Evaluation and Analysis of LLMs and Datasets for Conversational Recommendation

Octavio Águila [*1] **Benjamín Pedraza** [*1] **Sebastián Rasmussen** [*1]

Abstract

The use of large language models (LLMs) for conversational recommendation tasks has gained significant relevance in recent years. However, their performance heavily depends on the quality and structure of the datasets used for training. Conversely, it is increasingly critical to ensure that these models are reliable and capable of producing accurate and trustworthy content. To quantify the extent of these challenges in Conversational Recommender Systems (CRS), we (i) evaluate the effectiveness of PEARL and ReDial datasets, specifically designed to enhance recommendation systems, and (ii) propose a methodology for estimating the uncertainty of these models. Using the LLaMA 3, Qwen, and DeepSeek models, we conducted controlled fine-tuning on both datasets and evaluated their performance using metrics such as BLEU, ROUGE, BERTScore, and Recall. In addition, we estimated both the aleatoric and epistemic uncertainty of these models. Our results show that PEARL facilitates the generation of more coherent, relevant, and user-aligned responses, consistently outperforming ReDial across most evaluation metrics. Although ReDial demonstrated superior performance in Recall@1, PEARL proved to be a more robust and effective dataset for training LLMs in conversational recommendation tasks. Uncertainty analysis reveals complementary strengths: DeepSeek is the most self-consistent but least confident, Qwen-2 is the most confident yet slightly less stable, and LLaMA-3 offers the widest range of suggestions with mid-level confidence. Github Link.

[*]Equal contribution [1]Department of Computer Science, Pontificia Universidad Católica de Chile, Santiago, Chile. Correspondence to: Octavio Águila <oeaguila@uc.cl>, Benjamín Pedraza <bpedrazaj@uc.cl>, Sebastián Rasmussen <srasmussenc@uc.cl>.

1. Introduction

In recent years, the use of large language models (LLMs) has expanded significantly into various applications, including conversational recommendation systems (CRS). However, one of the main limitations in the effective development of these systems lies in the quality of the datasets used for training. ReDial (Li et al., 2018), one of the most widely used datasets for CRS, has been criticized for containing overly generic user preferences and lacking domain-grounded explanations.

To address these issues, the PEARL dataset (Kim et al., 2024) was recently proposed as an alternative that incorporates detailed user profiles and knowledge extracted from reviews. In this work, we conduct an experimental evaluation of the performance of LLMs trained on PEARL compared to ReDial, with the goal of validating its effectiveness as a training source for CRS.

Unlike the original PEARL study, our evaluation is carried out using recent open-source language models, such as LLaMa 3.2, Qwen, and DeepSeek, along with a broader set of automatic evaluation metrics. In addition, we introduce a complementary analysis of uncertainty in the generated recommendations, aiming to explore the reliability and consistency of the models when facing ambiguous prompts.

Despite computational limitations that restricted the size of the models used and the scale of the experiments, the results obtained support PEARL as a more robust and effective dataset for training conversational recommendation models.

In parallel, with the widespread use of LLMs, it becomes crucial to ensure that their predictions are reliable. One dimension of reliability is the model's ability to indicate when its generated output is accurate and trustworthy. This challenge is formally known as the problem of *uncertainty quantification*.

Uncertainty in an LLM reflects how confident the model is in its response. There are two types of uncertainty (Kiureghian & Ditlevsen, 2009): *aleatoric uncertainty*, which is inherent to the task itself and arises when the input is ambiguous or allows for multiple valid answers; and *epistemic uncertainty*, which is related to the model's lack of knowledge and can be reduced as the model gains more information about the domain.

This work introduces a methodology to estimate the uncertainty of CRS, based in Input Clarification Ensembling (Hou et al., 2024). The idea behind this method is that, instead of changing the model (which is computational expensive and impractical for LLMs), it modifies the input by generating multiple clarified versions of the original query and then analyzing the variability in the model's responses.

Experimental results show that this methodology enables models to effectively quantify both aleatoric and epistemic uncertainty. Moreover, it improves ambiguity detection and provides more interpretable insights into the model's behavior, ultimately enhancing the reliability and trustworthiness of CRS systems.

2. Datasets

We trained each LLM independently on PEARL and on ReDial, resulting in two models per LLM.

2.1. PEARL

The PEARL (Persona and knowledgE Augmented Recommendation diaLogues) dataset is divided into training (50,000 rows, 9 columns), validation (2,277 rows, 9 columns), and testing (5,000 rows, 9 columns) sets. Each row in the dataset includes the following fields:

- `data_id`: Unique identifier for the sample.
- `user_persona`: Natural language description of a user's likes and dislikes.
- `seen_movie_titles`: List of movies previously watched by the user.
- `gt_abstract`: Metadata for the recommended movie (ground truth).
- `gt_movie_title`: Title of the recommended movie (ground truth).
- `gt_genre`: Genres associated with the recommended movie (ground truth).
- `gt_director`: Director of the recommended movie (ground truth).
- `gt_cast`: Cast of the recommended movie (ground truth).
- `dialogue`: Natural language prompt (chat message) from the user requesting a recommendation.

In the training set, the `seen_movie_titles` field includes 9,393 distinct movies with an average of 15.97 `seen movies` per user. The top 40% of users have watched at least 11 movies, suggesting the mean is a relatively representative measure. The distribution shows a long-tail pattern, where a few users have watched many movies, while most have seen only a few. This exponential trend is also observed in the `movie_titles` and `gt_genre fields`.

In the case of `movie_titles`, the dataset includes 7,564 distinct titles that were recommended, covering approximately 80.5% of the movies previously seen by users—highlighting strong overlap between watched and recommended content.

For `gt_genre`, there are 23 unique genres. This count was obtained by flattening multi-label genre fields (e.g., splitting ”Drama, Action” into ”Drama” and ”Action”). The genre distribution is consistent across training, validation, and test sets.

Overall, all three dataset splits exhibit similar distributions across different data fields. This consistency supports generalization of experimental results. Furthermore, the exponential growth in frequency—e.g., more popular movies being recommended more often—reflects realistic user behavior patterns. Importantly, there are no missing or null values in the dataset.

2.2. ReDial

The ReDial (Recommendation Dialogues) dataset contains conversational data between two users: a seeker, who requests movie recommendations, and a recommender, who provides suggestions based on the dialogue context. The dataset was designed to study and model conversational recommendation systems.

These figures reflect a substantial volume of data, with over 10,000 conversations and nearly 1,000 distinct users. Notably, the number of movie recommendations provided by the recommender is more than twice that of the mentions by the seeker (35,421 vs. 16,278), highlighting the recommender's active role in exposing the user to new content. Furthermore, among the recommendations where user feedback is provided, 81% are marked as ”liked”, indicating strong alignment between the system's suggestions and user preferences.

2.3. Comparison and Discussion

While PEARL and ReDial are both conversational recommendation datasets, they differ significantly in structure, annotation, and conversational depth. PEARL provides single-turn dialogues with structured ground truth metadata, while ReDial contains multi-turn conversations with labeled user sentiment.

Structure: PEARL consists of one-turn recommendation prompts, whereas ReDial captures full conversations between a seeker and a recommender.

Annotations: PEARL includes detailed metadata for each recommended movie (e.g., genre, cast), while ReDial provides feedback tags such as ”Liked”, ”Seen”, or ”Not Seen”.

Scale: PEARL has 50k training examples, while ReDial includes 10,006 dialogues totaling over 180,000 utterances.

User context: PEARL explicitly encodes user personas (likes and dislikes); ReDial infers preferences through dialogue history.

These differences imply that models trained on each dataset may learn different strategies: PEARL favors structured, metadata-driven generation, while ReDial requires modeling dialogue flow and sentiment dynamics.

3. Methodology

3.1. Dataset comparison

3.1.1. Models

We fine-tuned three open-source language models using the Unsloth framework, which offers an efficient interface for parameter-efficient fine-tuning (PEFT). All models were trained using the same configuration to ensure fair comparison and reproducibility:

- **Base Models**:
  - **LLaMA** 3.2 3B Instruct
  - **Qwen** 2.5 3B Instruct
  - **DeepSeek** R1 Distill Llama 8B
- **Fine-tuning Strategy**:
  We employed **LoRA** (Low-Rank Adaptation) (Hu et al., 2021) for PEFT with the following configuration:
  - **Rank**: 16
  - **Alpha**: 16
  - **Dropout**: 0.0
  - **PEFT bias**: none
  - **Gradient checkpointing**: enabled (via Unsloth)

All models were fine-tuned using supervised learning (SFT) via Hugging Face’s SFTTrainer, with a maximum input length of 2048 tokens and a fixed random seed (42) for reproducibility.

3.1.2. Chat Template

We fine-tuned and evaluated the models separately on PEARL and ReDial.

To ensure comparability between datasets and compatibility with chat-tuned models, dialogue histories were reformatted into a consistent prompt-response format using the following template:

```
{ "role": ..., "content": ... }
```

This template was treated as part of the model setup and applied uniformly in an end-to-end manner across all models, regardless of their original formatting conventions.

Roles are either `user`, which contains the input utterance, or `assistant`, which contains the model’s response.

The key distinction between the two datasets lies in the presence of a special `system` role in PEARL, which appears only at the beginning of the dialogue. This message encodes user preferences—such as `Likes`, `Dislikes`, and `Seen`—and was preserved in the prompt to support context-aware recommendations based on prior user interests.

3.1.3. Evaluation Protocol

- **Sampling**: A random 20% of the test split was selected using `train_test_split()` from **scikit-learn**, with `random_state=42` to ensure reproducibility.
- **Metrics**: Evaluation was conducted only after training, using this held-out test set. No validation metrics were collected during training.
- **Tools**: The following libraries were used for evaluation:
  - `transformers` and `evaluate` (Hugging Face)
  - `rouge-score`, `nltk` (for BLEU and tokenization)
  - `bert-score` (for semantic similarity)

Evaluation focused on the models’ ability to generate relevant, diverse, and fluent recommendations. All metric results are reported in the Results section.

For each dialogue in the test set, we extract the dialogue history by selecting all but the last two messages:

$$\texttt{context} = \texttt{messages[:-2]}$$

The model is then conditioned on this context and prompted to generate the second-to-last message, `messages[-2]`, which typically corresponds to the final recommendation made by the assistant.

Evaluation metrics are computed by comparing the generated message with the ground-truth reference `messages[-2]`. The final message in the dialogue, `messages[-1]`, usually the user's response to the recommendation, is excluded from both input and evaluation.

For metrics that focus solely on the recommended movie titles rather than the full utterance—such as Recall@1—we applied a **two-step process**: **(1)** a title extraction function parsed assistant messages to identify candidate movie mentions; and **(2)** a title normalization function standardized these titles to ensure consistent comparison between references and predictions.

ReDial-specific Evaluation: For the ReDial dataset, the evaluation procedure differs due to the structure of its dialogues. Unlike PEARL, where messages alternate strictly between `user` and `assistant`, ReDial exhibits a more irregular turn-taking structure. As such, it is not always clear which assistant message corresponds to the primary recommendation.

To address this, we iterate backwards through each dialogue to locate the most recent `assistant` message that contains a valid movie recommendation.

The model is then prompted with the dialogue up to that point (`context`) and tasked with generating a response. Evaluation is performed by comparing the generated output with the retrieved `reference` message.

This targeted approach allows the evaluation to focus on recommendation-specific turns, better aligning with the task's objectives.

3.2. Uncertainty

3.2.1. Uncertainty Problem Formulation

Denote $\mathcal{C}(X) = \{C^{(1)}, \ldots, C^{(K)}\}$ as the set of K clarifications generated for the user prompt X, where each clarification is a reformulation of the original prompt that preserves its intended meaning.

For example, given the following user prompt:

> "I prefer movies with a more intense and action-packed storyline, with a focus on characters who have a strong sense of justice and courage."

One clarification could be the following:

> "I'm seeking a movie with an intense, action-packed storyline featuring characters driven by justice and courage."

Also denote $T^{(k)} \in \mathcal{T}$ as the movie title recommended by the model for each clarification k, and define the empirical distribution of titles as:

$$p(t) = \frac{1}{K} \sum_{k=1}^{K} \mathbf{1}\left[T^{(k)} = t\right]$$

Aleatoric uncertainty is measured using *Shannon entropy*:

$$\hat{U}_{\text{alea}}(X) = H(T) = -\sum_{t \in \mathcal{T}} p(t) \log p(t) \tag{1}$$

The main idea with (1) is that the greater the diversity of titles across clarifications, the higher the value of $H(t)$, and consequently, the greater the intrinsic ambiguity of the generated responses. Ideally, we would expect our models to recommend the same movie in different clarifications of a prompt, ensuring consistency in their recommendations.

Epistemic uncertainty is estimated by:

$$\hat{U}_{\text{epi}}(X) = -\frac{1}{K} \sum_{k=1}^{K} \bar{\ell}\left(Y^{(k)}\right) \tag{2}$$

where $Y^{(k)} = (y_1^{(k)}, \ldots, y_{N_k}^{(k)})$ is the model output (e.g., tokens of the response) for clarification k, and $\bar{\ell}$ denotes the average log-probability per token.

A large $\hat{U}_{\text{epi}}$ implies that, even after disambiguating X, the model assigns low probability (i.e., has little confidence) to its own output. This provides an indication of how well the model has learned its domain and how confident it is in its responses.

3.2.2. Prompts, filtering and samples for uncertainty

A total of 200 ambiguous questions were sampled from the PEARL dataset. These questions were selected by identifying prompts containing ambiguous words such as: *any, something, maybe, whatever, similar, either, idk, recommend,* and *like.*

Then, using the *LLaMA3 Instruct* model version, 4 clarifications are generated for each user prompt X. The clarification model is provided with the following prompt:

```
You are Clarify-Bot.
TASK: Rewrite ONLY the seeker's
last message so it has ONE clear
meaning.
RULES:
1.  NO greetings, questions,
pipes (|), or bullet lists.
2.  NO movie titles, years, or
recommendations.
3.  Single sentence, 6--40
words.
```

```
Return JUST the rewritten
sentence.
```

After obtaining the clarifications, we will focus on calculating the uncertainty of the three previously used models: DeepSeek, Qwen, and LLaMA 3. All three were previously fine-tuned using the PEARL dataset. We provide to each of these models as input the 800 questions clarifications (200 ambiguos questions $\times$ 4 clarifications), storing each response's log-probability and generated text.

Finally we extract the first movie title contained in the answer, and compute the $\mathbb{E}[\hat{U}_{\text{alea}}]$ for each model. We do the same with $\mathbb{E}[\hat{U}_{\text{epi}}]$.

4. Results

4.1. Dataset comparison

METRIC	LLAMA	QWEN	DEEPSEEK
BLEU SCORE	0.5521	0.5201	0.5544
ROUGE-1	0.4982	0.4710	0.5018
ROUGE-2	0.2500	0.2232	0.2524
ROUGE-L	0.3707	0.3516	0.3754
DISTINCT-1 INTRA	0.7938	0.81462	0.7952
DISTINCT-1 INTER	0.0808	0.0737	0.0786
DISTINCT-2 INTRA	0.9762	0.9843	0.9762
DISTINCT-2 INTER	0.2776	0.2482	0.2702
BERTSCORE F1	0.8631	0.8562	0.8640
BERTSCORE PRES.	0.8652	0.8584	0.8662
BERTSCORE REC.	0.8614	0.8544	0.8621
RECALL@1	0.013	0.009	0.004
NOVELTY	0.4452	0.4744	0.4402
SELF-BLEU	0.6808	0.7255	0.6900

Table 1. Metrics for each model independently trained only with PEARL

METRIC	LLAMA	QWEN	DEEPSEEK
BLEU SCORE	0.1144	0.1196	0.1293
ROUGE-1	0.1176	0.1250	0.1128
ROUGE-2	0.0274	0.0291	0.0294
ROUGE-L	0.1105	0.1177	0.1059
DISTINCT-1 INTRA	0.9704	0.9642	0.9617
DISTINCT-1 INTER	0.1726	0.1649	0.1468
DISTINCT-2 INTRA	0.9554	0.9705	0.9558
DISTINCT-2 INTER	0.4287	0.4145	0.3659
BERTSCORE F1	0.6937	0.6936	0.6949
BERTSCORE PRES.	0.7062	0.7044	0.7046
BERTSCORE REC.	0.6832	0.6843	0.6866
RECALL@1	0.013	0.019	0.022
NOVELTY	0.7577	0.7621	0.7731
SELF-BLEU	0.5323	0.5616	0.6131

Table 2. Metrics for each model independently trained only with ReDial

4.1.1. PEARL RESULTS (TABLE 1)

DeepSeek records the highest scores in BLEU (0.5544) and across all ROUGE variants, indicating a strong lexical match between its output predictions and the reference recommendations. LLaMA performs similarly but slightly behind, while Qwen the worst.

DeepSeek records the lowest Recall@1 score (0.004), meaning it often fails to include the correct movie among its top prediction. This may indicate to a tradeoff between producing fluent text and making accurate recommendations.

All three models achieve high BERTScore F1 values (above 0.85), with DeepSeek slightly ahead at 0.8640. The differences here are minimal, suggesting that all models produce semantically coherent responses.

Regarding the variety of tokens used, Qwen obtains an intra-sentence Distinct-1 score (0.8146). However, its lower inter-sentence diversity suggests it tends to repeat similar phrasing across different outputs.

Qwen achieves the highest novelty and self-BLEU scores, suggesting its responses are more creative but also somewhat repetitive. DeepSeek, by contrast, maintains a more consistent balance between novelty and repetition.

4.1.2. REDIAL RESULTS (TABLE 2)

Throughout all models, the results show a noticeable drop in performance across all evaluated metrics when compared to training on PEARL. This decline is the most evident in BLEU, ROUGE, and BERTScore values, reflecting the less structured nature of ReDial dialogues.

DeepSeek once again performs the best. It obtains the highest BLEU (0.1293) and BERTScore F1 (0.6949) among the three models, suggesting it maintains strong lexical and semantic alignment.

Recall@1 values shift in this setting. While DeepSeek had the lowest Recall@1 on PEARL, it achieves the highest score on ReDial (0.022). This suggests that DeepSeek may be more effective at understanding the unstructured dialogue.

Diversity metrics also show higher values compared to PEARL. Distinct-1 and Distinct-2 increase across all models, likely due to shorter, more varied responses in ReDial's conversations.

Overall, these results highlight how dataset structure influences model performance, with DeepSeek adapting better to ReDial's less rigorous format. However, all models struggle to match their performance on PEARL, indicating room for improvement in handling open-domain, dialogue-heavy recommendation contexts.

4.2. Uncertainty

The two columns in Table 3 capture *distinct* kinds of uncertainty in the recommendations. The quantity $\mathbb{E}[\hat{U}_{\text{alea}}]$ is the average aleatoric uncertainty: the Shannon entropy of the movie titles recommended across all clarifications of the same user turn. A high value means the model changes its suggestion whenever the prompt is rephrased, revealing lingering input ambiguity. By contrast, $\mathbb{E}[\hat{U}_{\text{epi}}]$ is our proxy for epistemic uncertainty, it is the mean negative token–log-probability, so larger numbers indicate that, after the input is clarified, the model still assigns low probability to its own output, i. e. it lacks confidence in the knowledge it draws on.

Modelo	$\mathbb{E}[\hat{U}_{\text{alea}}]$	$\mathbb{E}[\hat{U}_{\text{epi}}]$
LLaMA-3	1.858	46.376
DeepSeek	0.384	106.868
Qwen-2	0.764	29.090

Table 3. Aleatoric and epistemic uncertainty for each model.

DeepSeek shows the *lowest* aleatoric uncertainty (0.384 bits), signalling strong internal consistency: given a clarified prompt, it tends to recommend the same title every time. Yet it also exhibits the *largest* epistemic uncertainty (106.868), so the model is unsure of the very answer it repeats. Qwen-2 sits at the other extreme: it achieves the *lowest* epistemic uncertainty (29.090), making it the most self-confident system, while its aleatoric uncertainty is moderate (0.764), meaning it still offers several valid alternatives under rephrasings. LLaMA-3 lands in between: it produces the greatest diversity of titles (1.858 bits) but with a middle-of-the-road confidence level (46.376).

We can infer that if reliable confidence is paramount (e. g. risk-sensitive decision support) Qwen is preferable. Where traceable, deterministic output is required, DeepSeek provides stability but should be paired with external validation. Whenever breadth of suggestions is more valuable, such as exploratory recommendation scenarios,LLaMA-3 offers the richest set of plausible movies while keeping an acceptable level of self-trust.

5. Discussion and Conclusions

One consistent pattern across both datasets is the strong performance of DeepSeek across semantic and lexical similarity metrics. A likely explanation for this advantage lies in its larger parameter count (8B), compared to the 3B LLaMA and 3B Qwen models. This increased capacity likely enables better modeling of user preferences and dialog coherence, especially in the longer, context-rich conversations found in PEARL.

However, this increased fluency and diversity does not always translate to stronger recommendation accuracy. For instance, DeepSeek's Recall@1 on PEARL is the lowest among all models (0.004), despite its superior BLEU and BERTScore metrics. This highlights a potential trade-off between generating coherent, fluent responses and producing highly relevant recommendations.

Further analysis would be required to isolate the effect of model size from other factors such as architectural differences. Comparing all models at 3B would provide a clearer attribution of performance differences.

Considering the results obtained with both datasets highlight the importance of dataset characteristics in shaping model behavior. PEARL's format rewards models that can integrate long-term user profiles into fluent recommendations, whereas ReDial's less regular dialogue turns favor models capable of making high-precision predictions from sparse cues.

In the case of uncertainty experiments, computational capacity is vital to obtaining results that allow for reliable conclusions about each model's trustworthiness. This is very difficult to achieve using Google Colab's free tier, which was used in the present study.

First, the quality of the clarification generation depends on having a capable clarification model, with a sufficient number of parameters and an appropriate temperature setting, to produce clarifications that are diverse in grammar and structure while preserving meaning. The same applies to the models' responses to the clarifications: more parameters typically lead to richer and more informative outputs. We believe this aspect was limited by our access to models with a maximum of 8 billion parameters, which falls short compared to the larger models commonly used in industry and academic research.

The limitation is not only related to model size. Our study was also constrained to generating a maximum of four clarifications per user prompt due to the inference time required for generating both the clarifications and the corresponding model responses. We believe that this type of research must have access to a large infrastructure in order to present complete and generalizable results.

Acknowledgements

References

Hou, B., Liu, Y., Qian, K., Andreas, J., Chang, S., and Zhang, Y. Decomposing uncertainty for large language models through input clarification ensembling, 2024. URL https://arxiv.org/abs/2311.08718.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Kim, M., Kim, M., Kim, H., woo Kwak, B., Chun, S., Kim, H., Kang, S., Yu, Y., Yeo, J., and Lee, D. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset, 2024. URL https://arxiv.org/abs/2403.04460.

Kiureghian, A. D. and Ditlevsen, O. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. doi: https://doi.org/10.1016/j.strusafe.2008.06.020. URL https://www.sciencedirect.com/science/article/pii/S0167473008000556. Risk Acceptance and Risk Communication.

Li, R., Kawale, J., and McAuley, J. Towards deep conversational recommendations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 565–575, 2018.

A. Uncertainty Clarification Examples

Example Input and Clarifications from Dialogue #31069

Original Input:

”John Wick: Chapter 2” (2017) has great action, but the main protagonist’s motivation was less interesting to me and some action scenes felt far-fetched and unrealistic. I’m looking for a movie with more relatable characters and a less predictable plot. Do you have any other recommendations that align with those criteria?

Generated Clarifications:

- I seek a movie with relatable characters and a less predictable plot, where action scenes are grounded in reality.
- The main character’s motivations were less compelling and some action scenes felt unconvincing, leading me to seek a film with more believable characters and an unpredictable plot under 70 minutes.
- The main character’s motivations were less compelling, and some action sequences felt implausible; I seek a film with more believable characters and an unpredictable storyline.
- I’d like a movie with relatable characters and an unpredictable plot that avoids action scenes that feel far-fetched and unrealistic.

B. Uncertainty Clarification Examples

Qwen Example Answers to Clarification from Dialogue #31069 (Set 2)

Answer 1:

1. *The Bourne Ultimatum* (2007) This movie has relatable characters and a less predictable plot. The action scenes are grounded in reality, and the realistic portrayal of espionage and action sequences makes for an exciting and believable experience.

Answer 2:

Based on your preference for believable characters and an unpredictable plot, I would recommend *The Prestige* (2006). It has engaging character study, a complex plot, and excellent storytelling. The main characters’ motivations are compelling, and the film keeps you guessing until the end. Plus, it’s under 2 hours long, so it fits your time preference.

Answer 3:

I recommend *The Big Lebowski* (1998) because it has a unique and unpredictable storyline with well-developed characters. The characters’ motivations are very compelling, and the humor in the film is quite entertaining. Plus, it’s a great comedy and mystery movie that you should definitely watch.

Answer 4:

The Big Lebowski (1998): Quirky and offbeat characters, Unpredictable and darkly comedic plot, Avoids unrealistic action scenes.

The Adventures of Buckaroo Banzai Across the 8th Dimension (1984): Quirky and offbeat characters, Unpredictable and entertaining plot, Avoids unrealistic action scenes

The Princess Bride (1987): Quirky and offbeat characters, Unpredictable and entertaining plot, Avoids unrealistic action scenes

Qwen Example Answers to Clarification from Dialogue #31069

Answer 1:

89 minutes — Crime, Drama — I think you would really enjoy ”The Gondoliers” (1966). It has a fantastic cast with charismatic performances, especially from Jane Fonda and Richard Burton. The plot is relatable and grounded in reality, with some action scenes that are well-executed.

Answer 2:

How about giving ”The Mummy” a try? It has some funny moments and an unusual plot, and it’s under 70 minutes. The actors, especially Ben Affleck, deliver great performances. It’s definitely worth a watch if you’re in the mood for a fun and entertaining film.

Answer 3:

1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 1997 — — 199

Answer 4:

for you. ”(I’m sorry, but I can’t provide a recommendation for this movie. I need to know what movie you’re looking for in order to give you a recommendation.)