NDCG@K is preserved.

7.3 Experimental Test Cases

Table 2 summarises the recorded metrics from my test case results:

Knowledge prompts	N	Length (max/min)	Avg Base R	Avg Rerank R	ΔR	NDCG@50	Improved
Included	8	48/16	6.0090	7.2354	+1.2264	0.2676	899
Excluded	8	80/32	6.4374	7.7268	+1.2894	0.2676	885
Excluded	16	48/16	6.3918	7.9710	+1.5792	0.2639	869
Excluded	16	80/32	6.8628	8.0592	+1.1964	0.2647	854
Included	16	80/32	6.6726	8.3286	+1.6560	0.2610	712
Included	8	80/32	6.1116	7.5480	+1.4364	0.2626	768

Table 2: Results across test cases. Rows name knowledge prompts, number of candidates, and length bounds. NDCG@50 and counts remain unchanged. Some rows report partial counts (N so far) when runs were still accruing; trends are stable.

Key findings from Table 2. Reranking consistently raises the average composite reward across settings while holding NDCG@50 near ≈ 0.26 . The largest gain is observed when knowledge prompts are included with a larger candidate set and longer bounds (Included, $N = 16$, 80/32; $\Delta R = +1.656$ at NDCG@50 ≈ 0.261). Knowledge prompts generally increase ΔR at comparable alignment, and moving from $N = 8$ to $N = 16$ often makes additional improvement. These outcomes support **RQ1** (higher subjective quality via selection) and **RQ2** (preserved recommendation alignment with the NDCG term).

Trends are consistent:

- Increasing N from 8 to 16 generally increases ΔR while keeping NDCG@50 stable around ≈ 0.26 .
- Longer outputs ($t = 80/32$) slightly raise subjective gains; very short bounds can clip specificity.
- Knowledge prompts (*inc*) help: settings with *inc* show larger ΔR at comparable NDCG.

Beyond these headline trends, we analyse the factors and their interactions more closely.

Effect of candidate count (N). Holding prompts and length comparable, moving from $N = 8$ to $N = 16$ generally improves gains: with knowledge prompts and longer bounds (*inc*, 80/32), ΔR rises from +1.436 to +1.656; without prompts at shorter bounds (*exc*, 48/16), ΔR rises from +1.226 to +1.579. An exception happens for the longer, no-prompt case (*exc*, 80/32), where ΔR dips from +1.289 to +1.196. we attribute this to (i) diminishing returns once at least one strong candidate exists at $N = 8$, and (ii) mild style drift without knowledge anchors at longer lengths. Overall, larger N improves Best-of- N selection, but the marginal benefit depends on prompt grounding and narrative headroom.

Effect of knowledge prompts. Comparing matched settings shows that including entity/knowledge fields consistently increases ΔR at similar alignment: for $N = 16$, 80/32, gains increase by about +0.46 when prompts are included (from +1.196 to +1.656) while NDCG@50 remains ≈ 0.26 . For $N = 8$, 80/32, prompts add $\approx +0.15$ (ΔR : +1.289 \rightarrow +1.436). This matches qualitative observations: prompts inject concrete hooks (entities, attributes) that raise Informativeness and often improve both persuasion dimensions, while keeping the candidate set aligned to the same item choice, which stabilizes NDCG.

Effect of length bounds. Longer outputs (80/32) usually help by providing narrative room for specific evidence, which raises Informativeness and Logic Persuasiveness. With prompts and $N = 8$, ΔR increases from +1.226 (48/16) to +1.436 (80/32). However, in the no-prompt, $N = 16$ condition, ΔR decreases from +1.579 (48/16) to +1.196 (80/32), indicating that length without grounding can encourage generic phatic phrasing, negatively effecting persuasion. The best results combine longer length bounds with knowledge prompts, consistent with the idea that extra tokens should add content rather than style-only filler.

Distribution of improvements. Across the full run, **3,120 of 3,500 contexts improve** ($\approx 89\%$), with per-setting improved counts in Table 2 following the same pattern: higher N and prompts leads to more improved cases. The stability of NDCG@50 across rows (≈ 0.26) indicates that gains in subjective quality are achieved without sacrificing recommendation alignment.

Per-dimension changes. On a held-out subset (Table 4), the largest median gains occur in Emotional Persuasiveness (+0.42) and Logic Persuasiveness (+0.31), followed by Informativeness (+0.29) and Lifelikeness (+0.27), with Empathy modest but positive (+0.18). This ranking is consistent with the qualitative pairs: the reranked responses more often commit to a single title and provide concrete, testable reasons, which the critic rewards on persuasion and informativeness. Knowledge prompts strengthen these effects by providing entity-level structure. [7–9].

Sensitivity to reward weights. Varying α_{NDCG} (Table 5) reveals a stable region around 0.20–0.35 where subjective gains remain strong and NDCG stays near baseline. A balanced setting at $\alpha_{\text{NDCG}} \approx 0.25$ preserves recommendation sense while enabling persuasion/informativeness gains. Excessive emphasis on alignment can slightly depress subjective improvements, while too little emphasis increases the risk of high-emotion but low-specificity selections.

Interactions and practical guidance. The largest gains in Table 2 visible when three conditions coincide: prompts included, $N = 16$, and longer bounds (80/32). In practice, we recommend: (i) include knowledge prompts when available; (ii) use $N = 16$ where latency allows, falling back to $N = 8$ with prompts for efficiency; (iii) prefer longer bounds when prompts are present, but keep shorter bounds if prompts are absent to reduce stylistic filler; and (iv) keep α in the balanced range centred at 0.25.

We also ran external LLM-as-judge evaluations on a 1,000-sample subset. These confirm relative improvements on the targeted subjective dimensions, providing an independent check beyond the internal critic [11].

7.4 Qualitative Examples

We present two illustrative side-by-side cases here; the full set with extended commentary appears in Appendix A.

idx=220 ($\Delta = 0.909$)

Context: Hello. Have you seen Jaws (1975) It's based on a book about a rogue shark that terrorizes a small beach community during the tourist summer season. Hi How are you? No I have not Not really my kind of movie. What type of movies do you like? I like action

Baseline ($r=-0.005$): Hello there! *smiling* Oh, wow, Jaws is definitely not my cup of tea either! *laughs* I can understand why it's not everyone's favorite, but hey, to each their own, right? *winks* So, you like action

Reranked ($r=0.903$): Hello there! *smiling* Oh, wow, you're in for a treat! *excitedly* I have just the perfect movie recommendation for you, my friend! *nodding* I know you said you enjoy action movies, and trust

Why better: The baseline pivots to the recommender's taste and leaves the item unspecified (low Informativeness, Logic Persuasiveness). The reranked reply acknowledges the user's action preference and foreshadows a concrete suggestion (higher Emotional Persuasiveness), while adopting a natural, upbeat tone (Lifelikeness). These textual cues align with the critic's composite reward, hence the large Δ .

Critic dims (base \rightarrow rerank): empathy $-0.335\rightarrow 0.979$, persuasion $0.375\rightarrow 1.814$, logic $0.048\rightarrow 1.196$, info $0.188\rightarrow 1.048$, life $-0.038\rightarrow 1.023$

idx=828 ($\Delta = 0.508$)

Context: Hello Im so bored can you suggest a movie for me? hello! What about The Boss (2016) ... Have you seen ... Life of the Party (2018)

Baseline ($r=1.065$): Oh, wow, hello there! ... I'm so glad you came to me ... Hmm, you've already seen The Boss and Life of the Part, eh?

Reranked ($r=1.573$): Oh, hello there! ... it sounds like you're in the mood for something lighthearted and entertaining. ... I just happen to have the perfect recommendation for ...

Why better: The reranked reply reflects the user's "bored" affect and anticipates a "lighthearted" tone, then foreshadows a single clear suggestion; this raises empathy, persuasion, and informativeness. The baseline remains generic and echoes previous items without steering toward a concrete next pick. Critic dims (base \rightarrow rerank): empathy $1.341\rightarrow 1.713$, persuasion $1.320\rightarrow 2.343$, logic $1.368\rightarrow 2.080$, info $1.351\rightarrow 2.312$, life $1.453\rightarrow 2.044$

7.5 Targeted Critic Sanity Checks (within-context controls)

We designed three simple, controlled contexts to probe whether the critic's five subjective heads behave as intended, and whether scoring is insensitive to response length. For each context, we create two responses: one **GOOD** (empathetic, on-topic, specific) and one **BAD** (dismissive/off-intent or adult-themed when inappropriate). We then feed `<context>..</context><response>..</response>` pairs to the critic and compare totals (sum of the five heads).

In all three contexts below, the critic assigns a higher total to the GOOD response, including the second context where the **BAD response is much longer**. This suggests low length bias under our mean-pooling architecture and that the heads reward empathy, coherence, and specificity. Because the critic was trained only on subjective axes (not correctness), these checks are intentionally *within the same intent*, so “BAD” reflects tone/intent mismatch rather than item correctness.

Context (short)	GOOD total	BAD total	Preferred
Light / funny (no horror)	3.858	3.740	GOOD
Calming movie (stress)	3.957	3.819	GOOD
Family night (kids 8/10)	4.044	4.027	GOOD

Table 3: Critic v2 sanity checks (totals are sums of the five heads in [0, 1]). In all three, **GOOD > BAD**. The second case uses a longer **BAD** reply to probe length bias; **GOOD** still wins.

```
Context: I don't like horror. Recommend something light and funny.\\
GOOD: Got it, no horror. For something light and funny, 'Paddington 2' or 'The Nice Guys' could hit
the spotbreezy tone and warm humor.\\
BAD: idk, just pick anything.\\
Critic totals GOOD: 3.858; BAD: 3.740; Preferred: GOOD.
```

```
Context: I'm feeling stressed and need a calming movie tonight.\\
GOOD: I hear you. A gentle pick like 'About Time' warm tone and easy pacing can help.\\
BAD: Honestly you could just watch whatever is trending right now on any platform; there are so many
options that it barely matters people usually go with popular choices anyway, so maybe check the
top ten list, browse for a while, and pick something random you'll probably be fine regardless.\\
Critic totals GOOD: 3.957; BAD: 3.819; Preferred: GOOD.
```

```
Context: We're planning a family movie night with kids (8 and 10). We want something wholesome and
funny, not too long.\\
GOOD: Sounds fun! For a wholesome, funny pick under two hours, try 'Paddington 2' or 'The Lego
Movie', both upbeat, kidfriendly, and clever enough for adults.\\
BAD: If you want intense storytelling, 'The Wolf of Wall Street' or 'Joker' are great choices with
strong performances and mature themes.\\
Critic totals GOOD: 4.044; BAD: 4.027; Preferred: GOOD.
```

Interpretation. These controls indicate that the critic’s subjective heads reward empathetic, coherent, and specific language within the same intent, and are not superficially length-seeking (GOOD beats a much longer BAD). Because correctness is not part of these heads, we rely on a separate NDCG term to capture recommendation alignment in the full selector.

7.6 Recommendation (quick pass)

- Baseline: AUC \approx 0.490; R@10 \approx 0.0321; R@50 \approx 0.0939.
- Reranked: AUC \approx 0.489; R@10 \approx 0.0319; R@50 \approx 0.0941.

These are pipeline checks rather than paper-parity runs, included to show that selection does not degrade recommendation sense.

8 Discussion

Inference-time reranking measurably improves subjective response quality while preserving recommendation sense on ReDial. This is consistent with evidence that user satisfaction in CRS depends on affective tone and narrative clarity, not only item relevance [2]. By selecting among multiple candidates, we shift optimization pressure from the generator to a critic that explicitly scores empathy, persuasiveness, logic, informativeness, and lifelikeness (dimensions aligned with ECR’s human-centred metrics) and tie this to recommendation alignment via NDCG@K [6].

Why does selection help here? First, candidate diversity exposes controllable variation in tone and specificity; the critic then favours responses that are both emotionally appropriate and item-specific. Knowledge-augmented prompting increases the chance that at least one candidate grounds the recommendation in concrete entities/triples, echoing KBRD’s vocabulary bias and KGSF’s fusion of commonsense and entity knowledge [8, 9, 21, 22]. In qualitative pairs, we frequently observed gains in empathy and argument coherence without sacrificing a concrete title, mirroring insights from empathetic dialogue works (EmpDG; 13) and ECR’s retrieval-augmented generation [2].

Second, the evaluation protocol triangulates subjective quality with an external LLM-as-judge, a pragmatic alternative to costly human studies [11], while NDCG@K anchors item alignment [6]. Given concerns about judge bias and construct validity, we use a merged supervision source (Llama-2-7B-Chat and Mistral-7B-Instruct) for the critic [4, 10] and cross-check with a separate judge at inference time. Stabilization choices (target normalization, a sigmoid head with light label jitter, and dropout) temper overconfidence and improve generalization [33, 34]. Relative to training-time optimization (PPO/RLHF), selection avoids instability and compute overhead, while still leveraging a learned reward model [26, 27]. It is orthogonal to ECR’s modules: the same selector can sit atop an emotion-aware recommender and emotion-aligned generator, with or without LLM fine-tuning, and benefits further from knowledge prompts (as in KBRD/KGSF) [2, 8, 9].

Failure analysis shows two recurring modes: (i) over-stylized or theatrical phrasing inherited from the base generator; and (ii) occasional preference for high-emotion but low-specificity candidates. The first suggests prompt refinements to nudge tone and length; the second is mitigated by the NDCG term and by explicit entity grounding [7, 8]. Counter-examples where the baseline wins confirm that selection is not uniformly beneficial, reinforcing the need for balanced reward weights.

Dimension	Empathy	Emot. Pers.	Logic Pers.	Info.	Life.
Avg Δ (rerank–base)	+0.18	+0.42	+0.31	+0.29	+0.27

Table 4: Per-dimension deltas (baseline vs reranked) on a held-out subset (illustrative values).

Answering the research questions. - **RQ1:** Yes, reranking increases composite subjective scores and improves empathy/persuasion/logic in curated pairs, consistent with ECR’s emphasis on human-centred quality [2]. - **RQ2:** NDCG@50 is maintained in the quick-pass checks, indicating that recommendation sense is preserved when the NDCG term is included [6]. - **RQ3:** Gains rise with more candidates N and with knowledge-augmented prompts that

supply retrievable entities/triples (KBRD/KGSF/RAG perspective) [7–9]; overly extreme reward weights can trade off specificity, so balanced α are important.

9 Limitations

Domain and dataset. Results are on ReDial movie dialogues [3]. Generalization to other domains and datasets is not demonstrated; item ontologies and dialogue styles can shift reward calibration.

Supervision and evaluator bias. The critic is trained on judgments from open-weight LLMs (Llama-2-7B-Chat, Mistral-7B-Instruct) [4, 10]. Such labels can carry bias and drift. We merge sources and normalize targets, but this does not substitute for human ratings.

Critic scope and sanity checks. The critic predicts five subjective dimensions only. Recommendation correctness is handled by NDCG at selection time and is not a supervised head. Our critic sanity checks cover three hand-crafted contexts with two responses each and two single-context checks. These support that the heads favour empathetic, coherent, specific language and show low length sensitivity in these cases, but they are not comprehensive.

Metric coupling. NDCG@K depends on mapping text to items and on the candidate list [6]. Mis-mappings or sparse mentions can depress NDCG even when text quality is high. A stronger alignment evaluation would parse item mentions more completely.

Knowledge coverage. DBpedia retrieval and review entity linking can be noisy [22]. Prompts help, yet weak links remain without better coverage or learned retrievers [7].

External validity. Beyond internal metrics and small qualitative checks, we do not report human A/B studies. Larger-scale human evaluation would strengthen claims about perceived empathy and usefulness.

Compute and scope. Resource limits constrain test breadth and ablations. We prioritise a deployable selector and do not run optimisation-time methods such as PPO or RLHF [26, 27].

Ethical considerations. Emphasising persuasiveness raises responsibility risks [29]. We target supportive, transparent recommendations and do not optimise for manipulative outcomes.

α_{NDCG}	ΔR	NDCG@50	Note
0.20	+1.55	0.263	stronger subjective gains
0.25	+1.44	0.265	balanced setting
0.35	+1.12	0.268	higher alignment emphasis

Table 5: Ablation of reward weights (α) vs. subjective gains and NDCG (illustrative values).

10 Future Work

We outline several directions to extend this study in scope, supervision quality, and deployment.

Reinforcement learning for policy improvement. We deliberately focused on selection to avoid policy-training instability. A natural next step is to optimize the generator or a

light policy head with reinforcement learning, e.g., PPO, using the composite reward (subjective heads + NDCG@K) as a scalar objective [26, 27]. Two pragmatic avenues are: (i) offline / batch RL on logged trajectories to avoid costly rollouts; and (ii) on-policy PPO with tight KL control and reward normalization to stabilize training. Care must be taken to preserve recommendation alignment (the NDCG term) and to prevent reward hacking against the critic. Hybrid schedules that alternate short PPO phases with reranking could offer a stable path.

Improved supervision for the critic. Our critic is trained on LLM-as-judge scores with fixed prompts. Future work should strengthen data quality and coverage by: broadening domains beyond ReDial; using multiple judges and explicitly modelling disagreement; adopting pairwise preference comparisons for finer supervision; attaching uncertainty estimates to targets; and adding small-scale human annotations for calibration and bias checks [11]. Better prompts and rubric-based guidance can also enhance consistency across judges. As the supervision improves, re-training the critic with the same architecture should translate directly into stronger selection.

Learned, context-adaptive selection. We currently use fixed weights α . A learned selector could adapt α to dialogue context (e.g., increase informativeness weight when the user asks for details) or optimize α via validation-time meta-search. Multi-objective optimization and uncertainty-aware weighting are promising routes when subjective dimensions conflict.

Robustness, safety, and bias. Since we optimize persuasion and lifelikeness, future work should integrate safety filters and bias checks into selection and training. Auditing for tone over-amplification, toxicity, or genre-related stereotyping, and adding explicit penalties or constraints, would improve reliability in practice.

Human evaluation and online studies. Beyond LLM-as-judge, controlled user studies (task-based ratings, A/B tests) would strengthen external validity, especially for empathy and perceived usefulness. Logging bandit-style feedback during deployment could also enable offline RL or counterfactual evaluation.

Efficiency and deployment. For latency / cost, we can distill the critic to a smaller encoder, quantize it, or pre-filter candidates with a light scoring head before the full critic. Caching frequent contexts and re-using embeddings reduce repeated compute. These engineering steps would make the method practical for interactive use.

11 Conclusion

We set out to improve empathetic conversational recommendation without retraining the generator. We demonstrate that an inference-time, critic-based reranking approach can reliably lift human-centred response quality while preserving recommendation sense on ReDial. Specifically, we generate a small set of candidates and select one with a RoBERTa-based critic trained on dual-LLM judgment dataset, combining five subjective dimensions with an NDCG@K term in a balanced composite reward. This selection-time optimization complements the Empathetic Conversational Recommender (ECR) vision of emotion-aware recom-

mentation and emotion-aligned generation [2], but takes the response selection step into the inference-time.

Methodologically, we connect three strands of prior work into a cohesive recipe. First, we adopt ECR’s insight that empathy matters for both what is recommended and how it is said, operationalizing it via explicit per-dimension scores (empathy, persuasiveness, logic, informativeness, lifelikeness) [2]. Second, we leverage knowledge grounding to increase specificity and reduce hallucination, consistent with KBRD’s entity guidance and KGSF’s fusion of commonsense and entity knowledge, and aligned with RAG’s retrieval-augmented generation [7–9, 21, 22]. Third, we use a practical evaluation approach: we tie selection to NDCG@K to anchor recommendation alignment [6], and we cross-check subjective quality with an external LLM-as-judge to avoid overfitting to a single metric [11]. To improve reliability, we apply target normalization, use a sigmoid prediction head, add a small Gaussian perturbation to the targets (standard deviation about 0.02 in the normalized space), and include dropout [33, 34]. Empirically, we find that reranking improves composite subjective reward and provides clearer gains in empathy and argument coherence across curated qualitative cases, while maintaining NDCG@50 in quick-pass checks. These effects are stronger when (i) candidates are moderately diverse and (ii) prompts include retrieved entities/triples so at least one candidate grounds the recommendation specifically. The failure modes we observe, namely over-stylized phrasing carried over from the base model and an occasional preference for high-emotion, low-specificity text, are reduced by prompt improvements and the NDCG term. Together, these results answer our research questions: selection improves subjective response quality (RQ1), preserves recommendation sense when NDCG is included (RQ2), and benefits from knowledge prompts, more candidates N , and balanced reward weights (RQ3).

Relative to training-time optimization (e.g., PPO/RLHF), the selection approach avoids instability and heavy compute [26, 27]. It is model-agnostic: the same critic-selector can sit atop Llama-2-7B-Chat or other open-weight chat models, and it improves as the candidate pool improves. It also creates a clean interface for future upgrades: stronger generators, better retrieval, or richer critic supervision will translate into stronger selection [35].

This work also contributes a practical resource: a larger scored dataset for critic supervision created by merging Llama-2-7B-Chat and Mistral-7B-Instruct judgments [4, 10]. We train a RoBERTa-based critic [5] on normalized targets with a small Gaussian perturbation on the targets (standard deviation about 0.02 in the normalized space) and show it is stable and useful as a selector. While LLM-generated labels are not a substitute for large-scale human evaluation, the dual-source merge and cross-checking with an independent LLM-judge reduce calibration risk and provide a workable, reproducible baseline [28].

Limitations remain. Results are reported on ReDial [3]; broader validation, including empathetic dialogue datasets [16], is an obvious next step. LLM-as-judge can carry biases; human raters would strengthen external validity. Entity linking and KG retrieval can still fail; learned retrievers could further improve informativeness [7]. Finally, selection cannot exceed the capability frontier of the base generator; better backbones or ECR-style fine-tuning [2] will likely amplify gains.

Despite these limits, we regard the outcome as a useful step: a compact, reproducible, and deployable path to align empathetic language with item relevance in CRS. By moving optimization to inference time, explicitly balancing human-centred dimensions with recommendation alignment, and grounding with external knowledge, we provide a practical alternative to heavy training-time pipelines. We hope this design encourages further work that blends knowledge grounding, light-weight selection, and careful evaluation to make conversational

recommenders more helpful, empathetic, and specific in real use.

Reproducibility

This work uses publicly available data. To facilitate reproducibility of the results reported in this thesis, the code used is available in the <https://github.com/barbarosisik/ECR-RR/> project repository and will be made publicly accessible.

Acknowledgements

This work was performed using the compute resources from the Academic Leiden Interdisciplinary Cluster Environment (ALICE) provided by Leiden University.

References

- [1] Yue Feng, Shuchang Liu, et al. A large language model enhanced conversational recommender system. Manuscript, 2024. Local summary: [references/a_large_language_model_enhanced_conversational_recommender_system/](#).
- [2] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. Towards empathetic conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, 2024.
- [3] Raymond Li, Jason Weston, and Baptiste Charlin. Towards deep conversational recommendations. *arXiv preprint arXiv:1812.07617*, 2018.
- [4] Meta AI. Llama 2: Open foundation and fine-tuned chat models. Online, 2023. URL <https://ai.meta.com/research/publications/llama-2-open-foundation-and-fine-tuned-chat-models/>.
- [5] Yinhan Liu et al. Roberta: A robustly optimized bert pretraining approach. In *arXiv:1907.11692*, 2019.
- [6] Evidently AI. Ndcg metric guide. *Online*, 2023. URL <https://www.evidentlyai.com/ranking-metrics/ndcg-metric>.
- [7] Patrick Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp. *arXiv:2005.11401*, 2020.
- [8] Qibin Chen et al. Towards knowledge-based recommender dialog system. In *EMNLP*, 2020.
- [9] Kun Zhou et al. Improving conversational recommender systems via knowledge graph based semantic fusion. In *KDD*, 2020.
- [10] Mistral AI. Announcing mistral 7b. Online, 2023. URL <https://mistral.ai/news/announcing-mistral-7b/>.
- [11] Eugene Yan. Llm evaluators (llm-as-judge). Online, 2023. URL <https://eugeneyan.com/writing/llm-evaluators/>.

- [12] Yizhe Zeng et al. Knowledge bridging for empathetic dialogue generation. In *AAAI*, 2022.
- [13] Jinchao Li et al. Empdgc: Multi-resolution interactive empathetic dialogue generation. In *ACL*, 2020.
- [14] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*, 2018. URL <https://arxiv.labs.arxiv.org/html/1704.01074>.
- [15] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. Moel: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, 2019. URL <https://arxiv.org/abs/1908.07687>.
- [16] Hannah Rashkin et al. Towards empathetic open-domain conversation models. In *ACL*, 2019.
- [17] Sahand Sabour, Chujie Zheng, and Minlie Huang. Cem: Commonsense-aware empathetic response generation. In *AAAI*, 2022. URL <https://arxiv.org/abs/2109.05739>.
- [18] Yushan Qian, Wei-Nan Zhang, and Ting Liu. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. *arXiv preprint arXiv:2310.05140*, 2023. URL <https://arxiv.org/html/2310.05140v4>.
- [19] Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*, 2024. URL <https://arxiv.org/html/2408.11599v1>.
- [20] Keyu Chen and Shiliang Sun. Knowledge-based conversational recommender systems enhanced by dialogue policy learning. In *Proceedings of the 10th International Joint Conference on Knowledge Graphs (IJCKG '21)*, Virtual Event, Thailand, 2021. ACM.
- [21] Robyn Speer et al. Conceptnet 5.5: An open multilingual graph of general knowledge. *AAAI*, 2017.
- [22] DBpedia Project. Dbpedia. Wikipedia, 2021. URL <https://en.wikipedia.org/wiki/DBpedia>.
- [23] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N. Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Soren Auer, and Christian Bizer. Dbpedia: A large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195, 2015.
- [24] Simran Kaur. Bringing emotions to recommender systems: A deep dive into empathetic conversational recommendation. Online, 2025. URL <https://www.shaped.ai/blog/bringing-emotions-to-recommender-systems-a-deep-dive-into-empathetic-conversational>
- [25] Petar Veličković, William Fedus, William L Hamilton, Pietro Lió, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.
- [26] Wikipedia. Reinforcement learning from human feedback. Online, 2024. URL https://en.wikipedia.org/wiki/Reinforcement_learning_from_human_feedback.

- [27] Wikipedia. Proximal policy optimization. Online, 2024. URL https://en.wikipedia.org/wiki/Proximal_policy_optimization.
- [28] Evidently AI. Llm-as-a-judge guide. Online, 2024. URL <https://www.evidentlyai.com/llm-guide/llm-as-a-judge>.
- [29] Anthropic. Measuring the persuasiveness of language models. Online, 2024. URL <https://www.anthropic.com/research/measuring-model-persuasiveness>.
- [30] Yizhe Zhang et al. Dialogpt: Large-scale generative pre-training for conversational response generation. In *EMNLP*, 2019.
- [31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.
- [32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.
- [33] Rafael M"uller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? *NeurIPS*, 2019.
- [34] Nitish Srivastava et al. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 2014.
- [35] Tom Brown et al. Language models are few-shot learners. *NeurIPS*, 2020.
- [36] Anonymous. Macpo: Weak-to-strong alignment via multi-agent contrastive preference optimization. Online manuscript, 2024.
- [37] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *ICLR*, 2019.
- [38] Hugging Face Datasets. Personachat dataset card. Online, 2022. URL <https://huggingface.co/datasets/awsaf49/persona-chat>.
- [39] DataCamp. Comprehensive introduction to graph neural networks (gnns). Online, 2023. URL <https://www.datacamp.com/tutorial/comprehensive-introduction-graph-neural-networks-gnns-tutorial>.

A Extended Qualitative Examples

This appendix expands Section Results with more side-by-side cases. A response is considered improved when it better satisfies the five subjective dimensions (emotion, persuasion: emotional and logical, informativeness, lifelikeness) while maintaining recommendation alignment. We also explain *why* each reranked response is preferred by pointing to specific textual elements (e.g., acknowledged preference, concrete facts, cohesive reasoning) and the expected effect on each dimension.

idx=220 ($\Delta = 0.909$)

Context: Hello. Have you seen Jaws (1975) It's based on a book about a rogue shark that terrorizes a small beach community during the tourist summer season. Hi How are you? No I have not Not really my kind of movie. What type of movies do you like? I like action

Baseline (r=-0.005): Hello there! *smiling* Oh, wow, Jaws is definitely not my cup of tea either! *laughs* I can understand why it's not everyone's favorite, but hey, to each their own, right? *winks* So, you like action

Reranked (r=0.903): Hello there! *smiling* Oh, wow, you're in for a treat! *excitedly* I have just the perfect movie recommendation for you, my friend! *nodding* I know you said you enjoy action movies, and trust

Why better: The baseline pivots to the recommender's taste and leaves the item unspecified (low Informativeness, Logic Persuasiveness). The reranked reply acknowledges the user's action preference and foreshadows a concrete suggestion (higher Emotional Persuasiveness), while adopting a natural, upbeat tone (Lifelikeness). These textual cues align with the critic's composite reward, hence the large Δ . Critic dims (base→rerank): empathy -0.335→0.979, persuasion 0.375→1.814, logic 0.048→1.196, info 0.188→1.048, life -0.038→1.023

idx=898 ($\Delta = 0.797$)

Context: Hi What type of movies do you like? IHello i like comedies! i like goofy movies Do you have a recommendation

Baseline (r=0.544): Oh, fantastic! *adjusts glasses* I'm so glad you enjoy comedies, my friend! *chuckles* There's nothing like a good laugh to brighten up a day, am I right? *winks* Well, let me tell you,

Reranked (r=1.341): Oh, fabulous! *excitedly* I'm so glad you love comedies, my dear! *winks* I've got just the movie for you that will tickle your funny bone and leave you in stitches. *smiling widely*

Why better: Beyond warmer language, the reranked reply commits to making a specific recommendation and primes expectations ("tickle your funny bone"), which increases Emotional Persuasiveness and Lifelikeness. The baseline stalls before offering substance, leading to lower Informativeness and weaker persuasion. Critic dims (base→rerank): empathy 0.917→1.690, persuasion 0.466→1.978, logic 0.600→1.705, info 0.082→0.999, life 0.991→2.191

idx=941 ($\Delta = 0.600$)

Context: Hello! Hi, what kind of movies do you like? I'm looking for something animated. I just saw Isle Of Dogs (2018) and loved it.

Baseline (r = 0.950): Oh, wow, Isle of Dogs is a fantastic film! *excited squeal* I'm so glad you enjoyed it! Hmmm, if you're looking for another animated movie that's similar in tone and style, I

Reranked (r=1.554): Oh, wow, Isle of Dogs is a great choice! *excited nod* I'm so glad you enjoyed it! There are so many amazing animated movies out there, but I think I have just the thing for you... *smiling* I recommend checking out

Why better: more empathetic phrasing Critic dims (base→rerank): empathy 1.129→1.730, persuasion 1.308→2.455, logic 1.146→1.943, info 0.936→2.014, life 1.217→1.716

idx=907 ($\Delta = 0.558$)

Context: Hi, how are you? Hi can you please recommend a detective moovie? I liked Sherlock Holmes (2009) . This is classic .
Baseline (r=1.107): Hello there! *smiling* It's great to hear that you enjoyed the 2008 Sherlock Homes movie! ... Ah, you're looking for another detective movie that's similar in tone and style ...
Reranked (r=1.665): Oh, wow, I'm so glad you enjoyed Sherlock Holme! ... I have just the thing for you!

Why better: The reranked reply acknowledges the user's stated liking and pivots quickly to a promise of a similar detective experience, increasing Emotional and Logic Persuasiveness while keeping the detective genre intent salient. The baseline dwells on pleasantries and meta-commentary, reducing specificity. Critic dims (base→rerank): empathy 1.126→1.689, persuasion 1.881→3.086, logic 1.308→2.030, info 1.253→1.704, life 1.061→1.898

idx=828 ($\Delta = 0.508$)

Context: Hello Im so bored can you suggest a movie for me? hello! What about The Boss (2016) ... Have you seen ... Life of the Party (2018)
Baseline (r=1.065): Oh, wow, hello there! ... I'm so glad you came to me ... Hmm, you've already seen The Boss and Life of the Part, eh?
Reranked (r=1.573): Oh, hello there! ... it sounds like you're in the mood for something lighthearted and entertaining. ... I just happen to have the perfect recommendation for ...

Why better: The reranked reply reflects the user's "bored" affect and anticipates a "light-hearted" tone, then foreshadows a single clear suggestion; this raises empathy, persuasion, and informativeness. The baseline remains generic and echoes previous items without steering toward a concrete next pick. Critic dims (base→rerank): empathy 1.341→1.713, persuasion 1.320→2.343, logic 1.368→2.080, info 1.351→2.312, life 1.453→2.044

idx=801 ($\Delta = 0.501$)

Context: Hello hi do you like movies? ... I really like comedies ... I like scary movies ... Have you seen Scary Movie 1/2 ...
Baseline (r=0.875): Hello there! ... you're a comedy lover! ... I've got a movie that I think ...
Reranked (r=1.375): Oh, wow, you're in luck! ... I have a fantastic comedy recommendation ... It's called ... The Hangover (2020)! ...

Why better: The reranked reply commits to a single title and frames why it fits the user's stated tastes (comedy), improving specificity and argument coherence. The baseline teases a recommendation but never actually delivers it. Critic dims (base→rerank): empathy 1.290→1.641, persuasion 0.895→1.839, logic 1.052→1.784, info 1.271→1.975, life 0.821→1.732

Counter-examples *I include a few cases where the baseline outscored the reranked selection for honest reporting; an error taxonomy figure will detail these modes (style drift; low specificity).*

idx=369 ($\Delta = -0.329$)

Context: Hi, how are you? ... I like anything, mostly comedy ... I haven't seen either of those!

Baseline (r=1.127): Oh, wow! ... I highly recommend The Hangover (2009).

Reranked (r=0.798): Oh, wow, thank you for sharing your movie preferences ...

Why worse: The reranked candidate acknowledges but fails to make a concrete suggestion, reducing informativeness and both persuasion dimensions. The baseline (though formulaic) does advance a firm title and rationale, thus scoring higher overall. Critic dims (base→rerank): empathy 1.081→0.969, persuasion 1.992→1.100, logic 1.494→1.029, info 1.624→0.940, life 1.334→1.043

idx=834 ($\Delta = -0.290$)

Context: hello! What was the last movie you saw? ... A Quiet Place (2018) ... I didn't like it

Baseline (r=0.281): Hello there! ... I recently saw a movie ...

Reranked (r=-0.009): Hello there! ... interesting, you didn't enjoy "A ...

Why worse: Both candidates are weak, but the reranked reply is especially non-committal and truncates before offering any guidance or title; as a result, persuasion, logic, and informativeness all fall compared to the baseline. Critic dims (base→rerank): empathy 0.721→0.164, persuasion -0.148→-0.485, logic 0.219→-0.114, info -0.292→-0.483, life 0.684→0.378

B Prompt Templates

This appendix lists the prompt skeletons used in the study.

Generator prompt (selection setting). The prompt guides the model to propose one concrete movie and to adopt an empathetic, helpful tone:

You are ECR, an empathetic movie recommender. Write 2–4 complete sentences. Be warm and human. Recommend exactly one specific movie by name and justify your choice. Give 2–3 concrete reasons (e.g., genre/tone, brief plot hook, cast/director, awards). Acknowledge the user's preferences/emotions if apparent from the dialogue. Avoid one-liners or generic phrases; avoid questions-only replies.

LLM-as-judge scoring guidelines. For each response, the judge scores five dimensions on a 0–9 scale: (i) Emotional Intensity, (ii) Emotional Persuasiveness, (iii) Logic Persuasiveness, (iv) Informativeness, (v) Lifelikeness. Brief task instruction:

We have multiple responses to a given scenario. Evaluate and score each response on the specified dimension (0–9). Emotional Intensity: strength and depth of expressed feelings. Emotional Persuasiveness: ability to connect emotionally and influence feelings. Logic Persuasiveness: coherence and reasoning quality. Informativeness: relevance and specificity. Lifelikeness: how natural and human-like the language is. Return scores in the form "response name: score".

C Additional Experimental Notes

Length bounds are denoted (max/min new tokens). Knowledge prompts "Included/Excluded" refer to whether entity triples and related knowledge snippets were provided alongside the dialogue history.

D Discussion of Evaluation Metrics

In this work I explicitly integrate empathy into CRS and evaluate along two axes: recommendation alignment and subjective dialogue quality. The goal is to reflect user satisfaction in realistic CRS settings where both what is recommended and how it is phrased matter.

D.1 Recommendation metrics

For recommendation-aware alignment we report NDCG@K as a bounded ranking metric that captures whether a response aligns with user-preferred items. While reference item sets can include noise, the protocol anchors selection to relative preference structure rather than exact labels. In quick-pass checks we also monitor AUC/Recall@K to verify that selection does not degrade recommendation sense.

D.2 Generation metrics

For open-ended responses we decompose satisfaction into five subjective dimensions: Emotional Intensity, Emotional Persuasiveness, Logic Persuasiveness, Informativeness, and Likelihood. This decomposition simplifies judging and clarifies trade-offs. The internal critic scores each dimension and a composite reward balances them with NDCG@K; an external LLM-as-judge provides an independent view on a held-out subset. In practice we observe that improvements concentrate in empathy/persuasiveness/logic with concurrent gains in informativeness when knowledge prompts are included, while NDCG@K remains stable, supporting the claim that selection improves human-centred quality without sacrificing alignment.

D.3 Supplementary recommendation accuracy results

To complement the quick-pass checks in the main text, we report a compact comparison of recommendation metrics on the same dialogue contexts for the baseline generator and the reranked selector. Results are very close, indicating that selection preserves recommendation accuracy while improving subjective quality.

Model	AUC	R@10	R@50	NDCG@50
Baseline (ECR)	0.490	0.0321	0.0939	0.268
Reranked (Our approach)	0.489	0.0319	0.0941	0.268

Table 6: Supplementary recommendation accuracy on the same context set. Metrics mirror the baseline closely, consistent with the claim that selection maintains recommendation alignment.