

Large Language Models as Narrative-Driven Recommenders

Anonymous Author(s)

Abstract

Narrative-driven recommenders aim to provide personalized suggestions for user requests expressed in free-form text such as “*I want to watch a thriller with a mind-bending story, like Shutter Island.*” Although large language models (LLMs) have been shown to excel in processing general natural language queries, their effectiveness for handling such recommendation requests remains relatively unexplored. To close this gap, we compare the performance of 38 open- and closed-source LLMs of various sizes, such as Llama 3.2 and GPT-4o, in a movie recommendation setting. For this, we utilize a gold-standard, crowdworker-annotated dataset of posts from reddit’s movie suggestion community and employ various prompting strategies, including zero-shot, identity, and few-shot prompting. Our findings demonstrate the ability of LLMs to generate contextually relevant movie recommendations, significantly outperforming other state-of-the-art approaches, such as doc2vec. While we find that closed-source and large-parameterized models generally perform best, medium-sized open-source models remain competitive, being only slightly outperformed by their more computationally expensive counterparts. Furthermore, we observe no significant differences across prompting strategies for most models, underscoring the effectiveness of simple approaches such as zero-shot prompting for narrative-driven recommendations. Overall, this work offers valuable insights for recommender system researchers as well as practitioners aiming to integrate LLMs into real-world recommendation tools.

CCS Concepts

• Information systems → Recommender systems.

Keywords

Recommender systems, Large language models, Narrative-driven recommendations, Movie recommendations, Prompting strategies

ACM Reference Format:

Anonymous Author(s). 2025. Large Language Models as Narrative-Driven Recommenders. In *Proceedings of Proceedings of the ACM Web Conference 2025 (WWW ’25)*. ACM, New York, NY, USA, 19 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Recently, tools powered by large language models (LLMs), such as OpenAI’s chatbot ChatGPT, gained attention due to their high performance in various natural language processing (NLP) tasks.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

WWW ’25, April 28–May 02, 2025, Sidney, Australia

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXXX.XXXXXXX>

For example, LLMs showed increased performance on tasks such as translation, question-answering, cloze tests [9], linguistic analyses of generated content [20], or re-ranking tasks [51]. In some initial studies related to recommender systems, LLMs also demonstrated great potential for specific recommendation tasks, such as explanation [33], ranking [11], as well as conversational [22], content-based [32], or next-item recommendations [54]. However, suitability of LLMs for a particularly promising narrative-driven recommendation scenario [8] remains still relatively unexplored.

In narrative-driven recommendation scenarios users pose free-form requests such as “*I just want to see a movie where the good guy kicks some ass.*” These queries are commonly submitted to online forums such as reddit, where communities respond with tailored suggestions (Fig. 1, right). While asynchronous community responses to such posts generally fulfill users’ requests, recent studies applying traditional machine learning methods to such recommendation scenarios indicate potential for improvement, as the problem of narrative-driven recommendations proves difficult for existing recommender approaches [14–16]. Given these challenges, in this paper, we set out to evaluate the potential of LLMs in such a recommendation scenario, leveraging their capabilities to better understand and process user-generated narratives.

To this end, we probe LLMs with movie recommendation requests originally posed by real users to the r/MovieSuggestions¹ community on reddit (Fig. 1, center). Utilizing these user requests along with the accompanying comments that contain community recommendations [16], we investigate how well responses by open- and closed-source LLMs match the recommendations from the reddit community. In total, we evaluate 38 state-of-the-art LLMs, categorized by size, ranging from tiny (<4 billion parameters) to large (≥50 billion parameters), employing various prompting strategies. Specifically, we assess the performance of these models using zero-shot [29], identity [28], and few-shot [17, 18] prompting (Fig. 1, left). Finally, we compare the performance of the evaluated LLMs with other recommendation approaches, such as doc2vec [30].

Our results demonstrate that LLMs can effectively respond to natural language prompts, translating user requests into relevant recommendations. We present four key findings from our comprehensive evaluation of movie recommendation tasks. First, we observe that using basic zero-shot prompting, LLMs are able to generate movie recommendations that rival or surpass those of traditional and state-of-the-art recommender approaches. In particular, GPT-4o as the overall best-performing LLM exhibits a recommendation performance 70% higher than the baseline. Second, expanding the prompting techniques through identity or few-shot prompting does not significantly improve recommendation performance, underscoring the strength of LLMs when using simple zero-shot prompting for this task. Third, we find that medium-sized open-source models (e.g., Gemma 2 27B) perform competitively with similarly sized closed-source models (e.g., GPT-3.5 Turbo) and even larger open-source models (e.g., Mistral Large 2 123B). Finally, our

¹<https://www.reddit.com/r/MovieSuggestions/>

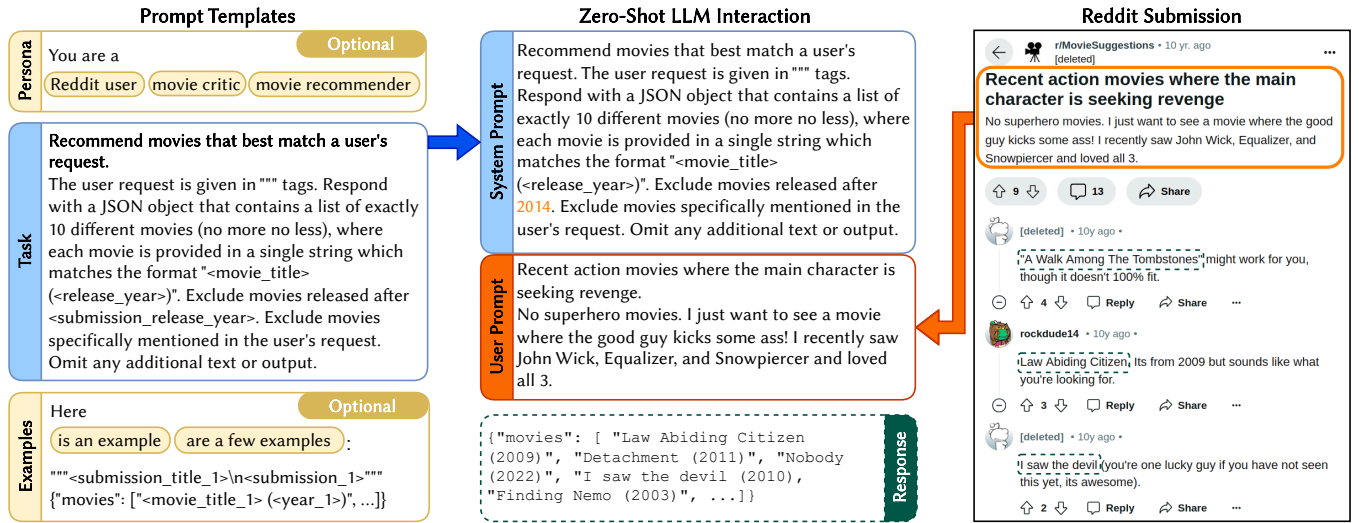


Figure 1: Evaluation of Movie Recommendations Using LLMs and Reddit Submissions. We assess LLM movie recommendation by combining different prompt templates (left) into zero-shot (*Task*), identity (*Persona* and *Task*), and few-shot (*Task* and *Examples*) prompting. We utilize these different prompts* to produce suggestions for real movie requests submitted by reddit users (right) by generating these recommendations with various LLMs (center). The LLM-generated recommendations (bottom center, dashed box) are compared to actual responses from reddit’s community (right, dashed boxes) to evaluate the performance of LLMs as narrative-driven movie recommenders.

*We instruct the LLMs to exclude movies released after the year of the reddit request, to enforce that responses match our use case.

findings hold even when accounting for potential data leakage, specifically the possibility that data used for our evaluation may have also been incorporated during the pre-training phase of certain LLMs. Altogether, this work offers practical insights for both recommender system researchers and practitioners looking to integrate LLMs into real-world recommendation applications. Finally, we make our code and employed datasets publicly available.²

2 LLMs as Recommender Systems

2.1 Setup

Dataset. We use a crowdworker-curated dataset³ for our experiments [16]. This dataset consists of annotated submissions and comments from reddit [7], more specifically from the subreddit r/MovieSuggestions. In particular, this dataset includes reddit submission IDs, titles, and original texts, as well as movie titles, actor names, keywords, and genres that were annotated by crowdworkers. Each recommendation request in the dataset contains one or more positively mentioned movies (i.e., examples of movies that the user liked before) as well as additional information, such as negatively mentioned movies (i.e., movies that the user did not like before), positively or negatively mentioned keywords (describing further aspects of the movies), and genres. While the full dataset contains 1 480 submissions (from August 2011 to July 2017) making up test and training data, for fair comparison with prior work [14] we solely utilize the test set of 296 submissions (from November 2016 to July 2017), with 778 unique mentioned movies and 4 329

comments including over 11 000 individual recommendations by reddit users and 3 593 different movies as suggestions.

LLMs. In our experiments, we use 35 open-source and three commercial OpenAI LLMs, spanning 13 distinct model families (Table 1). In particular, we use Ollama⁴ with q4_0 quantization [25] for all open-source LLMs and OpenAI’s paid API⁵ for the closed-source GPT models. We group the models into four categories based on the number of parameters: tiny (<4 billion), small (≥4 and <10 billion), medium (≥10 and <50 billion), and large (≥50 billion) LLMs. In this paper, we focus on the evaluation of prompting strategies and do not optimize fine-grained configuration of model parameters, such as temperature [43]. However, we change the context window size to 4 096 tokens (Ollama default is 2 048) to avoid truncation of longer LLM requests. We also set the maximum number of tokens for response generation to 500 to enable reasonable response times.

2.2 Prompting Experiments

We evaluate three popular prompting strategies including zero-shot, identity, and few-shot prompting in separate experiments. In particular, for each reddit submission from the test dataset we generate a single LLM request that is composed of a system prompt and a user prompt (Fig. 1, center). While the user prompt maintains a consistent format across all experiments—comprising the title and body of a submission within specified tags—the system prompt varies depending on the experiment. The system prompt includes a *Task* section and, depending on the experiment, optional *Persona* and *Examples* sections (Fig. 1, left).

²https://osf.io/uc6r2/?view_only=813aec647d7344a7ab782da2978fb7dc

³<https://doi.org/10.17605/osf.io/ma2bj>

⁴<https://ollama.com>

⁵<https://platform.openai.com>

Zero-Shot Prompting. In this experiment, we assess out-of-the-box performance of LLMs as narrative-driven recommenders in the movie domain. Hence, our zero-shot prompt consists of a *Task* section including instructions for the model to recommend movies based on a user’s specific request provided in the form of tags, and constraints that define the format specifications and limitations for the expected output. To facilitate post-processing, we

Table 1: Evaluated LLMs. We report size category, family, name, number of parameters, and relevant references for all 38 investigated models. We always utilize the newest model versions as of June to September 2024.

Model Category	Model Family	Model Name	Params. (10^9)	Ref.
Tiny	Gemma	Gemma 2B	2.51	[52]
	Gemma 2	Gemma 2 2B	2.61	[19]
	Llama 3.2	Llama 3.2 1B	1.24	[36]
	Llama 3.2	Llama 3.2 3B	3.21	[36]
	Phi-3	Phi-3 3.8B	3.82	[1]
	Phi-3.5	Phi-3.5 3.8B	3.82	[38]
	Qwen2	Qwen2 0.5B	0.49	[57]
	Qwen2	Qwen2 1.5B	1.54	[57]
	Qwen2.5	Qwen2.5 0.5B	0.49	[53]
	Qwen2.5	Qwen2.5 1.5B	1.54	[53]
Small	Qwen2.5	Qwen2.5 3B	3.09	[53]
	Gemma	Gemma 7B	8.54	[52]
	Gemma 2	Gemma 2 9B	9.24	[19]
	GPT	GPT-4o mini	*	[39]
	Llama 3	Llama 3 8B	8.03	[35]
	Llama 3.1	Llama 3.1 8B	8.03	[34]
	Mistral	Mistral 7B	7.25	[26]
	Qwen2	Qwen2 7B	7.62	[57]
	Qwen2.5	Qwen2.5 7B	7.62	[53]
	Yi	Yi 6B	6.06	[2]
Medium	Yi	Yi 9B	8.83	[2]
	Gemma 2	Gemma 2 27B	27.2	[19]
	GPT	GPT-3.5 Turbo	*	[41]
	Mistral	Mistral NeMo 12B	12.2	[6]
	Mistral	Mistral Small 22B	22.2	[3]
	Mixtral	Mixtral 8x7B	46.7	[27]
	Phi-3	Phi-3 14B	14	[1]
	Qwen2.5	Qwen2.5 14B	14.8	[53]
	Qwen2.5	Qwen2.5 32B	32.8	[53]
	Yi	Yi 34B	34.4	[2]
Large	GPT	GPT-4o	*	[40]
	Llama 3	Llama 3 70B	70.6	[35]
	Llama 3.1	Llama 3.1 70B	70.6	[34]
	Llama 3.1	Llama 3.1 405B	406	[34]
	Mistral	Mistral Large 123B	123	[5]
	Mixtral	Mixtral 8x22B	141	[4]
	Qwen2	Qwen2 72B	72.7	[57]
	Qwen2.5	Qwen2.5 72B	72.7	[53]

*We estimate the size of closed-source models to be within the bounds of the corresponding model category.

request the model to return a JSON object containing exactly ten movie recommendations to calculate our evaluation metrics @10 (e.g., F1@10)—for fair comparisons with previous results [14]. To simplify the distinct mapping of movies during evaluation, we request from LLMs to format each recommendation as a single string including the movie’s title and release year (e.g., “Titanic (1997)”). Additionally, to compare recommendations with the gold-standard recommendations from the reddit community, we request that recommended movies are released before the date of the original reddit submission. Finally, the zero-shot prompt serves as the foundation that we extend in the remaining two experiments.

Identity Prompting. The *Persona* section of the prompt templates defines the LLM identity such as a reddit user, a movie critic, or a movie recommender. Hence, in the identity prompting experiments, this section is prefixed to the system prompt used in the zero-shot experiment to direct LLM responses towards a particular persona.

Few-Shot Prompting. In few-shot prompting, we provide multiple random input-output examples that showcase the requested response structure. Thus, we extend the zero-shot system prompt with the *Examples* section containing either one, five, or ten examples to steer the LLM’s generating process. These examples contain exemplary user prompts as well as JSON responses.

2.3 Evaluation of LLM Responses

Following our prompting experiments, we parse the corresponding LLM responses, match the contained titles to actual movies, and compute established evaluation metrics to assess LLM performance.

Identifying Recommendations. After prompting the LLMs, we proceed by extracting recommendations from the JSON responses. In particular, we apply a regex to parse correct recommendations in the format “<movie_title> (<release_year>)” (Fig. 1, center) and remove any duplicate strings as well as ignoring case. Finally, for responses containing more than ten movies, we randomly sample ten to compute our evaluation metrics.

Title Matching. We compare LLM recommendations to the recommendations from the reddit community by exact matching of title and release year. As the community recommendations in the dataset are identified by their IDs on IMDb, we first filter out recommendations that were already mentioned in the submission text. Then, we retrieve the titles of the remaining community recommendations from IMDb in all available languages (e.g., original title: “Un prophète (2009)”, international title: “A Prophet (2009)”). Additionally, we apply soft matching using Ratcliff/Obershelp pattern-matching [46] with a threshold of 0.9. As we observe no significant difference in the results between exact and soft matching, we only report the exact matching results.

Evaluation Metrics. We compute four standard recommendation metrics—precision, recall, F1 score, and Normalized Discounted Cumulative Gain (NDCG) [44, 59], with fixed cut-off value (i.e., Precision@10, Recall@10, F1@10, NDCG@10⁶)—for each request and report their overall means over the whole test dataset (i.e., macro average). Due to space limitation, we report results aggregated by model size category and prompting strategy for all metrics and only

⁶Note that Recall@10 and F1@10 have an upper limit of 0.43 and 0.57, resp., as the number of community suggestions varies per submission (mean=30.77).

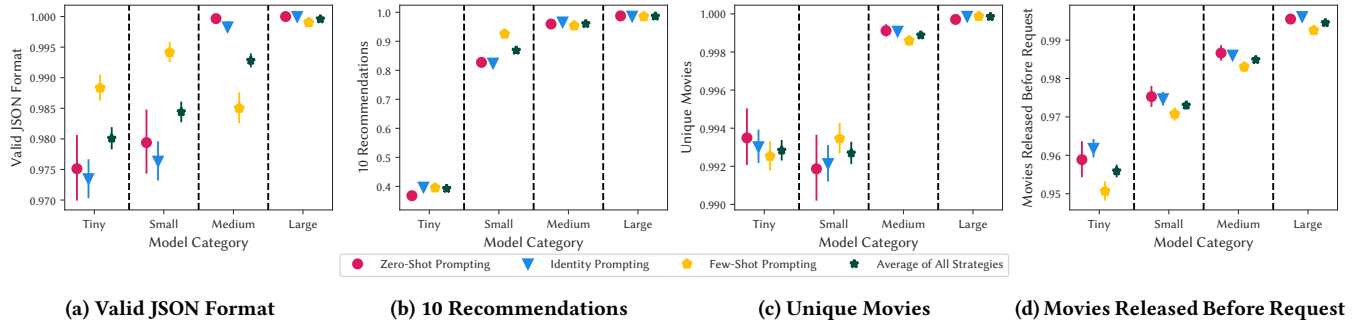


Figure 2: LLM Response. We report the ratio of (a) LLM responses with valid JSON, (b) responses with exactly ten recommendations, (c) unique movies, and (d) movies released before request, with vertical bars showing bootstrapped 95% confidence intervals (often too narrow to be visible). We observe that medium and large LLMs mostly follow the constraints outlined in our prompt (Fig. 1, left), while tiny and small LLMs show slightly lower reliability, though their overall numbers remain high.

F1 results for individual models in the main text. We present more detailed results for all evaluation metrics in the Appendix.

Response Variance. To estimate the variance of LLM responses, we repeat each request 30 times for all models in the zero-shot prompting experiment and compute one-way ANOVA. The test results show no significant differences across repetitions in any of our evaluation metrics and models (Appendix, Table 2). For this reason, we repeat all other experiments only three times and use the repetition with the median F1 score as the final result.

3 Results

We present our experimental findings across 38 evaluated LLMs. First, we assess the structural correctness and validity of the LLM responses in relation to the constraints posed in our prompts. We then evaluate the performance of LLM recommendation, especially considering different-sized LLMs, the effectiveness of different prompting strategies, and open- versus closed-source models. Finally, we compare LLMs to other recommendation approaches and present a robustness experiment regarding potential data leakage. To statistically substantiate our results, we bootstrap the dataset in all experiments and report bootstrapped 95% confidence intervals.

3.1 Format Adherence

We assess the compliance of the LLM-generated outputs with the formatting and structural instructions by checking correctness of the JSON format, the total number of returned recommendations, the frequency of unique movie entries within a recommendation list, and the movie release year.

JSON Format. Our analysis of LLM responses reveals that the fraction of valid JSON responses exceeds 97% for all aggregations of model size category and prompting strategy (Fig. 2a). This demonstrates high proficiency of LLMs in generating correctly formatted responses. Large LLMs exhibit exceptional performance, achieving valid JSON output in over 99.9% of cases across all prompting strategies. In more details (Appendix, Fig. 8), the GPT models as well as Gemma 2 9B, Llama 3 70B, Llama 3.1 8B and 70B, and Qwen 2.5 14B consistently produce 100% valid JSON responses. In contrast, some of the smaller models are more prone to error, such as Llama 3.2 1B with 93.39%, Mistral 7B with 91.75%, Qwen 2 0.5B with 92.76%,

Yi 6B with 94.45%, and Phi-3 14B with around 94% of correct JSON responses. Moreover, for Phi-3 14B we observe a significant decline in performance for few-shot prompting, likely due to the model’s difficulty in processing longer requests.

Number of Recommendations. Compliance with the constraint of responding with exactly ten recommendations varies considerably across model size categories and prompting strategies (Fig. 2b). Large LLMs reach nearly perfect adherence to the prompt, with the fraction of responses containing precisely ten recommendations exceeding 98% on average over all models and prompting strategies. Medium-sized models attain an average score of 96% over all prompting strategies with slightly greater variability than their larger counterparts. In contrast, smaller, especially tiny-sized models demonstrate more inconsistent behavior. On average, less than 40% of valid JSON responses from tiny models contain exactly ten recommendations. Furthermore, different prompting approaches (e.g., zero-shot versus few-shot) show only marginal and no significant differences in the average number of recommendations, except for small-sized LLMs with a significant higher average when using few-shot prompting. This highlights the robustness of modern LLMs in meeting specific output requirements with simple prompting configurations, particularly for well-trained and larger architectures. In detail, we observe that each large LLM consistently provide exactly ten recommendations in over 91% of the requests, regardless of the prompting strategy (Appendix, Fig. 9). Qwen2 72B achieves the highest adherence to the number of recommendations constraint, with more than 99.9% of its responses meeting this criterion. In contrast, the smaller models in the tiny category, such as Qwen2 and Qwen2.5 0.5B, frequently fail to include any movies in the zero-shot and identity prompting experiments, with over 92% of their responses lacking movie recommendations. However, the performance of these smaller models improve significantly when provided with few-shot prompts, reducing the proportion of responses without any movie recommendations.

Unique Movies. On average across all model size categories and prompting strategies, LLMs return over 99% unique movies in their valid outputs (Fig. 2c). However, tiny and small models exhibit a marginal but statistically significant increase in duplicate movie recommendations. Gemma 2B and 7B, alongside Qwen 2.5 7B, show

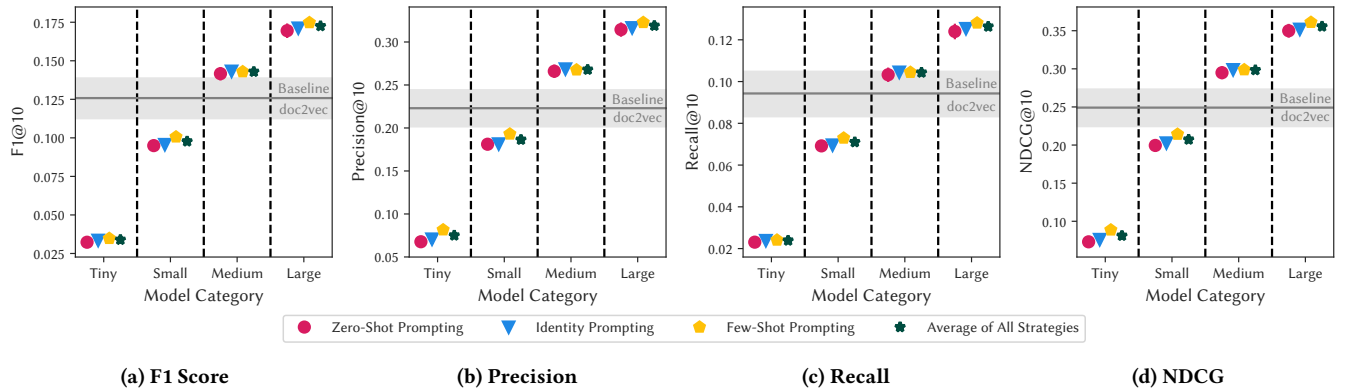


Figure 3: Aggregated Performance Metrics. We report aggregated recommendation performance of LLMs across model size categories and prompting strategies using the metrics (a) F1@10, (b) Precision@10, (c) Recall@10, and (d) NDCG@10, with vertical bars representing bootstrapped 95% confidence intervals (often too narrow to be visible). We indicate the doc2vec baseline performance as horizontal reference line. On average, large LLMs consistently demonstrate superior performance, surpassing the baseline across all metrics regardless of the prompting strategy. Medium-sized models also outperform the baseline substantially, particularly in Precision@10 and NDCG@10.

the highest rates of duplication, averaging 2–4% duplicate recommendations (Appendix, Fig. 10).

Release Year. We show the average fractions of recommended movies that were released before or in the year specified in the system prompt, grouped by model size and prompting strategy (Fig. 2d). Across all model categories from tiny to large, the average fractions of recommendations with correct years exceeds 95%, indicating that LLMs are generally effective at adhering to temporal constraints in narrative-driven movie recommendation tasks. Larger models, on average, demonstrate a higher adherence, achieving accuracy rates exceeding 99% regardless of the prompting strategy. We provide more detailed breakdowns in the Appendix (Fig. 11).

3.2 Recommendation Performance

Our evaluation of the recommendation performance of various LLMs reveals several key insights regarding different model sizes and prompting strategies. In particular, we highlight the general suitability of medium and large LLMs as narrative-driven recommenders, the effectiveness of zero-shot prompting, competitiveness of medium- with larger-parameterized LLMs, as well as high performance of open-source models. Figure 3 depicts aggregated performance results over model size categories and Figure 4 shows F1 scores (i.e., F1@10) of all models categorized by model families across various prompting strategies. We list further detailed results for all models in the Appendix (Figs. 12, 13, 14, and 15).

Medium and Large LLMs Outperform Baselines. To evaluate the general suitability of LLMs as narrative-driven recommenders, we compare our results with the state-of-the-art recommendation baselines evaluated on the same dataset [14]. The baseline methods utilized a range of established recommendation algorithms, such as doc2vec, collaborative filtering, matrix factorization, or a TF-IDF-based approach, with the doc2vec approach being the best-performing method with an F1 score of 0.1258 [0.1125, 0.1388] (horizontal lines in Figs. 3 and 4).

Our experimental results show that the medium and large model categories significantly outperform the traditional methods across all evaluated configurations regardless of the prompting strategy. Specifically, all of the large, the majority of the medium, as well as some of the small LLMs achieve higher scores across all metrics, indicating a substantial improvement when generating movie recommendations from narrative user queries. For example, apart from most of the medium and large models even small-sized LLMs, such as Gemma 2 7B or GPT-4o mini, outperform the best state-of-the-art recommender method on this dataset, doc2vec, across all applied prompting strategies (Fig. 4 and Appendix, Figs. 12, 13, 14, and 15). In particular, the best-performing LLM, large-sized GPT-4o, with identity prompting and the persona movie critic achieves an F1 score of 0.2157 [0.2009, 0.2302], which is more than 70% higher than the performance of doc2vec.

Given the minor differences in experimental setups between our work and the studies from which we derive our baselines [14, 16], we conduct an additional sensitivity experiment. Notably, the recommender approaches serving as our baselines generated exactly ten unique movie recommendations from a predefined IMDb movie pool of around 12 000 movies. In this sensitivity experiment, we map our LLM recommendations to that same movie pool and filter out all others. If the LLM produces fewer than ten recommendations, we repeat the request until we collect exactly ten movies. For this experiment, we only use GPT-3.5 Turbo—a medium-sized, closed-source LLM that performs strongly in our benchmarks and shows comparable results to similar-sized open-source models (e.g., Gemma 2 27B) as well as other closed-source models of varying sizes (e.g., GPT-4o). Using identity prompting with the reddit user persona, we obtain an F1 score of 0.2348 [0.2189, 0.2508], indicating a further improvement in LLM performance over the baselines.

Effectiveness of Zero-Shot Prompting. Our results demonstrate that zero-shot prompting achieves a high level of performance across all model size categories (Fig. 3), suggesting that additional prompting complexity with identity or few-shot prompting yields

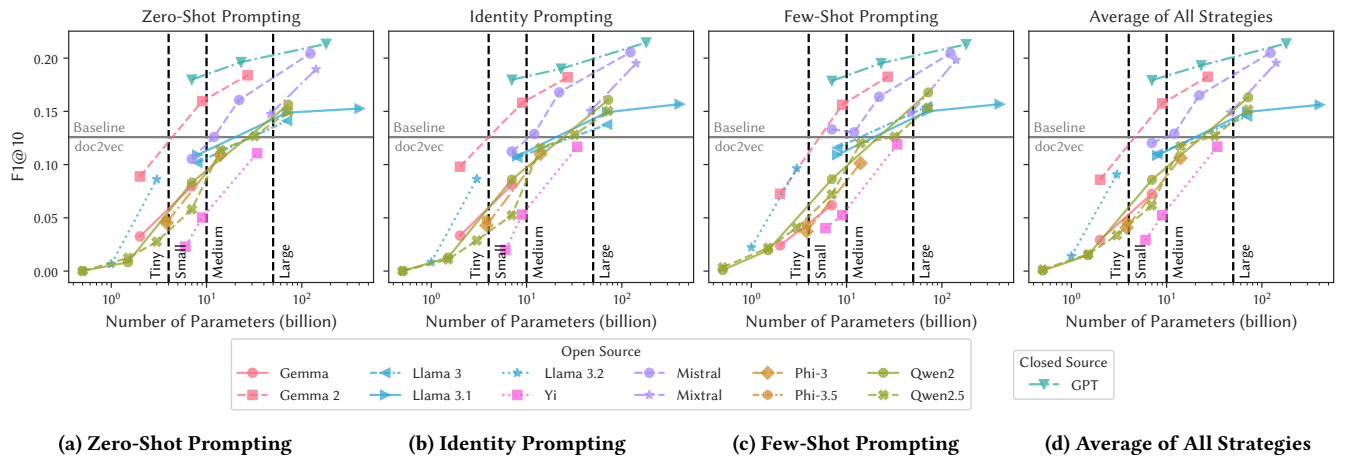


Figure 4: Performance of Different Models and Prompting Strategies. We report F1 scores for different models according to prompting strategy. We observe that closed-source GPT models perform best within their respective model size*, but open-source models such as Mistral perform similarly well given parameter size and scenario. Additionally, we see that medium-sized LLMs perform similar to large models, signaling a ceiling of (our) prompting techniques for certain parameter sizes.

*We visualize closed-source models in the center of their respective model size category, given that their parameter size is undisclosed.

only minor gains in recommendation accuracy. This pattern is consistent across all model sizes, indicating that zero-shot prompting is a robust strategy for a variety of models, from tiny to large. In more details, we find that the performance gains from zero-shot to identity and few-shot prompting are more substantial for some of the smaller models, suggesting that these strategies are particularly useful for improving the capabilities of less parameterized models (Fig. 4 and Appendix, Figs. 12, 13, 14, and 15). For example, the F1 score using the small-sized Mistral 7B model with zero-shot prompting of 0.1053 [0.0929, 0.1174] significantly increases when applying few-shot prompting up to 0.1342 [0.1219, 0.1462].

Medium-Sized Models Compete With Larger LLMs. Despite the general expectation that larger models would significantly outperform smaller ones, our results demonstrate that medium-sized models with parameter counts between 10 and 50 billion can achieve almost equivalent accuracy as their larger counterparts. The aggregated performance scores by model size category and prompting strategy highlight this finding (Fig. 3). Moreover, Figure 4 also illustrates this performance across various models and model families. Model families such as Gemma 2 are competitive with more computationally expensive LLMs such as the larger GPT or Mistral models, achieving similar F1 scores. More detailed results for all models in Appendix (Figs. 12, 13, 14, and 15) further highlight our findings. For example, the best-performing medium-sized GPT-3.5 Turbo model with an F1 score of 0.1932 [0.1876, 0.1986] on average over all prompting strategies, almost performs on par with the large-sized GPT-4o, which exhibits the highest score of all models with an F1 score of 0.2137 [0.2084, 0.2191].

Closed-Source Over Open-Source Models. We find that the closed-source GPT models show prominent performances compared to most of the similar sized open-source models within their respective model size categories (Fig. 4). The results further exhibit that medium-sized open-source models often outperform smaller closed-source models, suggesting that even without proprietary

data, open-source models can be competitive with appropriate parameter sizes. In more details, the large closed-source model GPT-4o achieves the highest F1 scores of over 0.21 throughout all prompting strategies (Appendix, Fig. 12). The best-performing open-source model, Mistral Large 2 123B, reaches an F1 score of around 0.20. Medium-sized open-source models such as Gemma 2 27B perform on a similar accuracy level as the closed-source medium-sized GPT-3.5 Turbo model. In the category of small models we find that the closed-source GPT-4o mini slightly but not significantly outperforms the best open-source counterpart, Gemma 2 9B.

3.3 Addressing Potential Data Leakage

Given the setup of our experiment, our results may be susceptible to data leakage due to the possible overlap of submissions in our reddit dataset with the LLMs' training data [48]. For this, we follow the methodology of prior work [49] to compare model performance before and after the knowledge cutoff to detect any significant pre- and post-cutoff differences.

Setup. As previously, we conduct this robustness analysis with GPT-3.5 Turbo. Given that GPT-3.5 Turbo's knowledge cutoff (i.e., September 2021) is outside the timeframe of our original dataset, we first use GPT-4o as a higher-parameterized expert model to extract recommendations from reddit posts made after this cutoff. Using this dataset, we then repeat our zero-shot experiment with GPT-3.5 Turbo and compare the results to the main zero-shot results.

Robustness Analysis Dataset. We consider all submissions from r/MovieSuggestions submitted from October 2021 to July 2024 that contain the word "request" in the submission title to ensure the post contains actual movie inquiries (92 040 submissions). We then keep all submissions with at least five comments with more upvotes than downvotes, focusing on high-quality and engaging user discussions (441 submissions remaining). In contrast to the crowdworker-labeled dataset for our previous experiments, we now employ GPT-4o to extract movie mentions from the requesting user's post as

well as from commenters’ replies (Appendix, Fig. 7 for the prompt). Finally, similar as in our original dataset, we exclude submissions with no movies mentioned in the request and less than ten recommendations made by commenters. This final robustness analysis dataset consists of 236 submissions.

Performance After Knowledge Cutoff. We rerun our zero-shot prompting experiment using GPT-3.5 Turbo on the robustness analysis dataset, achieving an F1@10 of 0.1667 [0.1513, 0.1814],⁷ which does not indicate a significant decline compared to our original results (F1@10: 0.1962 [0.1809, 0.2112]). However, we observe a minor decline in performance, potentially due to movies released after the knowledge cutoff being relevant recommendations for certain requests in the robustness analysis dataset. This supports the validity of our findings, even considering that parts of our evaluation dataset could have been included in the LLMs’ training data.

4 Discussion

Our findings reveal that LLMs are highly suitable as narrative-driven recommenders. LLMs demonstrate remarkable utility, especially compared to traditional recommender baselines, in generating contextually relevant recommendations from free-form narrative inputs, which can be particularly valuable in enhancing user experience in personalized movie suggestions. In particular, we compare the performance of LLMs as narrative-driven movie recommenders to state-of-the-art recommender approaches [14] and find substantial leaps in recommendation quality. We attribute this to the superior ability of LLMs to interpret complex natural language expressions and nuances, which traditional methods, such as matrix factorization or doc2vec, often fail to capture effectively. Unlike these traditional systems, which rely on extensive feature engineering and domain-specific fine-tuning, LLMs can extract meaningful contextual information directly from narrative requests due to their extensive pre-training.

Moreover, similarly to recent work that demonstrated the effectiveness of zero-shot prompting in various NLP tasks [47, 54], we find that more extensive prompting strategies such as identity or few-shot prompting do not outperform zero-shot prompting. These results suggest that even simple prompting strategies can effectively capture essential user preferences in narrative-driven recommendation tasks, reducing the need for extensive prompt engineering, model training, or fine-tuning. This result aligns with a recent study that evaluated the performance of various prompts for GPT-3.5 Turbo on multiple recommendation scenarios based on Amazon customer reviews, such as rating prediction [31].

Nevertheless, while we find that simple zero-shot prompting performs as well as other strategies such as few-shot prompting, more sophisticated techniques could potentially further enhance recommendation quality. For example, Wang and Lim [54] explored the application of LLMs utilizing GPT-3 for zero-shot next-item recommendation in the context of movie suggestions. The authors developed a three-step prompting strategy to execute subtasks that capture user preferences, select representative previously watched movies, and recommend a ranked list of ten movies. Their experimental evaluation on the MovieLens 100K dataset indicates that

⁷Note that Recall@10 and F1@10 now have reduced upper limits of 0.41 and 0.54, resp., due to an increase in community suggestions per submission (mean=37.09).

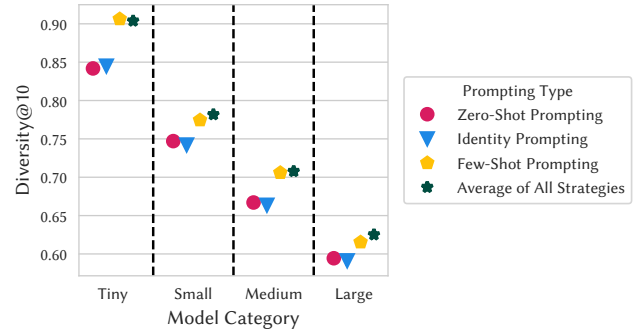


Figure 5: Inter-List Diversity for All Model Size Categories. We show inter-list diversity with vertical bars indicating bootstrapped 95% confidence intervals (often too narrow to be visible), observing a trend where large LLMs tend to produce less diverse responses across multiple experiment repetitions using the same prompt. Notably, few-shot prompting consistently results in higher diversity across all model sizes.

this method can outperform traditional recommendation models that require extensive training.

Furthermore, our experiments suggest that medium-sized LLMs, such as Gemma 2 27B, offer a compelling trade-off between computational efficiency and recommendation quality, competing with larger and more resource-intensive models. This finding relates to a study by He et al. [22] in the field of conversational recommenders, where medium-sized models such as GPT-3.5 Turbo produced results comparable to larger counterparts. Thus, medium-sized models emerge as viable alternatives for real-world applications, balancing performance and effectiveness.

Besides, while the closed-source GPT-4o model outperforms all other evaluated LLMs, the minimal performance gap between closed- and open-source models (e.g., Mistral 123B) highlights that open-source LLMs can provide high-quality movie recommendations without the significant computational or cost overhead of closed-source solutions. Liu et al. [32] corroborates this finding, showing that both open- and closed-source LLMs exhibit strong performance in content-based recommendation tasks.

In addition to recommendation performance, we also analyze the inter-list diversity of LLM recommendations (Fig. 5 and Appendix Fig. 16) as diverse recommendations typically enhance user experience in recommender systems [62]. Our investigation reveals that larger models exhibit lower diversity across prompting repetitions, indicating a higher tendency towards stable and repeated outputs. In contrast, smaller models, particularly tiny LLMs, show higher diversity, producing more varied recommendations across iterations. On top of that, few-shot prompting consistently increases recommendation diversity compared to zero-shot or identity prompting, likely due to the inclusion of different random examples across repetitions, as minor variations in prompts often lead to more output variability [60]. Importantly, this increase in diversity does not adversely impact recommendation accuracy across prompting strategies, as demonstrated by our reported results. Therefore, when system operators aim to improve user experience by increasing diversity, few-shot prompting or introducing minor variation in

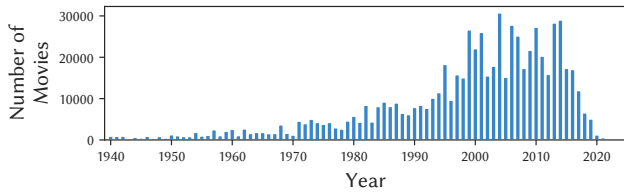


Figure 6: Release Years. Distribution of movies recommended by LLMs by their release year (visualized from 1940 to 2024) depicts strong preference for movies released after 2000.

prompts may be preferred over simple zero-shot prompting, as performance remains comparable while allowing for greater variation in the resulting LLM responses.

Finally, as recent works have highlighted the recency bias in LLMs [12], we investigate release years of the recommended movies (Fig. 6 and Appendix, Fig. 17). Our investigation indicates a noticeable preference for more recent movies, particularly those released after 2000. This suggests that LLMs tend to favor newer content, possibly influenced by inherent biases in the training data such as data availability or popularity of contemporary movies.

Limitations and Future Work. Our study has certain limitations, which we consider as opportunities for future research. First, our evaluation is restricted to a single domain and dataset from the r/MovieSuggestions subreddit. While this allows for a focused analysis of a wide range of open- and closed-source LLMs across various sizes, our work can be expanded by a broader validation of our approaches in similar recommendation scenarios where users formulate narrative requests for other domains, such as books or music (e.g., on the r/ifyoulikeblank subreddit). Second, we focus on three prompting strategies—zero-shot, identity, and few-shot prompting—while excluding more sophisticated approaches, such as chain-of-thought or tree-of-thought prompting, which have shown promise in tasks requiring strategic lookahead and may enhance recommender adaptability. Additionally, given the marginal differences in the results for our employed strategies, we do not investigate combinations of multiple prompting strategies (e.g., identity and few-shot). Moreover, the integration of multimodal inputs, such as images or audio, alongside textual user queries could further enrich the recommendation process, enhancing the system’s ability to respond to complex, multifaceted user preferences. Third, we did not test different individual LLM configurations, including parameters such as sampling method, temperature, or quantization details, which could further modify the output’s variability, diversity, and performance. Instead, we employed all LLMs using their default parameter settings, derived from official documentation or Ollama configuration files. When applying our findings to build real-world tools, practitioners should consider further evaluating these default settings to optimize recommendation quality and computational efficiency based on their specific use cases. Finally, we do not explore beyond recency bias to uncover other biases, such as skews in topic or popularity, that may be inherent in LLM responses. Exploring these biases further could yield valuable insights for both recommendation algorithms and pre-training of LLMs.

5 Further Related Work

The emergence of LLMs, such as BERT [13] or GPT [45], has led to new developments in NLP. Due to the significant increase in model sizes and the massive amount of training data, LLMs have shown ability of understanding and executing a wide variety of tasks, such as reasoning [10], software engineering [23] or language translation [9]. OpenAI’s ChatGPT as well as several competitive open-source LLMs, including Meta’s Llama, Mistral, and Microsoft’s Phi, have demonstrated exceptional performance, surpassing prior state-of-the-art NLP models across numerous tasks [21, 24, 56, 61].

LLMs are typically instructed through prompts—instructions to translate texts, answer questions, or write essays. Recent advancements in LLMs have sparked growing interest in enhancing their task performance through various prompting strategies [9], such as zero-shot, identity, or few-shot prompting. Further promising prompting strategies, such as chain-of-thought [55] or tree-of-thought [58], enhance the reasoning abilities of generative models, particularly in tasks such as question answering, by guiding them through human-like logical reasoning processes.

Additionally, recent work explored the suitability of LLMs for recommenders, revealing that models, such as Llama and GPT, improve user satisfaction with richer explanations [33], excel in data-sparse [42] and cold-start [50] scenarios, and can deliver highly relevant recommendations through conversational interactions [37].

This Work. In this paper, we are—to the best of our knowledge—the first to address the suitability of LLMs as narrative-driven recommenders. In particular, we utilize user-submitted text from reddit, consisting of narrative requests for movie recommendations, to evaluate the quality of LLM-generated recommendations.

6 Conclusion

In this paper, we evaluate the suitability of LLMs in a narrative-driven movie recommendation setting, comparing their performance to traditional state-of-the-art recommender systems. Our comprehensive analysis of 38 LLMs, both open- and closed-source, reveals that LLMs can effectively generate personalized movie recommendations from user-provided narratives, significantly outperforming traditional approaches, such as doc2vec. We find that while larger closed-source models generally demonstrate superior performance, medium-sized open-source models remain competitive, offering a viable trade-off between computational cost and recommendation quality. Notably, we observe minimal differences in effectiveness between zero-shot, identity, and few-shot prompting, indicating that simple approaches are sufficient for generating high-quality recommendations. These findings highlight the potential of LLMs to transform narrative-driven recommender systems, offering scalable solutions for integrating natural language capabilities into real-world applications. Our results also underscore the versatility of LLMs in real-world recommender applications. Minimizing prompt complexity is crucial, as it reduces operational overhead and latency, simplifying the integration of LLMs into existing recommendation frameworks. This finding is particularly valuable for researchers and practitioners aiming to incorporate LLMs in practical settings without the additional computational costs of model fine-tuning or complex prompt engineering.

References

- [1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, et al. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] <https://doi.org/10.48550/arXiv.2404.14219>
- [2] 01. AI, :, Alex Young, Bei Chen, Chao Li, et al. 2024. Yi: Open Foundation Models by 01.AI. arXiv:2403.04652 [cs.CL] <https://doi.org/10.48550/arXiv.2403.04652>
- [3] Mistral AI. 2024. AI in abundance: Introducing a free API, improved pricing across the board, a new enterprise-grade Mistral Small, and free vision capabilities on le Chat. <https://mistral.ai/news/september-24-release/>. Accessed on 2024-10-01.
- [4] Mistral AI. 2024. Cheaper, Better, Faster, Stronger. <https://mistral.ai/news/mixtral-8x22b/>. Accessed on 2024-10-01.
- [5] Mistral AI. 2024. Large Enough. <https://mistral.ai/news/mistral-large-2407/>. Accessed on 2024-10-01.
- [6] Mistral AI. 2024. Mistral NeMo. <https://mistral.ai/news/mistral-nemo/>. Accessed on 2024-10-01.
- [7] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 830–839. <https://doi.org/10.1609/icwsm.v14i1.7347>
- [8] Toine Bogers and Marijn Koolen. 2017. Defining and Supporting Narrative-Driven Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems* (Como, Italy) (RecSys '17). Association for Computing Machinery, New York, NY, USA, 238–242. <https://doi.org/10.1145/3109859.3109893>
- [9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- [10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, et al. 2022. PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs.CL] <https://doi.org/10.48550/arXiv.2204.02311>
- [11] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, et al. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) (RecSys '23). Association for Computing Machinery, New York, NY, USA, 1126–1132. <https://doi.org/10.1145/3604915.3610646>
- [12] Yashar Deldjoo. 2024. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. *ACM Trans. Recomm. Syst.* (Aug. 2024). <https://doi.org/10.1145/3690655>
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Computing Research Repository (CoRR)* (2018). <https://doi.org/10.48550/arXiv.1810.04805>
- [14] Lukas Eberhard, Kristina Popova, Simon Walk, and Denis Helic. 2024. Computing recommendations from free-form text. *Expert Systems with Applications* 236 (2024), 121268. <https://doi.org/10.1016/j.eswa.2023.121268>
- [15] Lukas Eberhard, Simon Walk, and Denis Helic. 2020. Tell Me What You Want: Embedding Narratives for Movie Recommendations. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media* (Virtual Event, USA) (HT '20). Association for Computing Machinery, New York, NY, USA, 301–306. <https://doi.org/10.1145/3372923.3404818>
- [16] Lukas Eberhard, Simon Walk, Lisa Posch, and Denis Helic. 2019. Evaluating Narrative-driven Movie Recommendations on Reddit. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) (IUI '19). Association for Computing Machinery, New York, NY, USA, 1–11. <https://doi.acm.org/10.1145/3301275.3302287>
- [17] Li Fei-Fei, R. Fergus, and P. Perona. 2006. One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 4 (2006), 594–611. <https://doi.org/10.1109/TPAMI.2006.79>
- [18] Michael Fink. 2004. Object Classification from a Single Example Utilizing Class Relevance Metrics. In *Proceedings of the 17th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) (NIPS'04). MIT Press, Cambridge, MA, USA, 449–456.
- [19] Google. 2024. Gemma 2 is now available to researchers and developers. <https://blog.google/technology/developers/google-gemma-2/>. Accessed on 2024-10-01.
- [20] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, et al. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. arXiv:2301.07597 [cs.CL] <https://doi.org/10.48550/arXiv.2301.07597>
- [21] Emman Haider, Daniel Perez-Becker, Thomas Portet, Piyush Madan, Amit Garg, et al. 2024. Phi-3 Safety Post-Training: Aligning Language Models with a "Break-Fix" Cycle. arXiv:2407.13833 [cs.CL] <https://doi.org/10.48550/arXiv.2407.13833>
- [22] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, et al. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (Birmingham, United Kingdom) (CIKM '23). Association for Computing Machinery, New York, NY, USA, 720–730. <https://doi.org/10.1145/3583780.3614949>
- [23] Xinyi Hou, Yanjie Zhao, Yue Liu, Zhou Yang, Kailong Wang, et al. 2024. Large Language Models for Software Engineering: A Systematic Literature Review. *ACM Trans. Softw. Eng. Methodol.* (Sept. 2024). <https://doi.org/10.1145/3695988>
- [24] Michael Ibrahim. 2024. Fine-Grained Languagebased Reliability Detection in Spanish New with Fine-Tuned Llama-3 Model. In *In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org*.
- [25] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, et al. 2018. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2704–2713.
- [26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, et al. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] <https://doi.org/10.48550/arXiv.2310.06825>
- [27] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, et al. 2024. Mixtral of Experts. arXiv:2401.04088 [cs.LG] <https://doi.org/10.48550/arXiv.2401.04088>
- [28] Junseok Kim, Nakyeong Yang, and Kyomin Jung. 2024. Persona is a Double-edged Sword: Enhancing the Zero-shot Reasoning by Ensembling the Role-playing and Neutral Prompts. arXiv preprint arXiv:2408.08631 (2024).
- [29] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio. 2008. Zero-Data Learning of New Tasks. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2* (Chicago, Illinois) (AAAI'08). AAAI Press, 646–651. <https://dl.acm.org/doi/10.5555/1620163.1620172>
- [30] Quoc Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*. PMLR, Beijing, China, 1188–1196. <https://proceedings.mlr.press/v32/le14.html>
- [31] Junling Liu, Chao Liu, Renjie Lv, Kang Zhou, and Yan Zhang. 2023. Is ChatGPT a Good Recommender? A Preliminary Study. arXiv:2304.10149 [cs.IR] <https://doi.org/10.48550/arXiv.2304.10149>
- [32] Qijiong Liu, Nuo Chen, Tetsuya Sakai, and Xiao-Ming Wu. 2024. ONCE: Boosting Content-based Recommendation with Both Open- and Closed-source Large Language Models. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) (WSDM '24). Association for Computing Machinery, New York, NY, USA, 452–461. <https://doi.org/10.1145/3616855.3635845>
- [33] Sebastian Lubos, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdenez, and Viet-Man Le. 2024. LLM-generated Explanations for Recommender Systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) (UMAP Adjunct '24). Association for Computing Machinery, New York, NY, USA, 276–285. <https://doi.org/10.1145/3631700.3665185>
- [34] Meta. 2024. Introducing Llama 3.1: Our most capable models to date. <https://ai.meta.com/blog/meta-llama-3-1/>. Accessed on 2024-10-01.
- [35] Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. <https://ai.meta.com/blog/meta-llama-3/>. Accessed on 2024-10-01.
- [36] Meta. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>. Accessed on 2024-10-01.
- [37] Cataldo Musto, Alessandro Francesco Maria Martina, Andrea Iovine, Marco de Gemmis, and Giovanni Semeraro. 2022. Tell Me What You Like: Introducing Natural Language Preference Elicitation Strategies in a Virtual Assistant for the Movie Domain. *Preprint submitted to Expert Systems with Applications* Available at SSRN 4140047 (2022). <https://doi.org/10.2139/ssrn.4140047>
- [38] Ollama. 2024. Phi-3.5-mini-instruct. <https://ai.azure.com/explore/models/Phi-3.5-mini-instruct/version/2/registry/azureml>. Accessed on 2024-10-01.
- [39] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>. Accessed on 2024-10-01.
- [40] OpenAI. 2024. Hello GPT-4o. <https://openai.com/index/hello-gpt-4o/>. Accessed on 2024-10-01.
- [41] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] <https://doi.org/10.48550/arXiv.2303.08774>
- [42] Dario Di Palma, Giovanni Maria Biancofiore, Vito Walter Anelli, Fedelucio Narducci, Tommaso Di Noia, et al. 2024. Evaluating ChatGPT as a Recommender System: A Rigorous Approach. arXiv:2309.03613 [cs.IR] <https://doi.org/10.48550/arXiv.2309.03613>
- [43] Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? arXiv preprint arXiv:2405.00492 (2024).
- [44] David Powers. 2008. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. *Mach. Learn. Technol.* 2 (01 2008).

- [45] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training. (2018).
- [46] John W Ratcliff and David E Metzner. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal* 13, 7 (1988), 46.
- [47] Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–7.
- [48] Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. To the cutoff... and beyond? a longitudinal perspective on LLM data contamination. In *The Twelfth International Conference on Learning Representations*.
- [49] Oscar Sainz, Jon Campos, Iker Garcia-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, et al. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 10776–10787. <https://doi.org/10.18653/v1/2023.findings-emnlp.722>
- [50] Scott Sanner, Krisztian Balog, Filip Radlinski, Ben Wedin, and Lucas Dixon. 2023. Large Language Models are Competitive Near Cold-start Recommenders for Language- and Item-based Preferences. In *Proceedings of the 17th ACM Conference on Recommender Systems* (Singapore, Singapore) (*RecSys '23*). Association for Computing Machinery, New York, NY, USA, 890–896. <https://doi.org/10.1145/3604915.3608845>
- [51] Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, et al. 2023. Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents. *arXiv:2304.09542 [cs.CL]* <https://doi.org/10.48550/arXiv.2304.09542>
- [52] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhatipatiraju, et al. 2024. Gemma: Open Models Based on Gemini Research and Technology. *arXiv:2403.08295 [cs.CL]* <https://doi.org/10.48550/arXiv.2403.08295>
- [53] Qwen Team. 2024. Qwen2.5: A Party of Foundation Models. <https://qwenlm.github.io/blog/qwen2.5/>. Accessed on 2024-10-01.
- [54] Lei Wang and Ee-Peng Lim. 2023. Zero-Shot Next-Item Recommendation using Large Pretrained Language Models. *arXiv:2304.03153 [cs.IR]* <https://doi.org/10.48550/arXiv.2304.03153>
- [55] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, et al. (Eds.), Vol. 35. Curran Associates, Inc., 24824–24837. https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
- [56] Shuyuan Xu, Wenyue Hua, and Yongfeng Zhang. 2024. OpenP5: An Open-Source Platform for Developing, Training, and Evaluating LLM-based Recommender Systems. *arXiv:2306.11134 [cs.IR]* <https://doi.org/10.48550/arXiv.2306.11134>
- [57] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. 2024. Qwen2 Technical Report. *arXiv:2407.10671 [cs.CL]* <https://doi.org/10.48550/arXiv.2407.10671>
- [58] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, et al. 2023. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, et al. (Eds.), Vol. 36. Curran Associates, Inc., 11809–11822. https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf
- [59] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore) (*SIGIR '08*). Association for Computing Machinery, New York, NY, USA, 603–610. <https://doi.org/10.1145/1390334.1390437>
- [60] Lexin Zhou, Wout Schellaert, Fernando Martínez-Plumed, Yael Moros-Daval, César Ferri, et al. 2024. Larger and more instructable language models become less reliable. *Nature* (2024), 1–8.
- [61] Yutao Zhu, Peitian Zhang, Chenghao Zhang, Yifei Chen, Binyu Xie, et al. 2024. IN-TERS: Unlocking the Power of Large Language Models in Search with Instruction Tuning. *arXiv:2401.06532 [cs.CL]* <https://doi.org/10.48550/arXiv.2401.06532>
- [62] Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web* (Chiba, Japan) (*WWW '05*). Association for Computing Machinery, New York, NY, USA, 22–32. <https://doi.org/10.1145/1060745.1060754>

Prompt for Labeling Robustness Analysis Dataset

Task	Detect all movies that are mentioned in a user's request. The user request is given in "" tags. Respond with a JSON object that contains a list of all mentioned movies, where each movie is provided in a single string which matches the format "<movie_title> - <IMDb_id>". Omit any additional text or output.
	Here is an example: ""Requesting beautiful, thoughtful, complex movies about the nature of consciousness and what it means to be alive.\nGhost in the Shell (1995) is one of my favourites and I love it. I also liked Upgrade and Ex Machina although they were significantly less hopeful."" {"movies": ["Ghost in the Shell - tt0113568", "Upgrade - tt6499752", "Ex Machina - tt0470752"]}
Examples	

Figure 7: Prompt for Robustness Analysis Dataset Creation. We visualize the system prompt for tagging movies in reddit submissions for our robustness experiment addressing data leakage (Sec. 3.3). It contains the task and format constraints as well as a few-shot example. When prompting GPT-4o as an expert labeler, we utilize this system prompt along user prompts containing reddit submission texts.

A Appendix

We show supplementary materials and figures for multiple subsections of the main body of our paper in this Appendix.

Evaluation of LLM Responses: Response Variance. For our analysis estimating the variance of LLM responses (Sec. 2.3), we show detailed results of our one-way ANOVA in Table 2.

Results: Format Adherence. While we show aggregated results for all metrics measuring formatting and structural correctness of LLM responses in the main part of the paper (Sec. 3.1), we hereafter visualize detailed results for all models regarding valid JSON format (Fig. 8), number of recommendations (Fig. 9), unique movies (Fig. 10), and movies released before the request (Fig. 11).

Results: Addressing Potential Data Leakage. We show the prompt we utilize to label our robustness analysis dataset using GPT-4o in Figure 7.

Results: Recommendation Performance. In our main result section, we report aggregate performance metrics as well as discuss F1@10 across models in regard to their model families and size category (Sec. 3.2). To supplement this, we now plot results for Precision@10, Recall@10, and NDCG@10—which do not differ substantially to F1@10—in similar style (Figs. 13, 14, 15) as well as further detailing results for F1@10 in a heatmap (Fig. 12).

Discussion: Diversity and Movie Years. We substantiate the aggregated plot of diversity for all model size categories (Sec. 4, Fig. 5) by visualizing inter-list diversity for individual models in Fig. 16. Furthermore, we lay out more details about the release year distribution of the movies contained in LLM responses in Fig. 17.

Table 2: Zero-Shot Performance Comparison Across 30 Repetitions Using One-Way ANOVA RM. This table presents the F-values (F Value) and associated p-values (Pr > F) for Precision@10, Recall@10, F1@10, and NDCG@10 metrics across 30 repetitions using zero-shot prompting for each LLM. The analysis is based on 29 numerator degrees of freedom (DF) and 8 555 denominator DF. The results exhibit no statistically significant differences in performance metrics between repetitions at the standard threshold of $p < 0.05$. To account for multiple tests we use Bonferroni correction and divide the significance level of 0.05 with the total number of tests (152) to obtain the final significance level of $p < 3.3 \cdot 10^{-4}$.

Model Category	Model Name	Precision@10		Recall@10		F1@10		NDCG@10	
		F Value	Pr > F	F Value	Pr > F	F Value	Pr > F	F Value	Pr > F
Tiny	Gemma 2B	0.5892	0.9607	1.0027	0.4612	0.8696	0.6667	0.8290	0.7267
	Gemma 2 2B	0.5518	0.9754	0.7571	0.8218	0.6942	0.8885	0.5749	0.9669
	Llama 3.2 1B	1.5303	0.0341	1.7137	0.0099	1.6844	0.0122	1.5194	0.0366
	Llama 3.2 3B	1.1144	0.3066	1.2545	0.1633	1.1839	0.2278	1.1578	0.2556
	Phi-3 3.8B	1.0710	0.3630	0.6720	0.9079	0.7118	0.8716	0.9528	0.5377
	Phi-3.5 3.8B	0.9381	0.5607	1.1038	0.3199	1.0670	0.3684	0.9265	0.5788
	Qwen2 0.5B	0.8661	0.6720	0.8565	0.6865	0.8247	0.7328	0.8677	0.6696
	Qwen2 1.5B	0.7861	0.7855	1.3928	0.0782	1.2753	0.1472	0.9708	0.5099
	Qwen2.5 0.5B	1.3712	0.0884	0.9001	0.6199	0.9464	0.5477	1.3100	0.1231
	Qwen2.5 1.5B	0.8943	0.6289	1.0432	0.4017	1.0125	0.4466	0.8725	0.6623
	Qwen2.5 3B	0.7361	0.8460	0.7078	0.8756	0.7026	0.8806	0.7395	0.8422
Small	Gemma 7B	0.9148	0.5970	0.9425	0.5537	0.9248	0.5815	0.6911	0.8914
	Gemma 2 9B	1.2181	0.1946	1.3006	0.1293	1.2950	0.1331	1.0597	0.3785
	GPT-4o mini	1.0238	0.4299	0.8861	0.6416	0.9437	0.5519	1.3070	0.1251
	Llama 3 8B	1.0565	0.3829	0.8686	0.6683	0.8875	0.6393	1.0257	0.4271
	Llama 3.1 8B	0.9477	0.5457	0.7396	0.8421	0.7734	0.8017	0.7961	0.7723
	Mistral 7B	1.4327	0.0621	1.3772	0.0855	1.3807	0.0838	1.2420	0.1736
	Qwen2 7B	0.9254	0.5805	0.9736	0.5056	0.9290	0.5748	0.7936	0.7756
	Qwen2.5 7B	0.7321	0.8503	0.6830	0.8986	0.6897	0.8926	0.8768	0.6558
	Yi 6B	0.6908	0.8916	0.7751	0.7996	0.7841	0.7881	0.7409	0.8406
	Yi 9B	1.0973	0.3281	1.1183	0.3018	1.1393	0.2766	1.2357	0.1789
Medium	Gemma 2 27B	1.0867	0.3419	1.0807	0.3499	1.0892	0.3387	1.0701	0.3643
	GPT-3.5 Turbo	0.8981	0.6230	0.9619	0.5235	0.9313	0.5712	0.9267	0.5785
	Mistral NeMo 12B	0.8833	0.6459	0.8893	0.6367	0.8701	0.6660	0.7129	0.8704
	Mistral Small 22B	0.9447	0.5504	0.9341	0.5669	0.8949	0.6279	0.9096	0.6052
	Mixtral 8x7B	0.6807	0.9005	0.6872	0.8949	0.6802	0.9010	0.8734	0.6609
	Phi-3 14B	0.7426	0.8387	0.8350	0.7181	0.8038	0.7619	0.6936	0.8890
	Qwen2.5 14B	1.6317	0.0175	1.3929	0.0782	1.4445	0.0579	1.7059	0.0105
	Qwen2.5 32B	1.3921	0.0785	1.1496	0.2648	1.1724	0.2397	1.5363	0.0328
	Yi 34B	0.4645	0.9937	0.5696	0.9690	0.5277	0.9824	0.7070	0.8764
Large	GPT-4o	1.1677	0.2448	1.2949	0.1332	1.2904	0.1363	0.7112	0.8722
	Llama 3 70B	0.7460	0.8348	0.8824	0.6472	0.8725	0.6623	0.5791	0.9651
	Llama 3.1 70B	1.4396	0.0596	1.3234	0.1147	1.3243	0.1141	1.4188	0.0674
	Llama 3.1 405B	1.0021	0.4620	0.9738	0.5051	1.0004	0.4647	0.9988	0.4671
	Mistral Large 123B	1.1131	0.3082	1.0666	0.3690	1.0983	0.3269	1.4827	0.0459
	Mixtral 8x22B	0.9408	0.5565	0.7787	0.7951	0.8451	0.7033	1.0889	0.3391
	Qwen2 72B	0.8674	0.6701	0.7723	0.8032	0.7891	0.7816	0.8315	0.7231
	Qwen2.5 72B	1.0417	0.4039	1.0040	0.4592	1.0458	0.3980	1.1607	0.2525

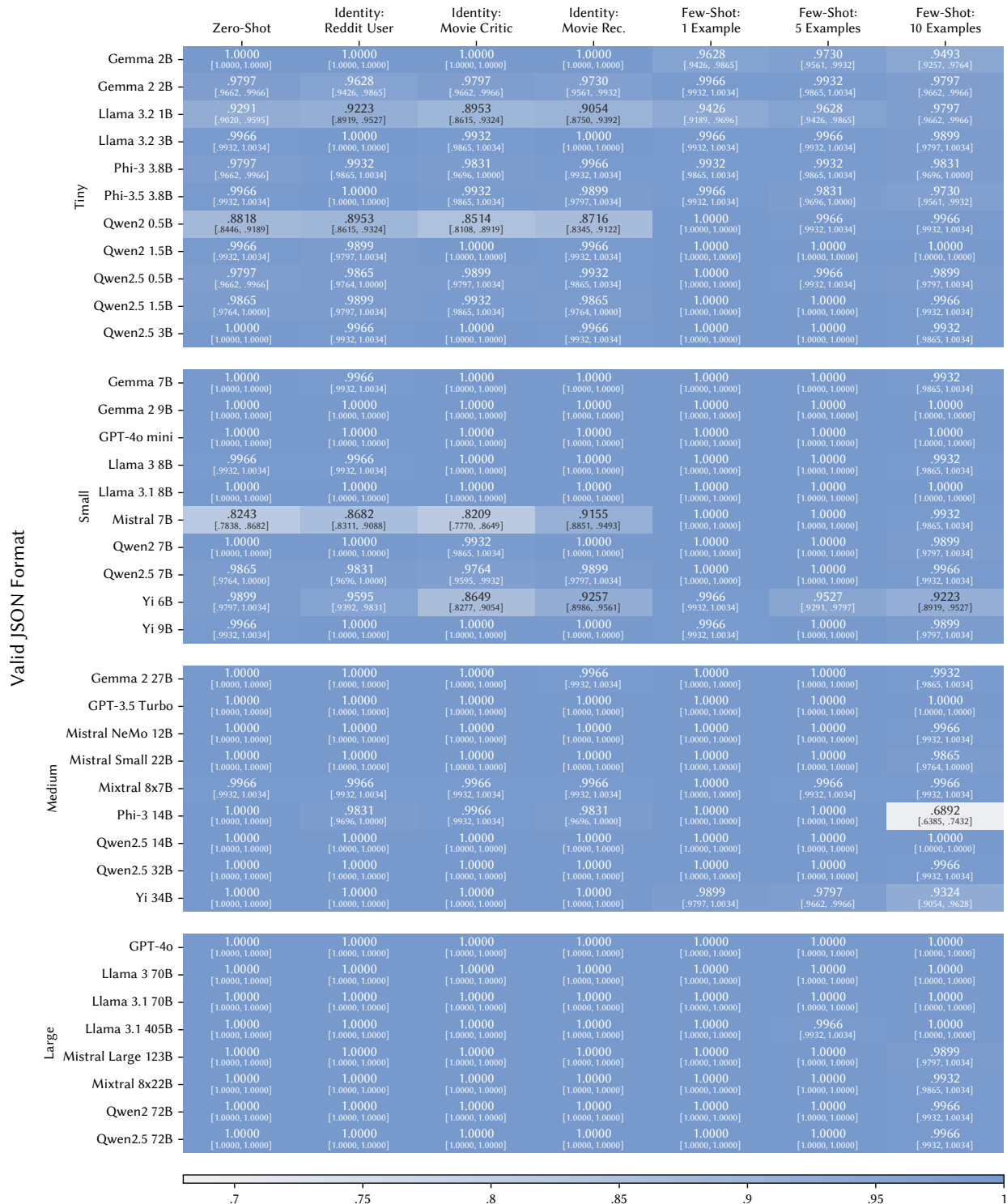
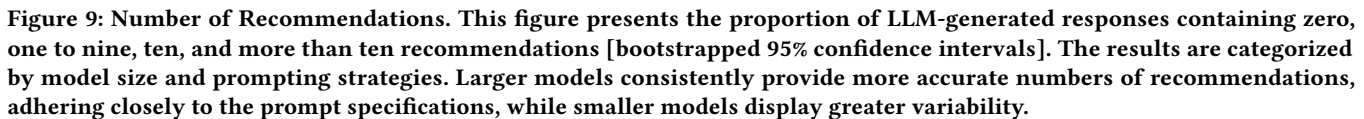


Figure 8: JSON Format. This figure shows the proportion of valid JSON responses generated by LLMs [bootstrapped 95% confidence intervals]. The results are categorized by model size and prompting strategies. Despite observing high proportion of valid output across all models, increased model size impacts the accuracy of valid output. Larger models achieve higher percentages of valid responses, demonstrating a benefit with scale.



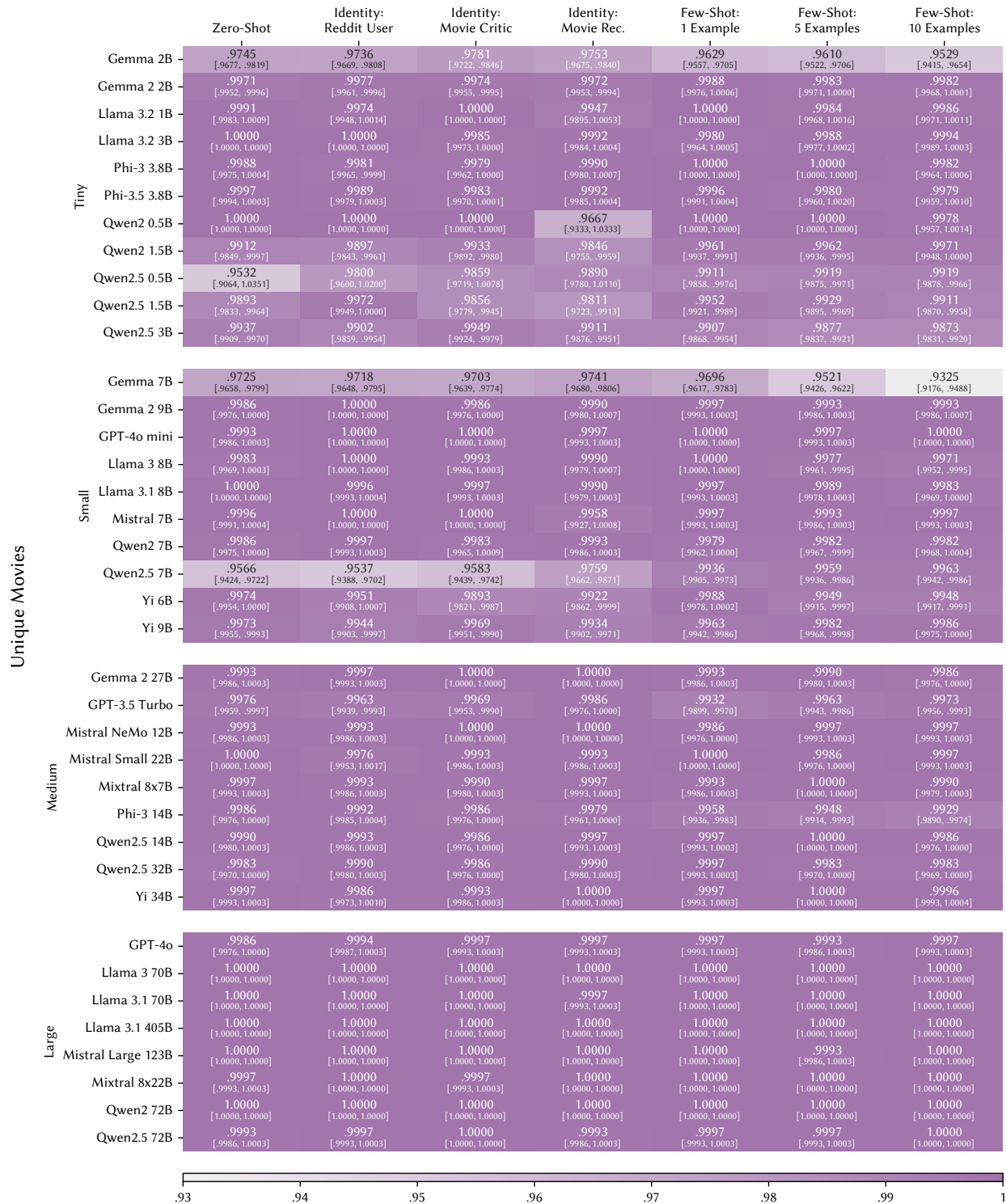


Figure 10: Unique Movies. This figure shows the average fraction of unique movie recommendations per request [bootstrapped 95% confidence intervals]. The results are categorized by model size and prompting strategies. Medium- and large-sized LLMs consistently returned over 99% unique movies within their valid JSON responses, while tiny and small models exhibited a slight but statistically significant increase in duplicate recommendations.

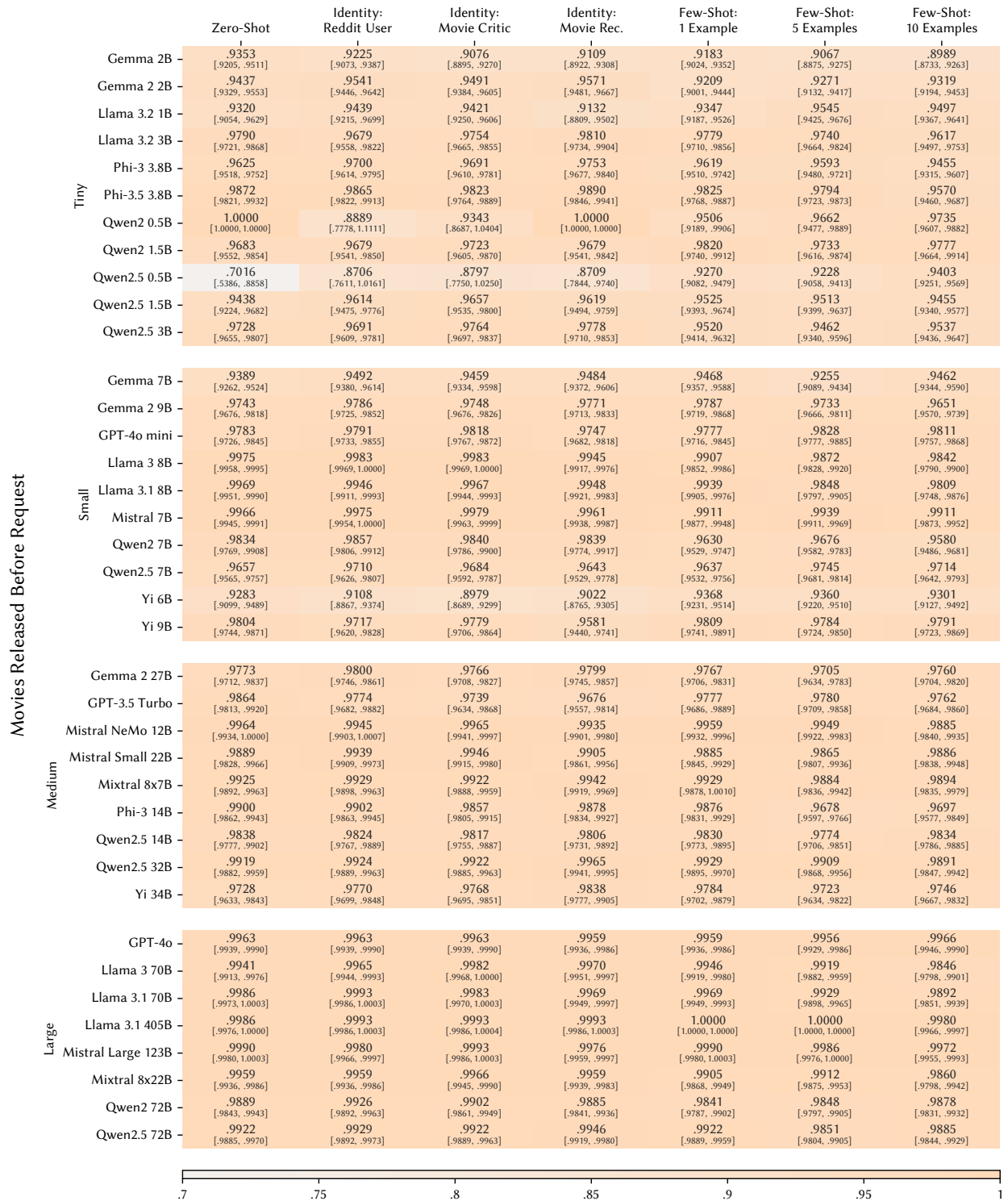


Figure 11: Movies Matching the Requested Year. This figure presents the fraction of movies that correspond to the requested release year [bootstrapped 95% confidence intervals]. The results are categorized by model size and prompting strategies. The results highlight the ability of larger LLMs to comply with the constraints stated in the prompt. Tiny model, specifically Qwen2.5 0.5B with zero-shot prompting, struggled to recommend movies in the requested timespan with around 30% of the recommended movies outside the given period.

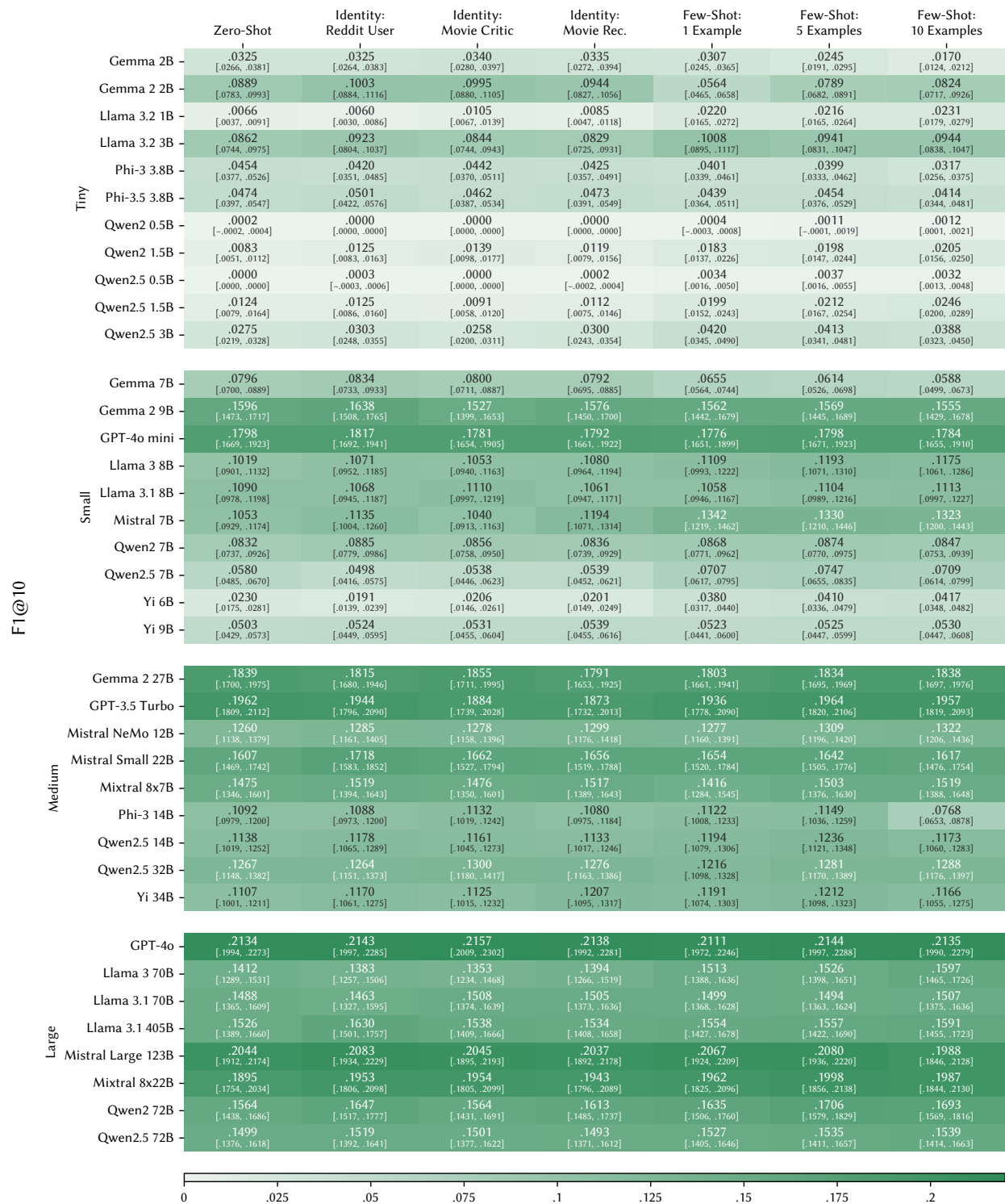


Figure 12: Detailed Recommendation Performance of Different LLMs and Prompting Strategies. We report F1@10 [bootstrapped 95% confidence intervals] for evaluated LLMs. The results are categorized by model size and prompting strategies.

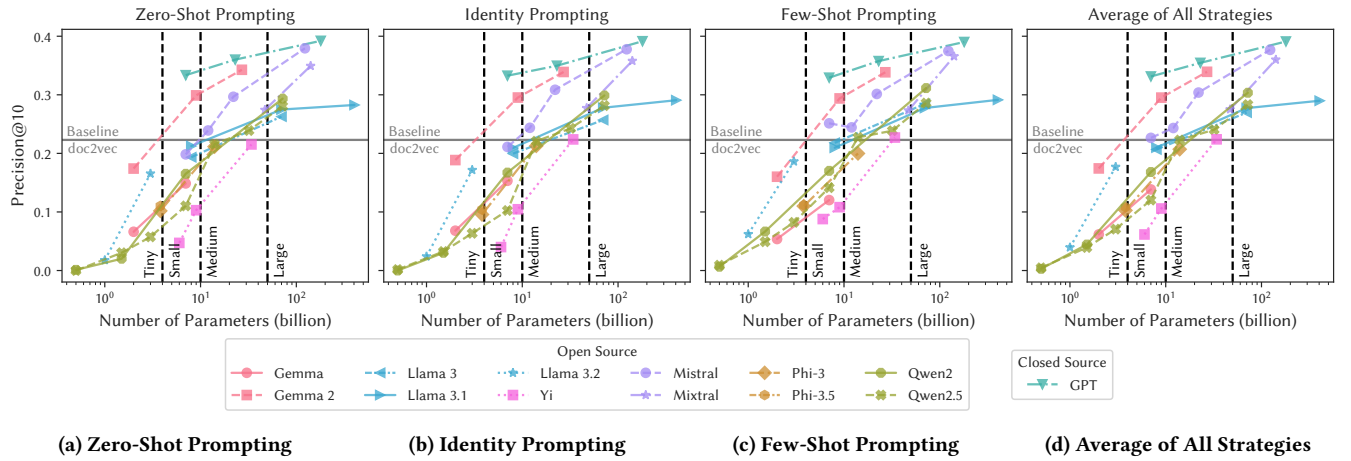


Figure 13: Precision@10 Across All Models. We show results for precision for all models and prompting strategies.

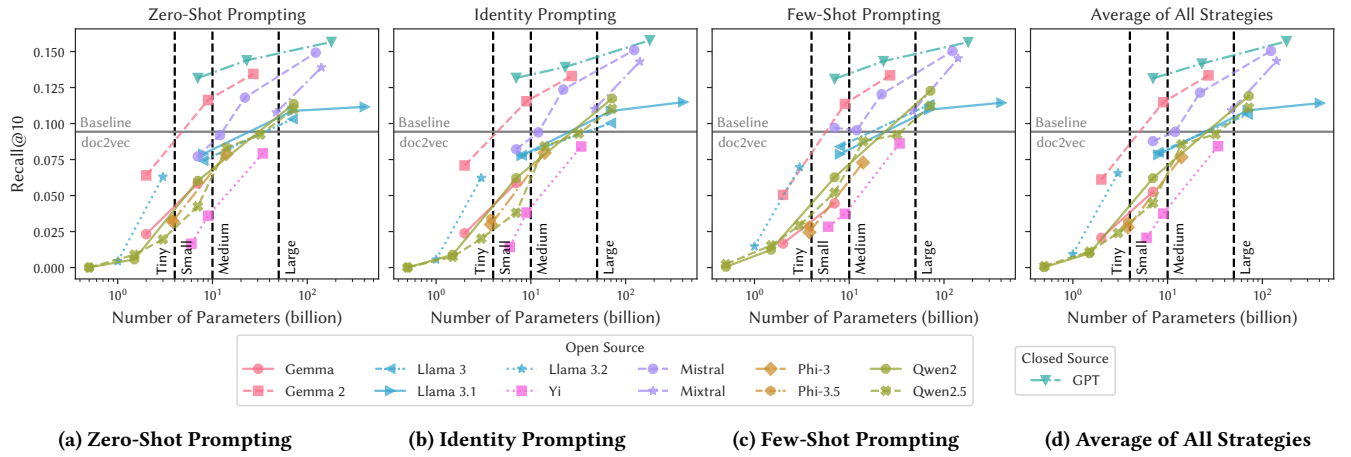


Figure 14: Recall@10 Across All Models. We show results for recall for all models and prompting strategies.

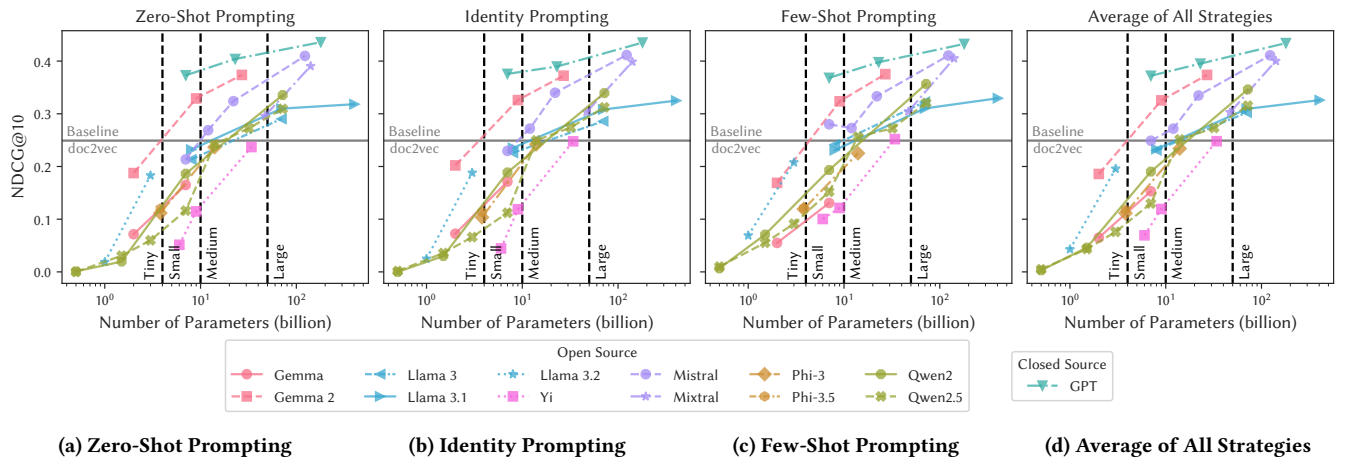


Figure 15: NDCG@10 Across All Models. We show results for NDCG for all models and prompting strategies.

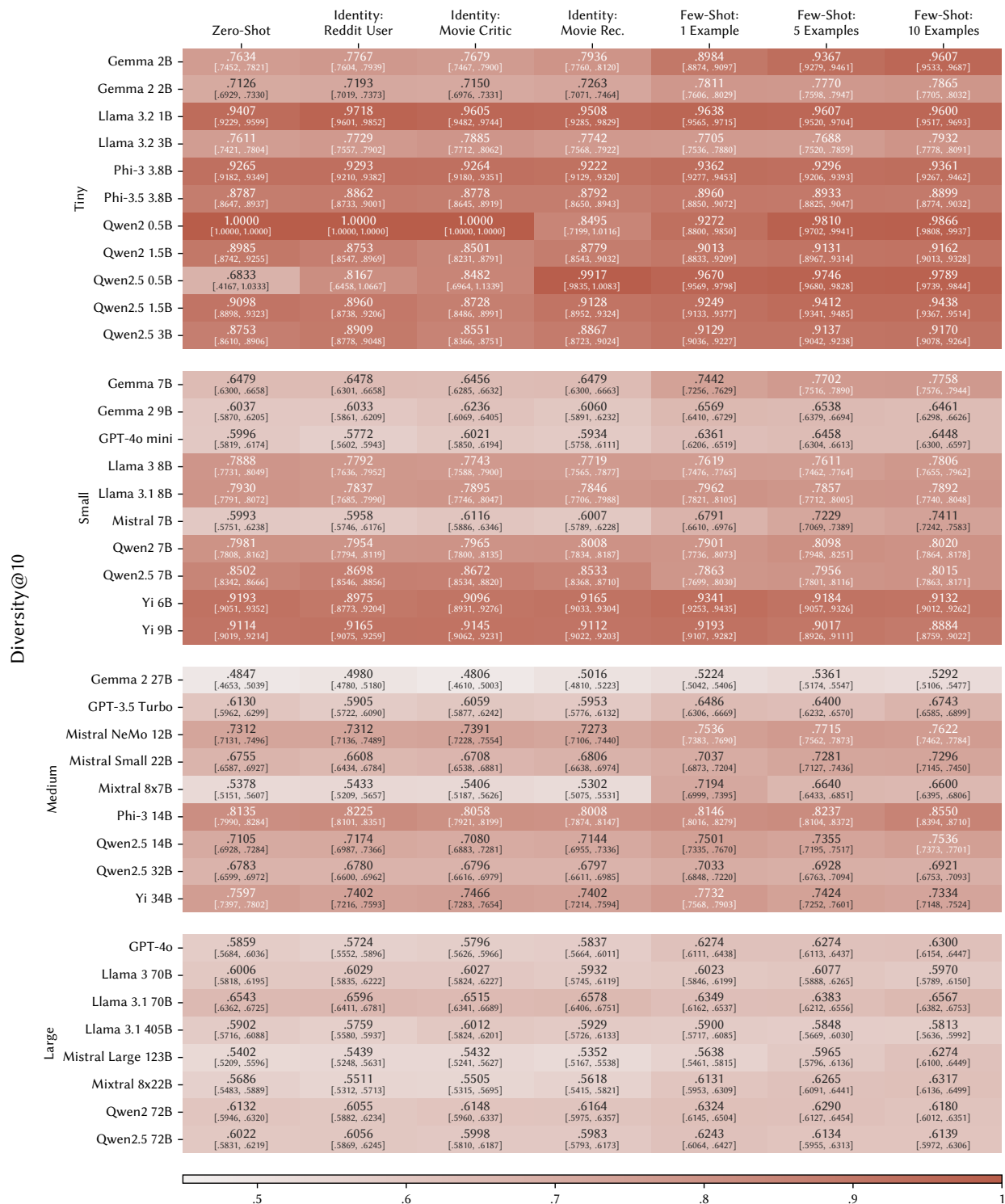


Figure 16: Inter-list Diversity Across Repetitions. The figure illustrates the inter-list Diversity@10 [bootstrapped 95% confidence intervals] of recommendations generated by different LLMs across 30 repetitions for zero-shot and three repetitions for other prompting strategies, showing the variability in the uniqueness of recommended items across runs.

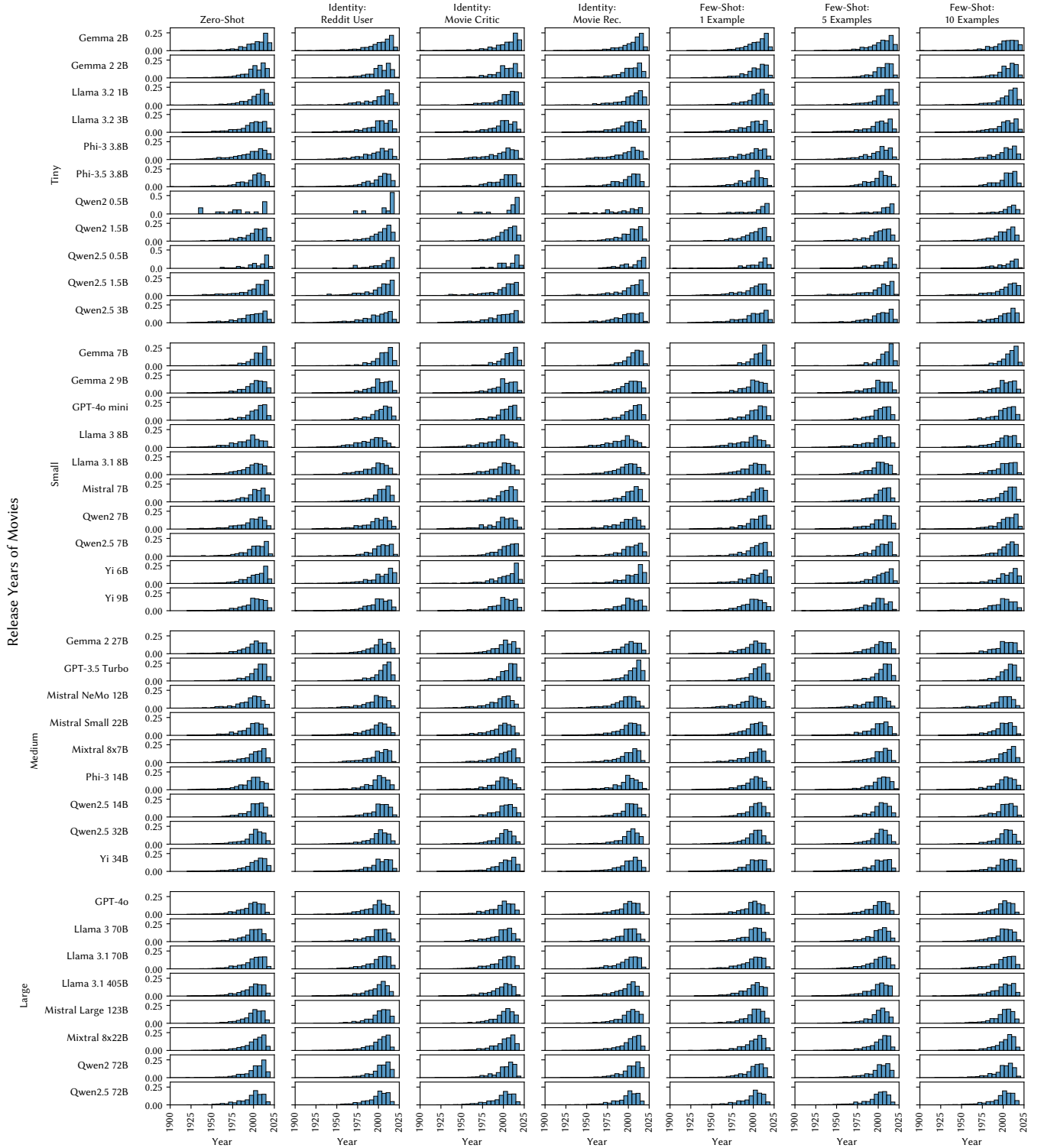


Figure 17: Distribution of Movie Release Years. This figure shows the distribution of movies recommended by LLMs based on their release years. The results are categorized by model size and prompting strategies. The x-axis represents the year of release, spanning from 1950 to 2025, while the y-axis indicates the fraction of movies released each year. The data, aggregated in 5-year-spans, exhibits a noticeable increase in movie releases over time, with a sharp rise in the number of releases after the year 2000.