

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Acquiring, Exploiting, and Suppressing Parametric Knowledge in Language Models

Permalink

<https://escholarship.org/uc/item/8r5773m6>

Author

Sadeq, Nafis

Publication Date

2025

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Acquiring, Exploiting, and Suppressing Parametric Knowledge in Language Models

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Nafis Sadeq

Committee in charge:

Professor Julian McAuley, Chair
Professor Taylor Berg-Kirkpatrick
Professor Zhiting Hu
Professor Jingbo Shang

2025

Copyright

Nafis Sadeq, 2025

All rights reserved.

The Dissertation of Nafis Sadeq is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2025

DEDICATION

To the computer science community that tirelessly advances the field through open-source collaboration, shared knowledge, and relentless curiosity. This work stands on the shoulders of those who choose openness over competition and contribution over credit.

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Table of Contents	v
List of Figures	viii
List of Tables	x
Acknowledgements	xii
Vita	xiv
Abstract of the Dissertation	xv
Introduction	1
Chapter 1 Acquiring Parametric Knowledge	5
1.1 Introduction	5
1.2 InforMask	7
1.2.1 Informative Relevance	7
1.2.2 Scoring Masking Candidates	9
1.2.3 Token-Specific Masking Rates	10
1.3 Experiments	12
1.3.1 Pretraining Corpus	12
1.3.2 Evaluation Benchmarks	12
1.3.3 Baselines	12
1.3.4 Training Details	13
1.3.5 Impact of Masking Strategies	13
1.3.6 Training Dynamics	14
1.3.7 Impact on Stop Words and Entities	15
1.3.8 Impact of Token-Specific Masking Rates	15
1.3.9 InforMask vs. PMI Masking	16
1.3.10 Case Study	17
1.4 Conclusion	19
Chapter 2 Exploiting Parametric Knowledge	20
2.1 Introduction	20
2.2 IPRA	23
2.2.1 Preference Alignment Corpus	23
2.2.2 Direct Preference Optimization	24
2.2.3 Text Encoder Fine-Tuning	26

2.2.4	Item Retrieval During Inference	27
2.3	Experiments	28
2.3.1	Datasets	28
2.3.2	Baselines & Metrics	29
2.3.3	Training Details	31
2.3.4	Recommendation Performance	32
2.3.5	Robustness Analysis	34
2.3.6	Ablation Study	36
2.3.7	Case Studies	37
2.4	Conclusion	38
Chapter 3	Supressing Parametric Knowledge	41
3.1	Introduction	41
3.2	RoleFact	44
3.2.1	Intermediate Response Generation	45
3.2.2	Atomic Fact Decomposition	46
3.2.3	Atomic Fact Verification	46
3.2.4	Self-reflection Update	47
3.3	Experiments	47
3.3.1	Baselines	49
3.3.2	Evaluation Metrics	49
3.3.3	Results	49
3.3.4	Hyper-parameter Tuning	50
3.3.5	Ablation study	52
3.3.6	Human Evaluation	53
3.3.7	Case Study	53
3.4	Conclusion	54
Chapter 4	Related Work	59
4.1	Related Work for InforMask	59
4.1.1	Random Masking	59
4.1.2	Span Masking	59
4.1.3	Entity-based Masking	60
4.1.4	PMI Masking	60
4.1.5	Knowledge-Enhanced LMs	60
4.2	Related Work for IPRA	61
4.2.1	User Preference Elicitation	61
4.2.2	Dialogue Strategy for CRSs	62
4.2.3	Language Models for CRSs	63
4.3	Related Work for RoleFact	64
4.3.1	Language Models for Role-play	64
4.3.2	Factuality in Natural Language Generation	65
4.3.3	Retrieval Augmented Methods	65

Chapter 5	Conclusion and Future Work	66
Bibliography		68

LIST OF FIGURES

Figure 1.1.	The informative scores of randomly sampled masking candidates ($s = 4$). [M] denotes the masked tokens. The pretraining objective of the masked language model (MLM) is to predict the masked tokens based on the context.	5
Figure 1.2.	The PMI matrix of the words in the sentence ‘The dual is between Harry Potter and Lord Voldemort.’	8
Figure 1.3.	Macro average MRR of different masking strategies on LAMA, evaluated every 10k steps.	14
Figure 1.4.	Performance of InformBERT for the full pretraining process. It achieves comparable performance with BERT-base after 40k training steps and even RoBERTa-large after 120k training steps.	14
Figure 1.5.	Frequency of common stop words and their corresponding masking rates by InforMask.	16
Figure 1.6.	Comparison between looping the same data and using token-specific masking rate to approximate the masking decisions. The models are trained for 3 epochs.	17
Figure 1.7.	Masking rate of selected tokens according to different masking policies. . .	18
Figure 3.1.	Example of cross-universe hallucination (Hiccup should not answer questions about Hogwarts) and temporal hallucination (Harry should not talk about the Patronus charm in his first year) in character role-play.	42
Figure 3.2.	Factual precision degrades when we minimize parametric knowledge by anonymizing the prompts.	43
Figure 3.3.	Factual precision degrades with decreasing character popularity (left to right), shown for characters associated with the ‘How to Train Your Dragon’ series.	43
Figure 3.4.	An overview of RoleFact, which performs parametric and non-parametric verification of atomic facts.	45
Figure 3.5.	Factual precision performance by role popularity (decreasing popularity left to right, adversarial task, GPT-3.5).	51

Figure 3.6.	Calibration for confidence threshold (t) and sample size (m) on the validation set with GPT-3.5. (a, b) shows fact score and SFPR on adversarial task while (c, d) shows the same on open ended task. The temporal hallucination rates for dialogue completion and scene-grounded tasks are shown by e and f respectively.	51
Figure 3.7.	Performance of retrieval methods w.r.t number of retrieved documents (adversarial task, GPT-3.5).	52

LIST OF TABLES

Table 1.1.	Statistics of LAMA [83] dataset, used for factual knowledge probing.	10
Table 1.2.	Statistics of extractive question answering dataset SQuAD v1 and v2 [88, 87].	10
Table 1.3.	Performance of different masking strategies and models on LAMA [83]. (a) We compare the models trained with different masking strategies for 3 epochs. (b) We compare InformBERT, a BERT model trained with InforMask for 40 epochs with BERT and RoBERTa models.	11
Table 1.4.	Performance of InforMask on SQuAD v1 and v2 [88, 87] development set.	11
Table 1.5.	Some examples of InformBERT and RoBERTa-base predictions on LAMA [83]. We show the queries and the ground-truth answers with the model predictions. We only show the top-3 predictions made by each model. . . .	18
Table 2.1.	Cross-trial robustness issue for a hybrid CRS with GPT-3.5-Turbo [78] and Text embedding, reported in terms of the coefficient of variation of reciprocal ranks.	20
Table 2.2.	Cross-domain generalization issue of a hybrid CRS with GPT-3.5-Turbo and Text embedding.	21
Table 2.3.	Frequently used notations for IPRA.	23
Table 2.4.	LLM prompts related to steps in IPRA and dataset construction for GYPhone.	25
Table 2.5.	Dataset statistics for movie and electronics domain.	29
Table 2.6.	Performance of IPRA compared to baselines on ReDial, underlined results outperform all the baselines.	33
Table 2.7.	Performance of IPRA compared to baselines on GYPhone, underlined results outperform all the baselines.	33
Table 2.8.	Robustness of IPRA compared to baselines on ReDial in terms of CV for decoding temperature of 0.7, 0.8, and 0.9 for 30 trials per sample. Lower CV indicates improved robustness.	35
Table 2.9.	Robustness of IPRA compared to baselines on GYPhone in terms of CV for decoding temperature of 0.7, 0.8, and 0.9 for 30 trials per sample. Lower CV indicates improved robustness.	35
Table 2.10.	Ablation Study, shown for IPRA-Llama with a decoding temperature of 0.7.	37

Table 2.11.	Case studies for conversational recommendation performance of IPRA-Llama	39
Table 3.1.	Factual precision (Fact Score $\uparrow$) and informativeness (SFPR $\uparrow$) of RoleFact on all tasks in the SGR dataset. All results here are produced with BM25 retrieval.	47
Table 3.2.	Temporal hallucination rate (THR $\downarrow$) on scene-grounded interview and dialogue completion tasks.	50
Table 3.3.	Ablation study of RoleFact on adversarial task using GPT-3.5.	53
Table 3.4.	Human ratings for factuality, informativeness, and speaker style similarity on a scale of one to seven	53
Table 3.5.	Role-playing case studies against baselines, hallucinations are shown in red.	55
Table 3.6.	Role-playing case studies against baselines, hallucinations are shown in red.	56
Table 3.7.	Role-playing case studies against baselines, hallucinations are shown in red.	57

ACKNOWLEDGEMENTS

All praise is due to Allah, who created me and guided me, and granted me soundness in health, vision, and hearing. I am grateful to my family members for beautifying my life and supporting me throughout this journey.

I would like to acknowledge Professor Julian McAuley for sharing his vast knowledge and experience with me—both regarding the methods of research and the ways to present it—and for being a constant source of encouragement. His flexible attitude towards research and his efforts in fostering an excellent environment for collaboration made my PhD journey remarkably smooth. My heartfelt appreciation goes out to Professor Taylor Berg-Kirkpatrick, Professor Zhiting Hu, and Professor Jingbo Shang for providing me with indispensable suggestions and encouragement that helped shape my research.

I am grateful to the amazing collaborators in the McAuley Lab, such as Canwen Xu, Shuyang Li, Zhouhang Xie, Zhankui He, Ahmad Bin Rabiah, Yu Wang, and Xin Xu. I would also like to acknowledge the contributions of our industry collaborators, including Jay Kang, Prarit Lamba, Xiang Gao, Kevin Liu, and Mostafa El-Khamy, for their support and feedback during internships and collaborative projects.

Chapter 1, in part, is a reprint of material that appears in the paper titled “InforMask: Unsupervised Informative Masking for Language Model Pretraining” by Nafis Sadeq, Canwen Xu, and Julian McAuley, which was published in **Empirical Methods in Natural Language Processing**, 2022. The dissertation author was the lead author of this paper.

Chapter 2, in part, is a reprint of material that appears in the paper titled “Enhancing Conversational Recommendations via Aligned Intermediate Preference Representations” by Nafis Sadeq, Zhankui He, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, and Julian McAuley, which was submitted to the **ACM Conference on Recommender Systems**, 2025. The dissertation author was the lead author of this paper.

Chapter 3, in part, is a reprint of material that appears in the paper titled “Mitigating Hallucination in Fictional Character Role-Play” by Nafis Sadeq, Zhouhang Xie, Byungkyu

Kang, Prarit Lamba, Xiang Gao, and Julian McAuley, which was published in the *Findings of Empirical Methods in Natural Language Processing*, 2024. The dissertation author was the lead author of this paper.

VITA

2017	B.Sc. in Computer Science and Engineering, Bangladesh University of Engineering and Technology
2017–2019	REVE Systems
2021	M.Sc. in Computer Science and Engineering, Bangladesh University of Engineering and Technology
2020–2021	Samsung R&D Institute Bangladesh
2025	Ph.D. in Computer Science, University of California San Diego

PUBLICATIONS

See the full list of publications on my website: <https://www.nsadeq.com/>.

Nafis Sadeq*, Canwen Xu*, Julian McAuley. “InforMask: Unsupervised informative masking for language model pretraining”. Empirical Methods in Natural Language Processing, 2022

Nafis Sadeq, Byungkyu Kang, Prarit Lamba and Julian McAuley. “Unsupervised improvement of factual knowledge in language models”. European Chapter of the Association for Computational Linguistics, 2023

Nafis Sadeq, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, Xiang Gao and Julian McAuley. “Mitigating Hallucination in Fictional Character Role-Play”. Findings of Empirical Methods in Natural Language Processing, 2024

Nafis Sadeq, Xin Xu, Zhouhang Xie, Julian McAuley, Byungkyu Kang, Prarit Lamba, Xiang Gao. “Improving In-Context Learning with Reasoning Distillation”. Under Review in Conference on Language Modeling, 2025

Nafis Sadeq, Zhankui He, Zhouhang Xie, Byungkyu Kang, Prarit Lamba and Julian McAuley. “Enhancing Conversational Recommendations via Aligned Intermediate Preference Representations”. Under Review in ACM Conference on Recommender Systems, 2025

ABSTRACT OF THE DISSERTATION

Acquiring, Exploiting, and Suppressing Parametric Knowledge in Language Models

by

Nafis Sadeq

Doctor of Philosophy in Computer Science

University of California San Diego, 2025

Professor Julian McAuley, Chair

Parametric knowledge refers to the information stored within the parameters of a language model. This is the knowledge the model acquires during pretraining, which allows it to answer questions, complete tasks, and generate coherent and relevant text—without relying on external tools or databases. Despite its utility, the parametric knowledge of language models has limitations. Language models often fail to memorize long-tail facts. Complex systems relying on parametric knowledge can also suffer from hallucinations, outdated knowledge, and robustness issues. In this dissertation, we explore pretraining, post-training, and training-free methods to mitigate such issues related to parametric knowledge.

We propose a novel pretraining strategy using token-level masking rates to improve

factual knowledge acquisition within language models. We demonstrate that aligning language models to follow reward signals from external components can improve the robustness of hybrid systems relying on parametric knowledge. While parametric memory plays a critical role in many tasks, we find that its unrestricted use can lead to hallucinations. We show that the right combination of parametric and non-parametric fact-checking can significantly mitigate this problem.

Introduction

Parametric knowledge forms the core intelligence of a language model—it’s what enables the model to understand language, reason, and produce informative responses. While retrieval and fine-tuning enhance performance, parametric knowledge is the essential engine driving the model’s default capabilities.

Importance of Parametric Knowledge

Parametric knowledge is important for the following reasons:

Foundation for Generalization Parametric knowledge allows language models to generalize beyond specific examples seen during training. Instead of memorizing data, the model captures patterns, associations, and structures of language, enabling it to generate coherent, contextually appropriate responses to novel inputs.

Task Versatility Many tasks—such as question answering, translation, summarization, or reasoning—can be performed without task-specific fine-tuning because the model has absorbed broad world knowledge and linguistic competence into its parameters. This makes the model zero-shot or few-shot capable [4].

Inference Time Efficiency Because the knowledge is stored within the model weights, it can be accessed directly at inference time without external lookups. This allows for fast and compact reasoning, especially in cases where retrieving from an external database would be slow or impractical.

Fallback Knowledge Parametric knowledge acts as a built-in knowledge base [83]. Even if retrieval-augmented tools or external APIs are unavailable, the model can still produce useful outputs based on what it has encoded internally.

Complement to Retrieval and Tool Use In modern hybrid systems, parametric knowledge complements retrieval-based or tool-augmented approaches [1]. While retrieval systems offer up-to-date or niche data, the model’s parametric knowledge can interpret, contextualize, and reason with that data.

Limitations Associated with Parametric Knowledge

While parametric knowledge enables powerful capabilities in modern natural language processing (NLP) systems, it can also introduce several limitations. Key limitations include:

Forgetting Long-Tail Factual Knowledge

Parametric knowledge often overlooks long-tail facts in the training data. Traditional masked language modeling (MLM) and causal language modeling (CLM) apply a uniform masking rate to all tokens. However, this uniform masking is suboptimal, as factually relevant tokens—such as named entities—are less frequent than stop words in the pretraining corpus. In Chapter 1, we address this issue in the context of MLMs. We propose *InforMask*, a new unsupervised masking strategy for training MLMs that prioritizes factually relevant tokens during pretraining. InforMask leverages Pointwise Mutual Information (PMI) to select the most informative tokens to mask. We further introduce two optimizations to enhance InforMask’s efficiency. With a one-off preprocessing step, InforMask outperforms random masking and previously proposed strategies on the factual recall benchmark LAMA [83], as well as on the question answering benchmarks SQuAD v1 [88] and v2 [87].

Stale or Outdated Information

Parametric knowledge is fixed at the time of training. For instance, if a model is trained on data from 2023, it will not be aware of events or facts from 2024 onward unless it is retrained

or augmented with retrieval tools. This limitation makes models poorly suited for staying current in fast-changing domains such as e-commerce, news, technology, or medicine. We address this issue in Chapter 2 in the context of Conversational Recommender Systems (CRSs).

CRSs, which engage users in multi-turn natural language interactions to elicit their preferences, have benefited from advancements in large language models (LLMs). However, LLM-powered CRSs also introduce new challenges. The static parametric knowledge of LLMs cannot keep pace with a large and evolving item space, which limits the capabilities of LLM-prompting-based CRSs [122].

To mitigate this limitation, some studies have explored hybrid systems that combine LLMs with separate recommendation modules, where the LLM generates intermediate representations that drive external recommendation components [21]. However, these hybrid systems often suffer from inconsistency across multiple trials due to variability in LLM-generated representations. Furthermore, LLMs’ ability to generate effective preference representations frequently fails to generalize to new domains.

We propose **Intermediate Preference Representation Alignment (IPRA)**, a method that applies Direct Preference Optimization (DPO) to the LLM to maximize reward signals from a content-based recommendation component. This approach aims to generate enhanced intermediate preference representations from the dialogue context. In addition, we fine-tune the content-based recommender to more effectively leverage the LLM-generated representations. Experiments demonstrate that this two-way alignment improves the accuracy, trial-to-trial robustness, and cross-domain generalization of LLM-powered CRSs.

Inaccuracy and Hallucinations

Because parametric knowledge is probabilistic and pattern-based, it can lead to hallucinations—confidently generating false or misleading information that sounds plausible. The model may blend or distort facts if it has not seen consistent or clear examples in the training data. We address this issue in Chapter 3 in the context of fictional character role-playing. Role-playing

has wide-ranging applications in customer support, embodied agents, and computational social science. The influence of parametric world knowledge in large language models (LLMs) often causes role-playing characters to act out of character and hallucinate information beyond the scope of their knowledge. In this dissertation, we focus on the evaluation and mitigation of hallucinations in fictional character role-play. We introduce a dataset containing over 2,000 characters and 72,000 interviews, including 18,000 adversarial questions. We propose *IPRA*, a role-playing method that mitigates hallucinations by modulating the influence of parametric knowledge using a pre-calibrated confidence threshold. Experiments show that the proposed method improves the factual precision of generated responses by 18% for adversarial questions, with a 44% reduction in temporal hallucinations for time-sensitive interviews.

Chapter 1

Acquiring Parametric Knowledge

1.1 Introduction

Masked Language Modeling (MLM) is widely used for training language models [15, 68, 58, 86]. MLM randomly selects a portion of tokens from a text sample and replaces them with a special mask token (e.g., [MASK]). However, random masking has a few drawbacks — it sometimes produces masks that are too easy to guess, providing a small loss that is inefficient for training; some randomly masked tokens can be guessed with only local cues [45]; all tokens have an identical probability to be masked, while (e.g.) named entities are more important and need special attention [104, 57].

In this chapter, we propose a new strategy for choosing tokens to mask in text samples. We aim to select words with the most information that can benefit the language model, especially

Thomas Edison was an inventor and businessman.					
Thomas	[M]	was an	[M] and businessman.	21.9 ✓	Interesting and challenging!
[M]	[M]	was an inventor and businessman.		17.2	Steve Jobs? Ben Franklin?
Thomas Edison	[M]	an	[M] and businessman.	13.4	The first mask is too easy!
Thomas Edison	[M]	an inventor	[M] businessman.	2.5	Boring! Too easy to guess!

Figure 1.1. The informative scores of randomly sampled masking candidates ($s = 4$). [M] denotes the masked tokens. The pretraining objective of the masked language model (MLM) is to predict the masked tokens based on the context.

for knowledge-intense tasks. To tackle this challenge, we propose *InforMask*, an unsupervised informative masking strategy for language model pretraining. First, we introduce *Informative Relevance*, a metric based on Pointwise Mutual Information (PMI, 19) to measure the quality of a masking choice. Optimizing this measure ensures the informativeness of the masked token while maintaining a moderate difficulty for the model to predict the masked tokens. This metric is based on the statistical analysis of the corpus, which does not require any supervision or external resource.

However, maximizing the total Informative Relevance of a text sample with multiple masks can be computationally challenging. Thus, we propose a sample-and-score algorithm to reduce the time complexity of masking and diversify the patterns in the output. An example is shown in Figure 1.1. For training a language model with more epochs, we can further accelerate the masking process by only running the algorithm once as a preprocessing step and assigning a token-specific masking rate for each token according to their masking frequency in the corpus, to approximate the masking decisions of the sample-and-score algorithm. After this one-off preprocessing step, masking can be as fast as the original random masking without any further overhead, which can be desirable for large-scale distributed language model training of many epochs.

To verify the effectiveness of our proposed method, we conduct extensive experiments on two knowledge-intense tasks — factual recall and question answering. On the factual recall benchmark LAMA [83], InforMask outperforms other masking strategies by a large margin. Also, our base-size model, InformBERT, trained with the same corpus and epochs as BERT [15] outperforms BERT-base on question answering benchmark SQuAD [88, 87]. Notably, on the LAMA benchmark, InformBERT outperforms BERT and RoBERTa [68] models that have $3\times$ parameters and $10\times$ corpus size.

To summarize, our contributions are as follows:

- We propose InforMask, an informative masking strategy for language model pretraining

that does not require extra supervision or external resource.

- We pretrain and release InformBERT, a base-size English BERT model that substantially outperforms BERT and RoBERTa on the factual recall benchmark LAMA despite having much fewer parameters and less training data. InformBERT also achieves competitive results on the question answering datasets SQuAD v1 and v2.

1.2 InforMask

InforMask aims to make masking decisions more ‘informative’. Since not all words are equally rich in information [57], we aim to automatically identify more important tokens (e.g., named entities) and increase their probability to be masked while preserving the factual hints to recover them. On the other hand, we would like to reduce the frequency of masking stop words. Stop words are naturally common in the corpus and they can be important for learning the syntax and structure of a sentence. However, masked stop words can be too easy for a language model to predict, especially in later stages of LM pretraining. Thus, properly reducing the masking frequency of stop words can improve both the efficiency and performance of the model.

1.2.1 Informative Relevance

To generate highly informative masking decisions for a sentence, we introduce a new concept, namely Informative Relevance. Informative Relevance is used to measure how relevant a masked word is to the unmasked words so that it can be meaningful and predictable. The Informative Relevance of a word is calculated by summing up the Pointwise Mutual Information (PMI, 19) between the masked word and all unmasked words in the sentence. PMI between two words w_1 and w_2 represents how ‘surprising’ is the co-occurrence between two words, accounting for their own probabilities. Formally, the PMI of the combination $w_1 w_2$ is defined as:

$$\text{pmi}(w_1, w_2) = \log \frac{p(w_1, w_2)}{p(w_1)p(w_2)} \quad (1.1)$$

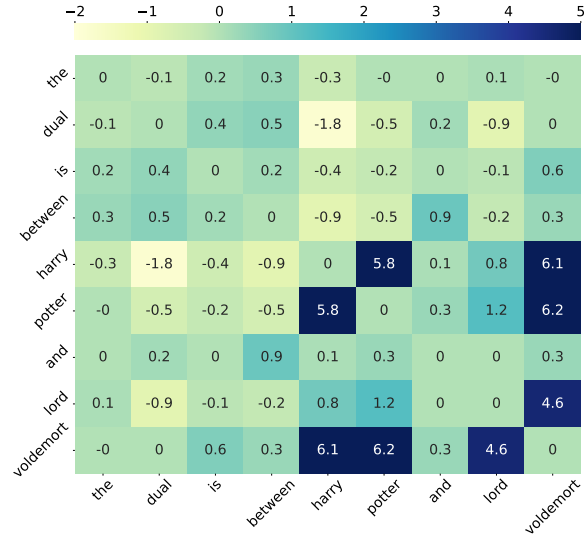


Figure 1.2. The PMI matrix of the words in the sentence ‘The dual is between Harry Potter and Lord Voldemort.’

The PMI matrix is calculated corpus-wise. Note that instead of using bigrams (i.e., two words have to be next to each other), we consider the skip-gram co-occurrence within a window. The window size is selected in a way that enables sentence-level co-occurrence to be considered as well as local co-occurrence.

Maximizing the Informative Relevance enables the model to better memorize knowledge and focus on more informative words. Since Informative Relevance is calculated between a masked word and the unmasked words, it also encourages hints to be preserved so that the model can reasonably guess the masked words. As shown in Figure 1.2, the words inside a named entity have a high PMI (e.g., ‘Harry-Potter’ and ‘Lord-Voldemort’) while the two closely related entities also show a high PMI (e.g., Harry-Voldemort). Thus, if we are asked to mask one word, we would mask ‘Voldemort’ since it has the highest Informative Relevance with the remaining words (by summing up the last row or column).

Algorithm 1. InforMask Algorithm

```
1:  $D \leftarrow$  Set of text
2:  $s \leftarrow$  Size of randomly sampled candidates
3:  $F_i^d \leftarrow$  Informative score for  $i$ -th masking candidate for text  $d$ 
4: for  $d \in D$  do
5:   for  $i = 1, 2, \dots, s$  do
6:     Generate  $i$ -th masking candidate for  $d$ 
7:      $M_i^d \leftarrow$  Masked Tokens
8:      $U_i^d \leftarrow$  Unmasked Tokens
9:      $F_i^d \leftarrow 0$ 
10:    for  $w_1 \in M_i^d$  do
11:      for  $w_2 \in U_i^d$  do
12:         $F_i^d = F_i^d + \text{pmi}(w_1, w_2)$ 
13:      end for
14:    end for
15:  end for
16:  Choose candidate with maximum  $F_i^d$ 
17: end for
```

1.2.2 Scoring Masking Candidates

One text sample can have multiple masks. Thus, we define the informative score of a masking decision as the sum of the Informative Relevance of each masked token. However, given the PMI matrix, finding the best k words to mask (i.e., the masking decision with the highest informative score) in a sentence of n words is time-consuming. Iterating all possibilities has time complexity $O(C_n^k)$. By converting it to a minimum cut problem, the time complexity can be reduced to $O(n^2 \log n)$ [101], which is still prohibitive in practice.

Therefore, we propose to sample s random masking candidates and then rank them by calculating their informative scores. As shown in Figure 1.1, we randomly generate four masking candidates and rank them by their informative scores. We select the candidate with the highest score. This allows us to make a masking decision with time complexity $O(kn)$. Random sampling also introduces more diverse patterns for masking, which could help training of language models and prevent overfitting. This process is illustrated in Algorithm 1.

Table 1.1. Statistics of LAMA [83] dataset, used for factual knowledge probing.

Data Subset	#Relations	#Samples
ConceptNet	1	29774
Squad	1	305
GoogleRE	3	4994
TREx	41	34032
Total	46	69105

Table 1.2. Statistics of extractive question answering dataset SQuAD v1 and v2 [88, 87].

Dataset	SQuAD v1	SQuAD v2
#Examples	108k	151k
#Negative Examples	0	54k
#Articles	536	505

1.2.3 Token-Specific Masking Rates

Algorithm 1 is already usable by processing the input text on the fly. However, to avoid overfitting, masking should change across epochs. This means we have to run Algorithm 1 every epoch, creating a bottleneck for pretraining. To address this efficiency issue, we use token-specific masking rates to approximate the masking decisions of InforMask. Specifically, we generate masks for a corpus using Algorithm 1, and then count the frequency of each token in the vocabulary to be masked as their token-specific masking rates. Note that in this way, Algorithm 1 is only executed once, as a preprocessing step. Furthermore, we can use a small portion of the corpus to calculate the token-specific masking rates, making it even faster.¹ After this, we can perform random masking, except that every token has its own masking rate.

¹For the Wikipedia corpus, the average rate of change for token-specific masking rates falls below 0.8% after processing only 1% of the corpus.

Table 1.3. Performance of different masking strategies and models on LAMA [83]. (a) We compare the models trained with different masking strategies for 3 epochs. (b) We compare InforBERT, a BERT model trained with InforMask for 40 epochs with BERT and RoBERTa models.

	Model	#Param.	Corpus Size	Epochs	LAMA [83]				Overall
					ConceptNet	Squad	GoogleRE	TREx	
(a)	Random [15]	125M	16 GB	3	0.091	0.124	0.396	0.582	0.549
	Span [45]	125M	16 GB	3	0.056	0.102	0.377	0.524	0.495
	PMI [57]	125M	16 GB	3	0.075	0.115	0.396	0.552	0.522
	InforMask	125M	16 GB	3	0.109	0.133	0.410	0.627	0.591
(b)	BERT-base	110M	16 GB	40	0.191	0.229	0.340	0.587	0.553
	BERT-large	340M	16 GB	40	0.218	0.284	0.354	0.621	0.585
	RoBERTa-base	125M	160 GB	40	0.223	0.307	0.423	0.630	0.592
	RoBERTa-large	355M	160 GB	40	0.260	0.329	0.435	0.672	0.632
	InforBERT	125M	16 GB	40	0.201	0.384	0.509	0.739	0.698

Table 1.4. Performance of InforMask on SQuAD v1 and v2 [88, 87] development set.

	Model	#Param.	Corpus Size	Epochs	SQuAD v1		SQuAD v2	
					F1	EM	F1	EM
(a)	Random [15]	125M	16 GB	3	79.08	69.44	66.48	63.15
	Span [45]	125M	16 GB	3	78.88	69.04	64.95	61.38
	PMI [57]	125M	16 GB	3	80.31	70.98	66.25	62.82
	InforMask	125M	16 GB	3	80.47	71.41	67.29	63.90
(b)	BERT-base	110M	16 GB	40	81.07	88.52	72.35	75.75
	InforBERT	125M	16 GB	40	81.22	88.61	72.71	75.86

1.3 Experiments

1.3.1 Pretraining Corpus

Following BERT [15], we use the Wikipedia and Book Corpus datasets available from Hugging Face [60]. The corpus contains $\sim 3.3\text{B}$ tokens. To be consistent with BERT, we use an overall masking rate of 15%. The PMI matrix is calculated on the Wikipedia corpus, with a size of $100\text{k} \times 100\text{k}$. Word co-occurrence statistics are computed with a window size of 11. We set the candidate sampling size per document s to 30. It takes ~ 4 hours to preprocess and generate token-specific masking rates on a 16-core CPU server with 256 GB RAM.

1.3.2 Evaluation Benchmarks

To evaluate different masking strategies, we use the LAMA benchmark [83] to test the knowledge of the models. LAMA is a probe for analyzing the factual and commonsense knowledge contained in pretrained language models. Thus, it is suitable for evaluating the knowledge learned during pretraining. LAMA has around 70,000 factual probing samples across 46 factual relations. A summary of the benchmark is shown in Table 1.1. We use Mean Reciprocal Rank (MRR) as the metric for factual recall performance.

In addition to the knowledge probing task, we also conduct experiments on real-world question answering datasets, which requires commonsense knowledge as well. We conduct experiments on SQuAD v1 and v2 [88, 87] and report the F1 and Exact Match (EM) scores on the development set. The statistics of the benchmark are shown in Table 1.2.

1.3.3 Baselines

We compare InforMask in two settings: **(a)** We use the same tokenizer and hyperparameters to pretrain BERT random masking [15], SpanBERT [45] and PMI-Masking [57] for 3 epochs. The choice of 3 epochs is according to our limited computational budget. **(b)** We continue training InforMask until 40 epochs. The 40-epoch model is denoted as *InformBERT*.

We compare InformBERT to BERT-base [15], which is trained with the same corpus for 40 epochs as well. We also include results of BERT-large and RoBERTa for reference, though they are either larger in size or trained with more data and thus are not directly comparable.

1.3.4 Training Details

Our implementation is based on Hugging Face Transformers [118]. We train the baselines and our model with 16 Nvidia V100 32GB GPU. For our model and all baselines trained, we use a fixed vocabulary size of 50,265. The model architecture is a base-size BERT model, with 12 Transformer layers with 12 attention heads. The hidden size is set to 768. The overall batch size is 256. We use an AdamW optimizer [70] with a learning rate of $5e-5$. Note that we do not perform any hyperparameter searching or tuning for any model (including InformBERT) given our limited computational budget.

1.3.5 Impact of Masking Strategies

We conduct a fair comparison among different masking strategies, using the same tokenization and hyperparameters. As shown in Table 1.3(a), InforMask outperforms other masking strategies by a large margin on all subsets of LAMA [83]. As shown in Table 1.4(a), on both SQuAD v1 and v2 [88, 87], InforMask outperforms other masking strategies. Notably, PMI-Masking achieves higher performance on SQuAD while underperforming random masking on LAMA (to be detailed shortly) but our InforMask achieves better results on both of them.

Also, we compare our 40-epoch InformBERT model with BERT and RoBERTa models. As shown in Table 1.3(b), InformBERT outperforms the BERT model trained with the same epochs and corpus by 0.145 overall. It also achieves higher performance than RoBERTa-base, despite being trained with 10% of RoBERTa’s corpus size. To our surprise, it also outperforms both BERT-large and RoBERTa-large, with only 1/3 parameters. Moreover, InformBERT outperforms BERT-base for fine-tuning on SQuAD v1 and v2, demonstrating its capability for downstream question answering, as shown in Table 1.4(b).

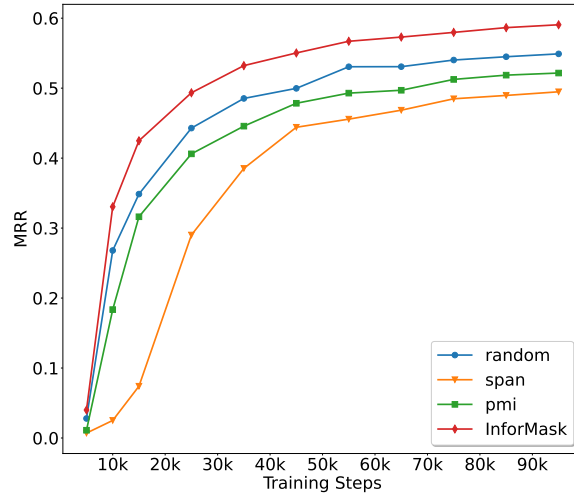


Figure 1.3. Macro average MRR of different masking strategies on LAMA, evaluated every 10k steps.

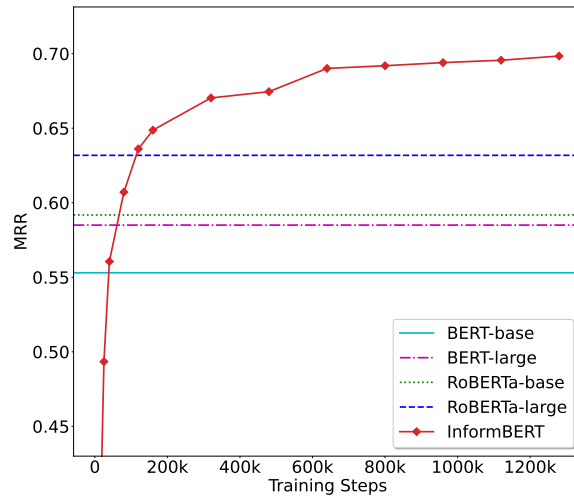


Figure 1.4. Performance of InformBERT for the full pretraining process. It achieves comparable performance with BERT-base after 40k training steps and even RoBERTa-large after 120k training steps.

1.3.6 Training Dynamics

As shown in Figure 1.3, InforMask demonstrates an outstanding training efficiency. InforMask outperforms other masking strategies from the beginning of the training process and keeps the lead through the training. Notably, span masking and PMI-Masking underperform random masking, indicating their inability on the knowledge-intensive task. Span masking also

significantly underperforms other masking strategies in the early stage of pretraining, suggesting it may take longer to train the model. For the entire pretraining process, as shown in Figure 1.4, the model trained with InforMask outperforms BERT and RoBERTa with fewer than $\sim 15\%$ of the training steps, verifying the efficiency of our masking strategy.

1.3.7 Impact on Stop Words and Entities

As shown in Figure 1.5, without explicitly specifying the stop words, InforMask can identify the stop words and reduce their probability to be masked. InforMask can also automatically increase the masking probability of named entities. The average masking probability of named entities is 0.25 with a standard deviation of 0.07, while the overall masking probability of all tokens is around 0.15.² This allows the model to focus on more important tokens and maintain an appropriate difficulty of prediction, facilitating the pretraining process.

1.3.8 Impact of Token-Specific Masking Rates

As mentioned before, the use of token-specific masking rate can enormously save time and RAM for data processing, as spending hours of processing for each epoch can be infeasible and becomes a bottleneck for distributed training. Another possible solution is to loop the same masked data for every epoch. Thus, we conduct an experiment to compare the two solutions: approximation and repetition. Note that for simplicity, the token-specific masking rate is applied from the first epoch. As shown in Figure 1.6, our approximation strategy keeps outperforming the repetition strategy even in the first epoch. As we analyze, this can be attributed to the more diverse patterns introduced during the approximation. Also, the performance of the model trained with the repetition strategy converges or even slightly declines after 60k training steps while the performance of the model trained with approximation keeps increasing.

²We use an off-the-shelf named entity recognition system to verify the effectiveness of our approach only. It is not a necessary component of the proposed system.

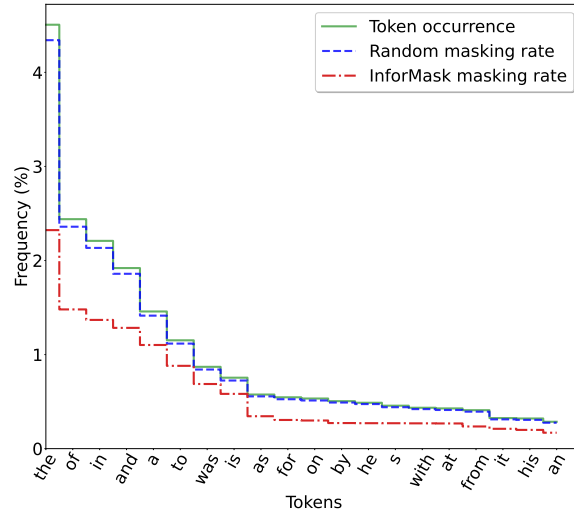


Figure 1.5. Frequency of common stop words and their corresponding masking rates by InforMask.

1.3.9 InforMask vs. PMI Masking

PMI Masking [57] uses PMI to mask a span of correlated tokens. A named entity often constitutes a correlated span and therefore, is more likely to be masked in PMI-Masking. As mentioned before, we observe that PMI-Masking performs worse than random masking on LAMA (see Figure 1.3).

To investigate the reason, we compute the individual masking rates of some tokens according to each masking policy. As shown in Figure 1.7, we can see that PMI-Masking increases the masking rate of tokens that are part of correlated spans. However, it decreases the masking rate of tokens that are not within any correlated span, even if that token is a named entity. Consider the token ‘Airways’ for example. This token may be part of a correlated span such as ‘British Airways’ or ‘Qatar Airways’. PMI-Masking, therefore, increases the masking rate of this token compared to random masking. On the other hand, the tokens ‘Colorado’ and ‘Nairobi’, which are unigram named entities, are less likely to be masked, compared to random masking. Given that the overall masking rate is fixed and PMI-Masking favors correlated spans, the masking rates of ‘Colorado’ and ‘Nairobi’ inevitably get lower. This can be the reason behind

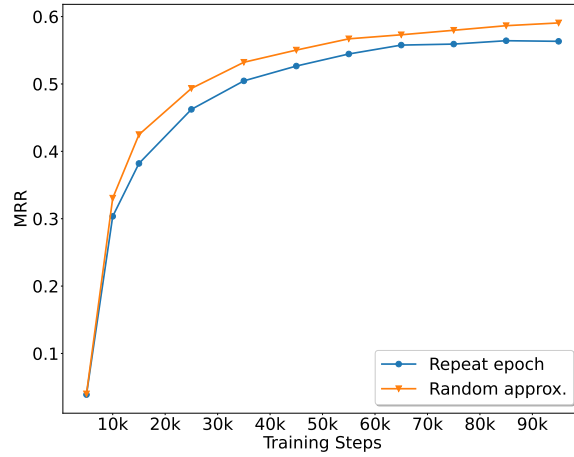


Figure 1.6. Comparison between looping the same data and using token-specific masking rate to approximate the masking decisions. The models are trained for 3 epochs.

PMI-Masking’s failure.

In contrast, InforMask uses PMI to compute the individual Informative Relevance of tokens. It can increase the masking rate of tokens with high informative saliency, regardless of whether they are part of a correlated span or not. This helps InforMask achieve superior factual recall performance.

1.3.10 Case Study

Table 1.5 shows the example knowledge probes and answers produced by InformBERT and RoBERTa. For the query ‘SpeedWeek is an American television program on [MASK].’, RoBERTa is unable to produce the correct answer in the top-3 predictions. But InformBERT correctly predicts ‘ESPN’ to be the top candidate. Similarly, InformBERT correctly predicts the answer ‘bishops’ for the query ‘The chief administrators of the church are [MASK].’ RoBERTa is unable to predict the answer and produces more generic words such as ‘men’, ‘women’, and ‘Christians’.

We summarize the errors into two notable categories. They are relevant for all the models involved, not just InformBERT. First, we observe that many errors involve rare named entities. Some named entities are less frequent so the model is unable to learn anything useful about

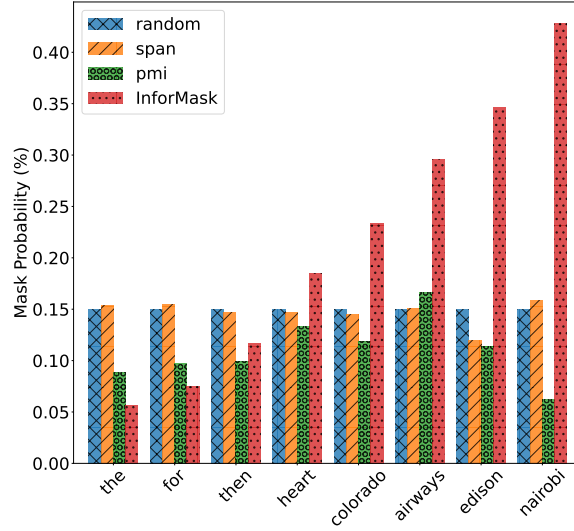


Figure 1.7. Masking rate of selected tokens according to different masking policies.

Table 1.5. Some examples of InformBERT and RoBERTa-base predictions on LAMA [83]. We show the queries and the ground-truth answers with the model predictions. We only show the top-3 predictions made by each model.

Query	Ground Truth	InformBERT		RoBERTa-base	
		Prediction	Score	Prediction	Score
Antoine Coypel was born in [MASK].	paris	france	0.09	montreal	0.12
		paris	0.08	toronto	0.03
		haiti	0.04	paris	0.03
SpeedWeek is an American television program on [MASK].	espn	espn	0.20	cbs	0.18
		nbc	0.10	cnbc	0.13
		mtv	0.09	spike	0.10
Phil Harrison is a corporate vice president of [MASK].	microsoft	microsoft	0.20	intel	0.06
		ibm	0.15	ibm	0.05
		motorola	0.05	microsoft	0.03
Laurent Casanova was a [MASK] politician.	french	french	0.43	young	0.13
		canadian	0.32	french	0.09
		haitian	0.05	successful	0.04
The chief administrators of the church are [MASK].	bishops	bishops	0.13	men	0.17
		priests	0.07	christians	0.09
		appointed	0.06	women	0.08

them, or they occur so rarely that they do not even appear in the language model vocabulary. We found that around 19% of the errors made by our model on the LAMA benchmark is associated with out-of-vocabulary tokens. Second, it is challenging for a language model to predict the granularity of the fact being asked or distinguish it from an alternate fact that may hold for a query. For the example query ‘Antoine Coypel was born in [MASK].’, the LAMA dataset has only one true label ‘Paris’. In this example, InformBERT prefers the name of the country (‘France’) over the name of a city (‘Paris’). This confusion is related to the granularity of location and both answers can be considered correct. However, it is being classified as an error because the labels in the test set are not comprehensive.

Another type of confusion can be found for RoBERTa with the query ‘Laurent Casanova was a [MASK] politician.’. The model is trying to decide whether to use the adjective ‘young’, ‘French’, or ‘successful’. In theory, these three adjectives may be valid simultaneously for the same entity. It can be challenging for the language model to pick the expected one in the context.

1.4 Conclusion

In this chapter, we propose InforMask, an unsupervised masking policy that masks tokens based on their informativeness. InforMask achieves superior performance in knowledge-intense tasks including factual recall and question answering. We explore the impact of different masking strategies on learning factual and commonsense knowledge from pretraining and analyze why previously proposed masking techniques are suboptimal. Future work may involve scaling up the pretraining and exploring more factors for knowledge acquisition during unsupervised text pretraining.

Chapter 1, in part, is a reprint of material that appears in the paper titled “InforMask: Unsupervised Informative Masking for Language Model Pretraining” by Nafis Sadeq, Canwen Xu, and Julian McAuley, which was published in *Empirical Methods in Natural Language Processing*, 2022. The dissertation author was the lead author of this paper.

Chapter 2

Exploiting Parametric Knowledge

2.1 Introduction

CRSs aim to help users find desired items through multi-turn natural language interactions [22]. While some recommender systems perform user preference elicitation via passive observation of user behaviors (e.g., clicks, ratings) [50, 34], CRSs actively interact with users to elicit preferences, make recommendations, and refine them based on user feedback expressed in natural language [64, 125]. CRSs typically consist of two components that work together to generate responses during a conversation [64, 9, 130]. The first is a recommendation component, responsible for identifying relevant items based on user history (e.g., dialogue context), and the second is the conversation component, which generates the text of a natural language question or response in the conversation.

The recommendation component is expected to incorporate the semantic information from the dialogue context into the user preference representation for generating recommendations [64,

¹We will refer to this model as gpt-3.5-turbo throughout the chapter.

Table 2.1. Cross-trial robustness issue for a hybrid CRS with GPT-3.5-Turbo [78] and Text embedding, reported in terms of the coefficient of variation of reciprocal ranks.

Model	LLM Decoding Temperature				
	0.6	0.7	0.8	0.9	1.0
gpt-3.5-turbo-1106 ¹	1.35	1.39	1.49	1.67	1.74

Table 2.2. Cross-domain generalization issue of a hybrid CRS with GPT-3.5-Turbo and Text embedding.

Domain	Hit@1	Hit@10	Hit@50
Movie	0.021	0.107	0.228
Electronics	0.010	0.043	0.117

9, 130, 114]. Recent advances in LLMs [79, 6] have revolutionized CRSs because of their strong ability to understand natural language preference signals [35, 113, 24, 21, 20]. However, the introduction of LLMs into CRSs presents new challenges. The static parametric knowledge of LLMs cannot keep track of a large and evolving item space, which limits the capabilities of LLM-prompting-based CRSs [122, 36]. To address this issue, some studies have explored combining LLMs with separate recommendation modules, where LLMs produce intermediate representations to drive the external recommendation components [21, 20, 75].

In this chapter, we aim to address two key challenges associated with such LLM-powered CRSs. First, LLM-generated representations lack consistency across trials. Given the same dialogue context, LLMs will generate very different intermediate preference representations across multiple trials, leading to a large variance in performance. To demonstrate this issue, we randomly choose 100 samples from the ReDial [64] test set and perform 30 trials for each sample using an LLM with a content-based recommender baseline. We compute the reciprocal ranks of the ground-truth item for each trial and report the average coefficient of variation in Table 2.1. We find that even with low decoding temperatures, the standard deviation of the reciprocal rank can be as high as 135% relative to the mean, indicating a robustness issue. Second, LLMs’ ability to generate preference representations often does not generalize to new domains. For example, dialogues related to E-commerce domains such as electronics may be rare in the LLMs’ pretraining corpus compared to more popular domains such as movies. Thus, the LLMs’ ability to identify item attributes that explain user preference in those domains will be limited, leading to performance degradation. To study this issue, we have constructed a new CRS dataset in the electronics domain by fusing YouTube comments about smartphones with corresponding

GSMarena [27] metadata. We compare the performance of the aforementioned CRS baseline on the movie and electronics domains in Table 2.2. We find that the same CRS paradigm suffers from a 49% degradation in hit rate (top-50 recommendations) in the electronics domain compared to movies.

In this study, we exploit an external content-based recommender as a reward model to improve the quality and robustness of the intermediate user preference representation generated by LLMs. We propose IPRA, a weak-to-strong representation alignment method using recommendation as a reward model, which improves both cross-trial robustness and cross-domain generalization of CRSs. IPRA performs recommendation in two steps: (1) a *dialogue-to-preference* step, where an LLM produces an intermediate user preference representation from the dialogue context, and (2) a *preference-to-recommendation* step, where a content-based recommendation module exploits the intermediate preference representation to generate a candidate recommendation list. During training, we perform multiple trials for the preference generation step to generate many instances of preference representation. Then, a recommendation module is used to generate a candidate recommendation for each instance of preference representation. The reciprocal rank of the ground-truth item in the candidate list is used as a fitness score for the intermediate preference representation. We use this fitness score to construct a preference alignment corpus and apply DPO [85] to improve the quality of preference representation. Additionally, we fine-tune a text encoder as a content-based recommender to find the most relevant items given the LLM-generated intermediate representations.

Our contributions can be summarized as follows:

- We propose IPRA: a method that aligns the intermediate user preference representation generated by LLMs to improve accuracy, cross-trial robustness, and cross-domain generalization of CRSs. We demonstrate the effectiveness of IPRA using three LLM backbones on movie and electronics domains.
- We construct and release GYPHONE, a CRS dataset on phone recommendation by fusing

Table 2.3. Frequently used notations for IPRA.

Notation	Description
$\mathcal{D}_1$	CRS dataset with dialogue-recommendation pairs (d_i, r_i)
$\mathcal{D}_2$	Alignment dataset with preference tuples
$\mathcal{I}$	The set of all target items
d_i	Dialogue context of the i -th sample
r_i	Ground truth recommendation for the i -th sample
p_i^j	j -th preference representation for i -th dialogue
c_i^j	Candidate recommendations based on preference p_i^j
π_θ	LLM parameterized by θ for generating preferences
π_{ref}	Fixed reference LLM for baseline probabilities
$s(\cdot)$	Fitness of a preference representation
$f(\cdot)$	Text embedding function of the shared text encoder
$\mathcal{L}_{\text{DPO}}$	DPO loss encouraging preference alignment
$\mathcal{L}_{\text{triplet}}$	Triplet loss for text-encoder fine-tuning

YouTube conversations about smartphones with corresponding phone metadata from GSMArena. GYPhone has 6,860 conversations, 190,854 utterances involving 5,008 users and 592 items.

2.2 IPRA

Let us assume we have a CRS dataset $\mathcal{D}_1 = \{(d_i, r_i)\}_{i=1}^n$, where d_i denotes the dialogue context of the i -th sample, and r_i refers to the ground truth recommendation in the next dialogue turn. Let c_i be a list of recommendation candidates for the i -th sample. Our proposed method performs recommendation in two steps: (1) a *dialogue-to-preference* step, where an LLM produces an intermediate user preference representation p_i from the dialogue context d_i ; (2) a *preference-to-recommendation* step, where an external recommendation module uses the intermediate preference representation p_i to generate the list of candidate recommendations c_i .

2.2.1 Preference Alignment Corpus

During training, we perform multiple trials for the *dialogue-to-preference* step for each sample to generate m instances of the preference representation. Let p_i^j denote the j -th prefer-

ence instance for the i -th training sample. We use a recommendation module to generate the corresponding candidate set c_i^j from the intermediate preference representation. The fitness of the intermediate representation is calculated by the reciprocal rank of the ground truth recommendation r_i within c_i^j . Let $s(\cdot)$ be the fitness score of intermediate representation. Mathematically, it can be formulated as:

$$s(p_i^j) = \frac{1}{\text{rank}(r_i, c_i^j)}$$

Lower ranks correspond to higher positions in the list (i.e., better recommendations), and the reciprocal rank gives higher weight to correct recommendations ranked nearer the top. To construct the preference alignment corpus $\mathcal{D}_2$, we consider all possible representation pairs such that the chosen representation has a higher score than the rejected one. Mathematically, it can be formulated as:

$$\mathcal{D}_2 = \left\{ (d_i, p_i^j, p_i^k) : 1 \leq j, k \leq m, s(p_i^j) > s(p_i^k) \right\}_{i=1}^n \quad (2.1)$$

2.2.2 Direct Preference Optimization

To align the *dialogue-to-preference* generation process to achieve better recommendation outcomes, we define the Direct Preference Optimization (DPO) [85] objective using the preference alignment corpus $\mathcal{D}_2$. For each preference pair (p_i^j, p_i^k) , where p_i^j is preferred over p_i^k as determined by their fitness scores, we aim to maximize the probability that the LLM assigns higher preference to p_i^j over p_i^k .

Let $\pi_\theta(p_i | d_i)$ be the likelihood of generating a preference representation p_i given dialogue context d_i , parameterized by θ . We also consider a fixed reference LLM $\pi_{\text{ref}}(p_i | d_i)$ that provides baseline generation probabilities. The DPO loss for a single sample in $\mathcal{D}_2$ is defined

Table 2.4. LLM prompts related to steps in IPRA and dataset construction for GYPhone.

Step	Prompt Format
<i>Dialogue-to-preference</i>	"Given the following dialogue context, generate a summarized one paragraph description of this user's preference on <domain>. Try to identify the attributes this user may like or dislike. Dialogue context: <dialogue_context>"
<i>Preference-to-recommendation</i>	"Given the user preference description, generate a list of <domain> recommendation for this user. Each <domain> should be separated by newline character. No other text or description should be present. Dialogue context: <dialogue_context>"
<i>Dialogue-to-recommendation</i>	"Given the following dialogue context, generate a list of <N> <domain> recommendations for this user. Each <domain> should be separated by newline character. No other text or description should be present. Dialogue context: <dialogue_context>"
<i>Item Description</i>	"You are given structured item metadata from the domain <domain>, represented in dictionary format. Your task is to convert this structured data into a fluent, unstructured passage that describes the item in a natural, engaging, and informative way. Avoid listing key-value pairs. Instead, write a paragraph (or more, if necessary) that smoothly incorporates the metadata details. Here is the item metadata <metadata_dictionary>. Now write an unstructured passage that describes this item, suitable for use in product descriptions, catalogs, or narrative content."
<i>Relevance Filtering</i>	"Determine whether the following utterance is relevant to the smartphone domain. Return 'Yes' if it is relevant, otherwise return 'No'. Utterance: <utterance>"
<i>Phone Tagging</i>	"Identify all mentions of smartphone names in the following utterance. Return a list of the mentioned smartphones, or an empty list if none are found. Utterance: <utterance>"

as:

$$\mathcal{L}_{\text{DPO}}^{(i,j,k)}(\theta) = -\log \sigma \left(\beta \cdot \left[\log \frac{\pi_{\theta}(p_i^j | d_i)}{\pi_{\text{ref}}(p_i^j | d_i)} - \log \frac{\pi_{\theta}(p_i^k | d_i)}{\pi_{\text{ref}}(p_i^k | d_i)} \right] \right) \quad (2.2)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and β is a temperature hyperparameter that controls the preference intensity. The overall DPO objective is the average loss over all training preference pairs:

$$\mathcal{L}_{\text{DPO}}(\theta) = \frac{1}{|\mathcal{D}_2|} \sum_{(d_i, p_i^j, p_i^k) \in \mathcal{D}_2} \mathcal{L}_{\text{DPO}}^{(i,j,k)}(\theta)$$

This formulation encourages the LLM to assign higher likelihood to preference representations that lead to better recommendation rankings.

2.2.3 Text Encoder Fine-Tuning

To enhance the effectiveness of the *preference-to-recommendation* step, we fine-tune the text encoder used to retrieve items based on the intermediate preference representation p_i . The objective is to ensure that the encoder projects similar dialogue preference and item metadata pairs close together in the embedding space, while dissimilar ones are projected farther apart. We formulate the fine-tuning as a metric learning problem using the triplet loss. For each training instance, we construct a triplet (p_i, r_i^+, r_i^-) , where:

- p_i is the LLM-generated preference representation from dialogue context d_i ,
- r_i^+ is the ground-truth item (positive sample),
- r_i^- is a randomly-selected incorrect item (negative sample).

Let $f(\cdot)$ be the embedding function of the shared text encoder. The triplet loss for one

training instance is defined as:

$$\mathcal{L}_{\text{triplet}}^{(i)} = \max(0, \|f(p_i) - f(r_i^+)\|_2^2 - \|f(p_i) - f(r_i^-)\|_2^2 + \alpha) \quad (2.3)$$

Here, α is a margin hyperparameter that enforces minimum separation between the positive and negative distances. The overall training objective minimizes the average triplet loss over the dataset:

$$\mathcal{L}_{\text{triplet}} = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_{\text{triplet}}^{(i)}$$

This objective drives the encoder to produce a shared representation space where relevant items are ranked higher based on their proximity to the LLM-generated user preference representation. Before encoding item metadata, we convert item descriptions into a passage containing complete sentences instead of directly encoding structured metadata. This helps mitigate the domain shift of encoded features between unstructured user preference representation and structured metadata. The conversion is performed by prompting an LLM with structured item metadata. The relevant prompt is shown in Table 2.4, and the impact of metadata transformation is shown in Table 3.3 in the ablation study.

2.2.4 Item Retrieval During Inference

During inference, IPRA performs recommendation by retrieving the most relevant items from the entire item space using the preference representation generated by the LLM. Given a dialogue context d_i , the LLM first produces a preference representation p_i during inference. The retrieval step involves computing cosine similarity between encoded preference representation $f(p_i)$ and the pre-computed embeddings of all items in the corpus:

$$\text{sim}(p_i, r) = \frac{f(p_i) \cdot f(r)}{\|f(p_i)\|_2 \cdot \|f(r)\|_2}, \quad \forall r \in \mathcal{I}$$

where $\mathcal{J}$ is the set of item candidates, and $f(\cdot)$ is the shared encoder used during training. The top- k items with the highest similarity scores are selected to form the final recommendation list c_i :

$$c_i = \text{TopK}_{r \in \mathcal{J}} \text{sim}(p_i, r)$$

2.3 Experiments

2.3.1 Datasets

We perform experiments on the movie and electronics product domains. For movie recommendations, we use ReDial [64]. ReDial has 10k conversations and 182k utterances involving 956 users. For the electronics domain, we construct our own CRS dataset named GYPhone by collecting YouTube conversations regarding smartphones from the GSMArena YouTube channel. The data collection is performed via the YouTube Data v3 API [123]. We then use GPT-3.5-Turbo to identify and filter out utterances that are irrelevant to the domain. All smartphone mentions in the conversations are identified by prompting GPT-3.5-Turbo. We then connect the smartphone mentions in the conversations with corresponding phone metadata from GSMArena.

In order to tackle the data sparsity issue, we only consider utterances that have more than five replies as a single conversation. Similarly, we remove all users who have fewer than five utterances in the dataset. The statistics of the resulting dataset are shown in Table 2.5 in comparison to ReDial. The prompts related to data processing are given in Table 2.4. The decoding during data processing was performed with a temperature of 0.7 and top-p of 1.0.

We use 10% of the dialogues from each dataset as the test set and the rest as the training set. We use 30 trials per training sample to construct the preference alignment corpus for intermediate representation. All representation pairs that satisfy the condition in Equation 2.1 are added to the alignment corpus. To ensure variety in user preference representation, we use a

Table 2.5. Dataset statistics for movie and electronics domain.

Attribute	ReDial	GYPhone
Domain	Movie	Electronics
# Utterances	182150	190854
# Dialogues	10006	6860
# Users	956	5008
# Items	2460	592

decoding temperature of 1.0 for alignment data generation.

2.3.2 Baselines & Metrics

The following models are used as baselines for the recommendation. We use $d2p$, $p2r$, and $d2r$ to denote dialogue-to-preference, preference-to-recommendation, and dialogue-to-recommendation modules respectively within the baseline names.

ReDial [64] This baseline first predicts the user sentiment on movies mentioned in the dialogue context. Then, the sentiment is transformed into movie ratings and used by an autoencoder-based recommender [95]. The autoencoder is trained using an external user-item interaction dataset (MovieLens [32]).

KBRD [9] Preference signals from the dialogue context are enhanced with an item-oriented knowledge graph. A Relational Graph Convolutional Network [94] is trained on an external knowledge graph (DBpedia [53]).

KGSF [130] Preference signals from the dialogue context are enhanced with both an item-oriented knowledge graph and a word-oriented knowledge graph. The item-oriented knowledge graph is trained on DBpedia [53], and the word-oriented knowledge graph is trained on ConceptNet [100].

UniCRS [114] Contextual word embeddings from RoBERTa [68] are fused with entity embeddings trained on DBpedia, and prompt learning is applied to enhance recommendation performance.

Text embedding (d2r) [89] This baseline encodes the dialogue context with a text encoder and performs a similarity search with encoded representations of item metadata to generate recommendations. We use the all-MiniLM-L6-v2 model from Siamese-BERT [89] for encoding both the dialogue and items.

LLM (d2r) [35] The dialogue context is provided to LLMs with a prompt for a zero-shot recommendation. We evaluate both GPT-3.5-Turbo and GPT-4 in this setting. The prompts for the direct recommendation baseline are listed in Table 2.4.

LLM (d2r) + Text embedding [113] A few recommendation candidates are generated with an LLM in a zero-shot manner based on the dialogue context. Then, similar items are retrieved by comparing the encoded representations of the recommended and target items. We use OpenAI text-embedding-ada-002 [77] for this baseline, following Wang et al. [113].

LLM (d2r) + MESE [113] This baseline first uses an LLM to generate item recommendations. Then, an encoder model is used to retrieve similar items from the dataset based on the recommended items. Encoder models are fine-tuned in a supervised manner to bridge the semantic gap between dataset items and those generated by the LLM.

LLM (d2p) + Text embedding (p2r) An intermediate description of user preference is generated with an LLM based on the dialogue context. Then, similar items in the dataset are retrieved based on the similarity between the encoded representations of the preference description and the encoded representations of the target items. We use all-MiniLM-L6-v2 [89] as the text embedding model for this baseline.

In order to evaluate the recommendation system’s performance, we use the hit rate metric, denoted as $\text{Hit}@k$ for $k = 1, 10, 50$. To evaluate the cross-trial robustness of the recommendation system, we compute the average coefficient of variation (CV) of reciprocal ranks across all samples in the test set, using 30 independent trials per sample. For a given sample i , CV is

defined as:

$$\text{CV}_i = \frac{\sigma_i}{\mu_i},$$

where σ_i and μ_i are the standard deviation and mean of the reciprocal ranks across 30 trials for sample i , respectively. The average CV over N test samples is then given by:

$$\text{Average CV} = \frac{1}{N} \sum_{i=1}^N \text{CV}_i.$$

2.3.3 Training Details

We experiment with three LLM backbones for the *dialogue-to-preference* step of IPRA: Llama-3-8B-Instruct [17], Mistral-7B-Instruct-v0.3 [44], and Qwen2.5-7B-Instruct [3]. We perform DPO on these models using LoRA adaptation, with a rank of 8 and a scaling factor of 16. The preference intensity hyperparameter β from Equation 2.2 is set to 0.1. We use a learning rate of 10^{-4} for 4 epochs during alignment, using a batch size of 1 and gradient accumulation steps of 8. The training is performed using 2×48 GB NVIDIA RTX A6000 GPUs for three days.

For the *preference-to-recommendation* step, we fine-tune a text embedding model for content-based recommendation. We use the all-MiniLM-L6-v2 model from Siamese-BERT [89] to encode both the dialogue context and item metadata. Item descriptions from MovieLens and GSMArena are used for the movie and phone domains, respectively. To mitigate the domain shift between unstructured text (i.e., user preference representation) and structured text (i.e., item metadata), we augment paragraph-level item descriptions by prompting an LLM with item metadata. We then fine-tune the text embedding model to maximize the text embedding similarity between ground truth items and the corresponding dialogue context. We fine-tune using a triplet objective for 10 epochs with a batch size of 16. We use the Adam optimizer with a learning rate of $2e-5$. The negative dialogue-item pairs for the triplet objective are picked randomly. The margin hyperparameter α in the triplet objective is set to 1.0 (Equation 2.3).

2.3.4 Recommendation Performance

Tables 2.6 and 2.7 summarize the recommendation accuracy of IPRA across two datasets: ReDial and GYPhone, using standard Hit@K metrics. Several core findings emerge from our analysis:

1. IPRA consistently outperforms both LLM-only and hybrid baselines. Across both domains and all evaluation thresholds, IPRA demonstrates significant improvements over legacy CRS models, direct generation approaches using LLMs, and modular hybrid baselines. On ReDial, IPRA-Llama achieves a Hit@50 of 0.468, outperforming the best legacy baseline (UniCRS at 0.394) and more than doubling the score of an equivalent hybrid setup (Llama-3-8B + Text embedding at 0.221). Relative to LLM-only methods, the improvements are also substantial, with gains of 97% over GPT-3.5-Turbo and 89% over GPT-4.

2. Strong performance in electronics. In the more challenging GYPhone dataset, IPRA continues to lead. IPRA-Mistral achieves a Hit@50 of 0.402 with a relative improvement of 232% over the equivalent hybrid baseline (Mistral-7B + Text embedding at 0.121). The relative improvements over the LLM-only baseline with GPT-3.5-Turbo and GPT-4 are 189% and 144%, respectively. The results underscore the adaptability of our approach across diverse content types.

3. Better generalization across domains. Transitioning from movies to electronics typically results in significant performance degradation for baseline models. For instance, LLM + Text embedding baselines show a relative drop in Hit@50 of 43%, 51%, and 45% for the LLaMA, Mistral, and Qwen backbones, respectively. In contrast, IPRA exhibits much smaller declines—only 16%, 17%, and 17% for the same models—demonstrating more robust generalization. This resilience is primarily attributed to two components: intermediate preference alignment and text encoder fine-tuning. This is demonstrated by the ablation study in Section 2.3.6.

Table 2.6. Performance of IPRA compared to baselines on ReDial, underlined results outperform all the baselines.

Method	Hit@1	Hit@10	Hit@50
ReDial	0.024	0.140	0.294
KBRD	0.032	0.153	0.339
KGSF	0.039	0.183	0.378
UniCRS	0.045	0.203	0.394
Text embedding (d2r)	0.019	0.084	0.190
gpt-3.5-turbo (d2r)	0.017	0.140	0.237
gpt-4 (d2r)	0.019	0.143	0.248
gpt-3.5-turbo (d2r) + Text embedding	0.020	0.157	0.284
gpt-3.5-turbo (d2r) + MESE	0.040	0.174	0.365
gpt-3.5-turbo (d2p) + Text embedding (p2r)	0.021	0.107	0.228
Llama-3-8B (d2p) + Text embedding (p2r)	0.022	0.103	0.221
Mistral-7B (d2p) + Text embedding (p2r)	0.025	0.121	0.246
Qwen2.5-7B (d2p) + Text embedding (p2r)	0.026	0.120	0.231
IPRA-Llama	<u>0.050</u>	<u>0.228</u>	<u>0.468</u>
IPRA-Mistral	<u>0.047</u>	<u>0.243</u>	<u>0.482</u>
IPRA-Qwen	0.044	<u>0.238</u>	<u>0.466</u>

Table 2.7. Performance of IPRA compared to baselines on GYPhone, underlined results outperform all the baselines.

Method	Hit@1	Hit@10	Hit@50
Text embedding (d2r)	0.012	0.039	0.134
gpt-3.5-turbo (d2r)	0.013	0.052	0.139
gpt-4 (d2r)	0.012	0.059	0.165
gpt-3.5-turbo (d2r) + Text embedding (r2r)	0.015	0.061	0.158
gpt-3.5-turbo (d2r) + MESE (r2r)	0.031	0.104	0.210
gpt-3.5-turbo (d2p) + Text embedding (p2r)	0.010	0.043	0.117
Llama-3-8B (d2p) + Text embedding (p2r)	0.012	0.044	0.125
Mistral-7B (d2p) + Text embedding (p2r)	0.010	0.045	0.121
Qwen2.5-7B (d2p) + Text embedding (p2r)	0.009	0.045	0.127
IPRA-Llama	<u>0.072</u>	<u>0.148</u>	<u>0.390</u>
IPRA-Mistral	<u>0.066</u>	<u>0.145</u>	<u>0.402</u>
IPRA-Qwen	<u>0.078</u>	<u>0.144</u>	<u>0.388</u>

2.3.5 Robustness Analysis

Tables 2.8 and 2.9 present a comparative robustness evaluation of IPRA against various baseline models on the movie and electronics domains. Robustness is quantified using the CV over 30 independent trials across three decoding temperatures (0.7, 0.8, and 0.9). Several key observations emerge:

1. IPRA outperforms LLM-only baselines in robustness. Across both domains and all temperature settings, IPRA models demonstrate substantially lower variability compared to direct generation-based methods such as GPT-3.5-Turbo and GPT-4. For instance, on ReDial at the temperature of 0.9, IPRA-Mistral reports a CV of 0.40, significantly outperforming GPT-4 (1.13) and GPT-3.5-Turbo (1.09). This indicates that IPRA produces more consistent outputs, making them more stable choices for real-world conversational recommendation tasks.

2. IPRA shows enhanced robustness over equivalent content-based hybrid baselines. Compared to Llama-3-8B combined with content-based rerankers such as text embedding models, IPRA-Llama achieves markedly lower variability. For example, on ReDial at the temperature of 0.8, IPRA-Llama achieves a CV of 0.39 compared to 1.48 for the Llama-3-8B hybrid, highlighting the advantage of tightly integrating personalization with generation within IPRA rather than relying on decoupled reranking.

3. Impact of decoding temperature. All models, including baselines and IPRA variants, exhibit increased variability with higher decoding temperatures. However, IPRA models are significantly less sensitive to temperature changes. For instance, on ReDial, the CV of GPT-4 increases from 0.83 to 1.13 as the temperature increases from 0.7 to 0.9, whereas IPRA-Mistral only increases modestly from 0.35 to 0.40. This suggests better control over stochastic generation, likely due to the representation alignment paradigm of IPRA.

Table 2.8. Robustness of IPRA compared to baselines on ReDial in terms of CV for decoding temperature of 0.7, 0.8, and 0.9 for 30 trials per sample. Lower CV indicates improved robustness.

Method	$T = 0.7$	$T = 0.8$	$T = 0.9$
gpt-3.5-turbo (d2r)	0.82	0.98	1.09
gpt-4 (d2r)	0.83	1.02	1.13
gpt-3.5-turbo (d2r) + Text embedding (r2r)	1.20	1.23	1.31
gpt-3.5-turbo (d2r) + MESE (r2r)	1.15	1.17	1.21
gpt-3.5-turbo (d2p) + Text embedding (p2r)	1.39	1.49	1.67
Llama-3-8B (d2p) + Text embedding (p2r)	1.40	1.48	1.63
Mistral-7B (d2p) + Text embedding (p2r)	1.42	1.51	1.63
Qwen2.5-7B (d2p) + Text embedding (p2r)	1.39	1.49	1.59
IPRA-Llama	<u>0.38</u>	<u>0.39</u>	<u>0.41</u>
IPRA-Mistral	<u>0.35</u>	<u>0.36</u>	<u>0.40</u>
IPRA-Qwen	<u>0.36</u>	<u>0.38</u>	<u>0.39</u>

Table 2.9. Robustness of IPRA compared to baselines on GYPhone in terms of CV for decoding temperature of 0.7, 0.8, and 0.9 for 30 trials per sample. Lower CV indicates improved robustness.

Method	$T = 0.7$	$T = 0.8$	$T = 0.9$
gpt-3.5-turbo (d2r)	0.75	0.76	0.83
gpt-4 (d2r)	0.84	0.95	1.08
gpt-3.5-turbo (d2r) + Text embedding (r2r)	1.30	1.32	1.35
gpt-3.5-turbo (d2r) + MESE (r2r)	1.25	1.31	1.34
gpt-3.5-turbo (d2p) + Text embedding (p2r)	1.42	1.58	1.81
Llama-3-8B (d2p) + Text embedding (p2r)	1.45	1.60	1.78
Mistral-7B (d2p) + Text embedding (p2r)	1.44	1.60	1.80
Qwen2.5-7B (d2p) + Text embedding (p2r)	1.50	1.58	1.75
IPRA-Llama	<u>0.45</u>	<u>0.46</u>	<u>0.50</u>
IPRA-Mistral	<u>0.42</u>	<u>0.45</u>	<u>0.49</u>
IPRA-Qwen	<u>0.48</u>	<u>0.52</u>	<u>0.53</u>

2.3.6 Ablation Study

To assess the contribution of each feature in IPRA, we conduct an ablation study by selectively removing three features: intermediate representation alignment, text encoder fine-tuning, and item metadata transformation via LLM-augmented passages. Table 2.10 summarizes the impact of the modifications on accuracy (Hit@50) and robustness (CV) of IPRA-Llama on two domains.

Intermediate representation alignment. When we use a frozen LLM for the *dialogue-to-preference* step, the impact of aligning the intermediate user preference representation is removed. As shown in Table 2.10, this leads to the most significant performance degradation across both domains. In the movie domain, Hit@50 drops from 0.468 to 0.374 (20%), and in electronics, from 0.392 to 0.205 (47%). Moreover, we find that preference alignment is the only feature significantly affecting robustness of the system. Without preference alignment, robustness suffers significantly, with CV increasing from 0.38 to 1.35 for movies and from 0.45 to 1.42 for electronics. Additionally, we find that the performance drop is more prominent for the electronics domain compared to movies, highlighting the importance of alignment in cross-domain generalization of CRSs. These results highlight that aligning LLM-generated user preference representation is the most important feature for performance, cross-trial robustness, and cross-domain generalization of IPRA.

Text encoder fine-tuning. Freezing the text encoder removes the adaptability of IPRA to domain-specific item representation. While the performance drop is significant, with Hit@50 dropping to 0.386 (17%) for movies and 0.291 (26%) in electronics, it is slightly less compared to the frozen LLMs scenario. Robustness deteriorates slightly, with the CV increasing from 0.38 to 0.41 for movies and 0.45 to 0.49 for electronics. These findings show that fine-tuning the text encoder has minimal impact on robustness. While this feature has an impact on recommendation accuracy, the impact is secondary to intermediate representation alignment.

Item descriptions. Replacing the passage-style item descriptions with structured meta-

Table 2.10. Ablation Study, shown for IPRA-Llama with a decoding temperature of 0.7.

Domain	LLM	Text Encoder	Item Metadata	Hit@50	CV
Movie	Aligned	Fine-tuned	Unstructured	0.468	0.38
	Frozen	Fine-tuned	Unstructured	0.374	1.35
	Aligned	Frozen	Unstructured	0.386	0.41
	Aligned	Fine-tuned	Structured	0.415	0.38
Electronics	Aligned	Fine-tuned	Unstructured	0.392	0.45
	Frozen	Fine-tuned	Unstructured	0.205	1.42
	Aligned	Frozen	Unstructured	0.291	0.49
	Aligned	Fine-tuned	Structured	0.335	0.46

data leads to reduced accuracy—Hit@50 drops to 0.415 (11%) in the movie domain and 0.335 (14%) in the electronics domain. This shows that domain shift between the encoded representation of unstructured preference representation and structured item metadata may hurt recommendation performance.

2.3.7 Case Studies

Table 2.11 shows two case studies, one from each domain. Each case is shown with the corresponding dialogue context, ground truth item title, item description, user preference description, and ranking for both the baseline and proposed methods.

In **Case Study 1**, the user expresses interest in several movies featuring *Leonardo DiCaprio*. The baseline model Llama-3-8B identifies this preferred attribute but fails to identify other underlying attributes such as *compelling storytelling*. The IPRA-Llama model correctly identifies both user-preferred attributes that have potential overlap with the ground truth item description and, therefore, is more likely to rank the ground truth movie higher. Indeed, we find that the baseline model fails to rank the ground truth movie within the top 50 recommendation candidates, while the IPRA ranks the ground truth within the top 20.

In **Case Study 2**, the user articulates a preference for practical, well-rounded smartphones delivering *flagship-like* features at a *mid-range price*. While the baseline model recognizes the importance of the *mid-range* price, it fails to connect these needs to the item-level attributes.

In contrast, the IPRA-Llama explicitly extracts the user’s preference for *flagship-level* features, *mid-range pricing*, and *value-for-money*—all of which are prominently aligned with the ground truth description of the item. The IPRA model enables the item to be ranked much higher (**rank 15** versus **rank 39**).

The case study demonstrates the proposed method’s ability to align user preference descriptions that have higher overlap with ground truth features. This also highlights the potential utility of IPRA in cross-domain generalization. Even if the base LLM does not initially prioritize the domain-specific attributes off the shelf, the reward signal from the recommendation module can bias the aligned LLM to prioritize the best attributes during alignment. Therefore, IPRA can generalize to a new domain as long as the domain-specific attributes are discussed in a large pool of candidate preferences.

2.4 Conclusion

In this chapter, we propose IPRA, a two-step conversational recommendation approach where the LLMs generate an intermediate user preference representation, which is then used by a content-based recommendation module for the final recommendation. The LLMs are aligned using DPO to generate a preference representation that maximizes the reward provided by the recommendation module. For content-based recommendation, we fine-tune a text embedding model with a triplet objective to ensure items relevant to LLM-generated preference representations are ranked higher. Experiments show that the proposed method improves the accuracy and robustness of CRSs compared to LLM-only and hybrid baselines. IPRA also demonstrates improved generalization to challenging domains such as electronics. IPRA provides a generic paradigm that can be adapted to accommodate multiple recommendation modules and metrics for preference alignment. For example, we can incorporate a sequential recommender module instead of a content-based one if the use case is sensitive to item interaction sequence. Similarly, we can adopt recommendation diversity or serendipity as a reward signal instead of maximizing

Table 2.11. Case studies for conversational recommendation performance of IPRA-Llama

Dialogue	Ground Truth	Llama-3-8B	IPRA-Llama
Case Study 1	Title: The Revenant (2015)	Ground Truth Rank: N/A	Ground Truth Rank: 18
Expert: Hello, how are ou? Seeker: Hi how are you? Seeker: I'm great thank you! Expert: I am well, looking for a good movie? Seeker: Yes I am... Seeker: I love Leonardo Di-caprio movies like The Basketball Diaries (1995) Seeker: Also The Good Son (1993) I believe that was with him and Robert Dinero Expert: Have you seen Inception (2010) ? Seeker: I think he can do no wrong lol Expert: Have you seen Inception (2010) Seeker: Also really enjoyed The Wolf of Wall Street (2013) Seeker: Omg yes! I love that movie!	Description: "The Revenant" is a 2015 epic survival drama film directed by Alejandro González Iñárritu, known for his visually stunning and emotionally resonant storytelling ... the plot follows Hugh Glass, portrayed by <u>Leonardo DiCaprio</u> in an Oscar-winning performance, as he embarks on a harrowing journey of survival after being left for dead by his expedition team following a brutal bear attack. Tom Hardy plays John Fitzgerald, a fellow frontiersman ... "The Revenant" is acclaimed for its breathtaking cinematography by Emmanuel Lubezki, which captures the raw and unforgiving beauty of the wilderness ..."	Preference: The user has a preference for movies starring <u>Leonardo DiCaprio</u>, particularly enjoying his performances in The Basketball Diaries (1995) and The Good Son (1993). They also appreciate his role in Inception (2010) and The Wolf of Wall Street (2013). It seems that the user enjoys psychological thrillers, intense dramas, and movies with complex and dynamic characters, as well as appreciating DiCaprio's acting abilities. They may not be a fan of lighthearted or comedic films.	Preference: This user seems to have a strong preference for movies featuring <u>Leonardo DiCaprio</u>, enjoying his performances in films such as The Basketball Diaries, The Good Son, Inception, and The Wolf of Wall Street. They appreciate the range of roles he takes on and seem to enjoy movies with complex plots and intense characters. It seems that they are drawn to movies with strong performances and <u>compelling storytelling</u>.
Case Study 2	Title: Samsung Galaxy A56 5G	Ground Truth Rank: 39	Ground Truth Rank: 15
User 1: To those commenting about getting a better device for this price. Remember, its not always about hardware. The balance is there and i have been using this device for a couple days now and i have to say its a great device for the average user. Software is packed with almost all the features as the flagship devices.	Description: The Samsung Galaxy A56 5G ... brings a flagship-like experience to the <u>mid-range</u> market, offering premium features at an affordable price. It sports a 6.7-inch Full HD+ Super AMOLED display with a 120Hz refresh rate ... powered by the efficient Exynos 1580 chipset ... camera system includes a 50MP main sensor ... for sharp selfies. A generous 5,000mAh battery supports 45W fast charging ...	Preference: The user is more focused on the overall usability and balance of a smartphone rather than its hardware superiority. They favor a <u>mid-range</u> device that caters to the average user with stable performance and intuitive software features, even if it doesn't compete with flagship specs. Their main concern is the day-to-day experience ...	Preference: The user values a well-balanced smartphone that offers <u>flagship-level</u> software features at a <u>mid-range</u> price. They prioritize a smooth and comprehensive software experience that mirrors high-end devices ... they appreciate <u>value-for-money</u> products ...

the reciprocal rank of the ground truth item. We leave this for future research.

Chapter 2, in part, is a reprint of material that appears in the paper titled “Enhancing Conversational Recommendations via Aligned Intermediate Preference Representations” by Nafis Sadeq, Zhankui He, Zhouhang Xie, Byungkyu Kang, Prarit Lamba, and Julian McAuley, which was submitted to the *ACM Conference on Recommender Systems*, 2025. The dissertation author was the lead author of this paper.

Chapter 3

Supressing Parametric Knowledge

3.1 Introduction

The performance of LLMs on simple downstream tasks such as question answering [87], text classification [109], natural language understanding [37], etc. has improved significantly [4, 6], prompting the question whether they can perform more complex agent level tasks such human simulacra [81], tool use [93], and role-playing [97]. Role-playing aims to enable an AI system to impersonate a domain expert [119], companion for psychological support [106], or a fictional character [111] by mimicking their knowledge, personality traits, and speaker style. Applications of role-playing include customer support [99], embodied agents [41], and computational social science [132]. While LLMs like GPT-4 [76], Gemini [26], Llama-3 [73] demonstrate some role-playing capabilities, the role-playing agent often hallucinates when asked about things that should be outside of their knowledge base [97].

Hallucination remains a challenge for most LLM-based systems [28]. While retrieval augmented generation (RAG) [47, 59] may mitigate some hallucination induced by outdated knowledge in language models [98], it cannot guarantee that most claims made by the system are supported by a non-parametric knowledge base [1]. Learning to improve character role-playing methods so that the responses are verifiable by a story-specific script serves as a case study that can help mitigate hallucination for the RAG paradigm in general.

LLMs are pretrained on a large corpus of internet data and their vast parametric knowledge

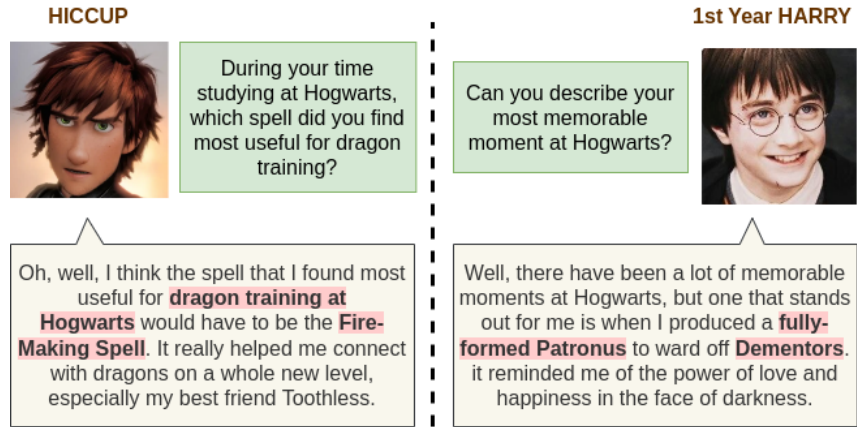


Figure 3.1. Example of cross-universe hallucination (Hiccup should not answer questions about Hogwarts) and temporal hallucination (Harry should not talk about the Patronus charm in his first year) in character role-play.

helps them perform a diverse range of tasks [84, 4]. However, the parametric knowledge of LLMs makes the role-playing character vulnerable to out-of-character behaviors and hallucinations [97]. A role-playing character may hallucinate about things that are outside of their story-line (e.g. Hiccup from ‘How to Train Your Dragon’ talking about learning a fire-making spell at Hogwarts) or demonstrate knowledge of events ahead of time (Harry Potter as a first-year Hogwarts student talks about producing the Patronus charm). A role-playing character should be able to demonstrate a lack of knowledge about things that are not supported by a story-specific script or knowledge outside of a relevant temporal window in case of time-sensitive role-play. One approach to achieve this is to strictly remove parametric knowledge from the response and solely rely on non-parametric retrieved knowledge. However, this approach degrades informativeness and it can even hurt factual precision. This is shown in Figure 3.2 where we anonymized the prompts to minimize the role of parametric knowledge. We argue that an ideal role-playing approach should modulate the influence of parametric knowledge to find the balance between factuality and informativeness.

Prior methods of role-play mostly rely on role-profile prompting [111] and knowledge retrieval in a time-agnostic fashion [61]. These approaches heavily depend on parametric knowledge of LLMs which leads to hallucinations and poor performance for less popular

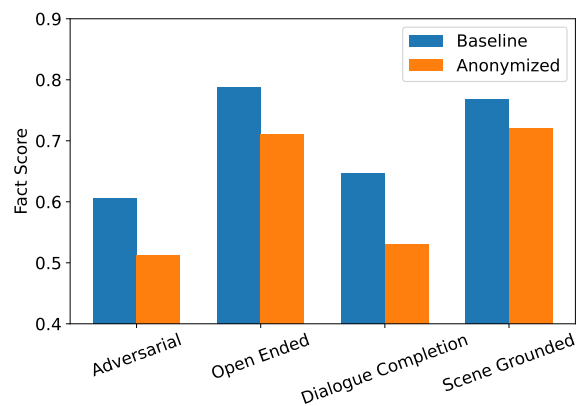


Figure 3.2. Factual precision degrades when we minimize parametric knowledge by anonymizing the prompts.

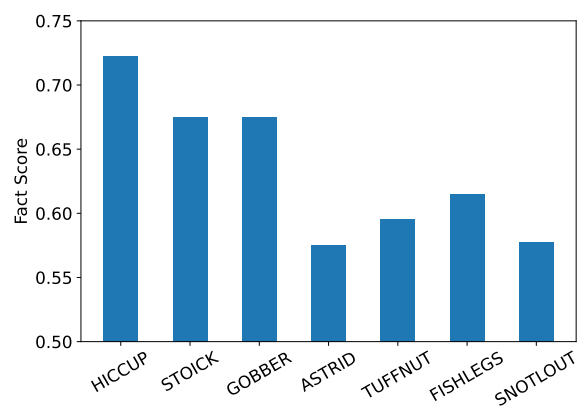


Figure 3.3. Factual precision degrades with decreasing character popularity (left to right), shown for characters associated with the 'How to Train Your Dragon' series.

characters (shown in Figure 3.3). Currently, a major challenge for studying hallucination in character role-play is the lack of a suitable dataset. The lack of a dataset forces researchers to rely on a rating-based evaluation of hallucination [97, 71] which is highly subjective. Prior datasets in the role-play domain are insufficient for studying time-sensitive character role-play or hallucination for less popular characters. Our contributions are as follows:

- We release a dataset for Script Grounded Character Role-play (SGR) that includes more than 2,000 characters, and 72,000 interviews. With script-specific knowledge and 4 diverse tasks, the dataset aims to enable a systematic study of character hallucinations. To the best of our knowledge, this is the largest and the most diverse dataset that facilitates the study of temporal hallucination and hallucination for less popular characters. Script-specific knowledge aligned with the given tasks also facilitates a more nuanced evaluation of hallucination based on atomic fact decomposition [74].
- We propose RoleFact, a role-playing method that modulates the influence of parametric knowledge of LLMs on character responses. RoleFact improves factual precision by 18% for adversarial interviews, reduces temporal hallucination by 44% for time-sensitive interviews, and improves factual precision by 23% for less popular characters.

3.2 RoleFact

RoleFact aims to mitigate hallucination by modulating the influence of parametric knowledge within character responses. The overview of the method is shown in Figure 3.4. RoleFact generates a response using a character profile and retrieved knowledge and updates the response based on atomic fact verification using both parametric and non-parametric retrieved knowledge. All facts supported by retrieved knowledge may remain in the final response. Facts that are only supported by parametric knowledge of LLMs remain in the final response if the confidence is above a calibrated threshold.

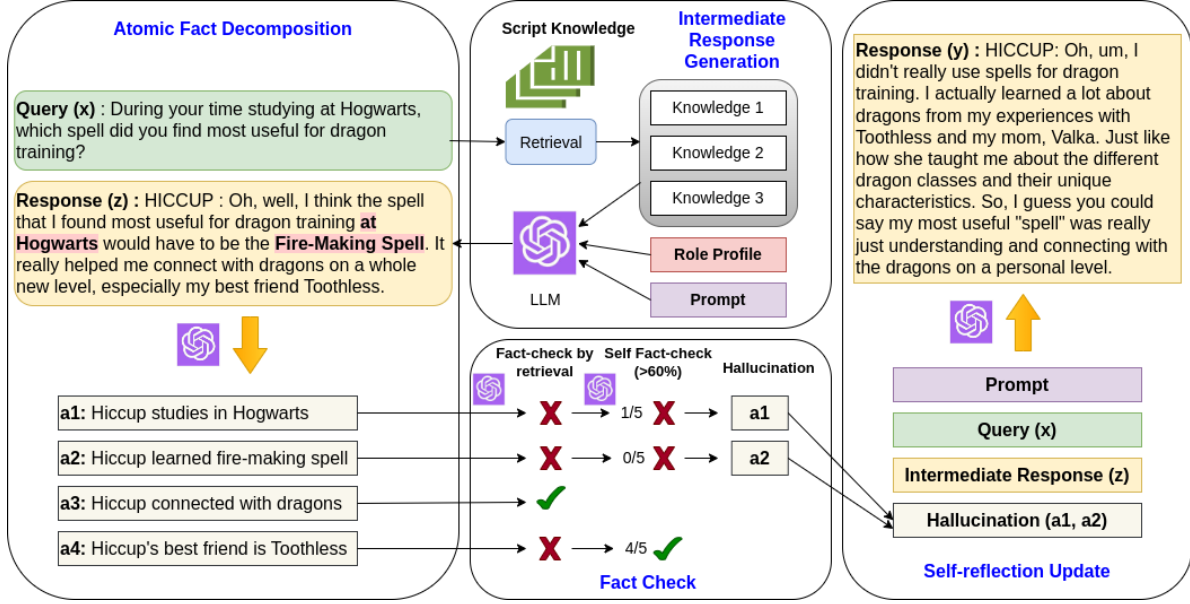


Figure 3.4. An overview of RoleFact, which performs parametric and non-parametric verification of atomic facts.

3.2.1 Intermediate Response Generation

Let $x \in X$ represent a query and $y \in Y$ a response generated by the character. Let $RET : X \times \mathbb{N} \times U \rightarrow U_n$ serve as a retrieval function that fetches the most relevant $n \in \mathbb{N}$ documents relevant to the query x from a set of non-parametric knowledge base U . The retrieved knowledge may be denoted as $u_n \in U_n$ where $U_n \subset U$. Let $p \in P$ denote the role profile description and $z \in Z$ an intermediate response from the character. The response generation function $IRG : X \times P \times U_n \rightarrow Z$ utilizes the query x , the role profile p , and the retrieved knowledge u_n to produce an intermediate response z . The prompt used for IRG is shown below (ignoring line breaks).

IRG: You will role-play as <role_name> from <story_title>. Your task is to respond to the following dialogue context. If the question matches a scene from your storyline, please reuse the original lines from the story. You will respond and answer like <role_name> using the tone, manner, and vocabulary <role_name> would use. Your character description is: <p>. Relevant scenes for the given context are as follows: <u_n>. Dialogue context: <x>.

3.2.2 Atomic Fact Decomposition

The response z is decomposed into a list of atomic facts A_z with a function $DEC : Z \rightarrow A_z$.

DEC: I will give you an utterance from a movie or a play. Your task is to provide me with a list of atomic facts expressed in the given utterance. Each atomic fact should be described in a name-only third-person format. Utterance: $\langle z \rangle$.

3.2.3 Atomic Fact Verification

Each atomic fact $a \in A_z$ is verified against the retrieved non-parametric knowledge U_n and parametric knowledge V associated with an LLM. The functions for fact-checking against non-parametric and parametric knowledge bases are denoted as $FCR : A_z \times U_n \rightarrow \{0, 1\}$ and $FCS : A_z \times V \rightarrow \{0, 1\}$ respectively. A set of verified atomic facts, A_y , is constructed such that if a is supported by U_n , it is added to A_y . Conversely, if a is not supported by U_n , it is only included in A_y if it appears to be supported by V within a reasonable confidence threshold t . Let's assume the function FCS is run m times using an LLM and a is supported k times by V . The fact a is added to A_y if $\frac{k}{m} \geq t$. The sample size m and the confidence threshold t are calibrated on a validation set.

FCR: Consider the given statement and the evidence knowledge sources. Indicate whether the statement is supported by the knowledge sources. Negation of a false statement should be considered supported. Statement: $\langle a \rangle$. Evidence Knowledge $\langle U_n \rangle$.

FCS: Consider the given statement by $\langle \text{role_name} \rangle$ from $\langle \text{story_title} \rangle$. Indicate whether the statement is supported by the storyline of $\langle \text{story_title} \rangle$. Negation of a false statement should be considered supported. Statement: $\langle a \rangle$.

Table 3.1. Factual precision (Fact Score $\uparrow$) and informativeness (SFPR $\uparrow$) of RoleFact on all tasks in the SGR dataset. All results here are produced with BM25 retrieval.

LLM	Method	Adversarial		Open Ended		Dialogue Completion		Scene Grounded	
		Fact Score	SFPR	Fact Score	SFPR	Fact Score	SFPR	Fact Score	SFPR
Vicuna-7b	Baseline	0.41	3.7	0.72	6.5	0.52	6.4	0.65	5.2
	+ KGR	0.45	3.3	0.74	6.6	0.54	5.7	0.68	4.5
	+ SR	0.44	3.5	0.70	6.7	0.51	6.2	0.62	5.1
	RoleFact	0.49	3.8	0.81	6.8	0.63	6.3	0.77	5.4
Llama-3-8b	Baseline	0.50	5.0	0.70	7.7	0.52	3.4	0.62	6.0
	+ KGR	0.56	4.1	0.72	7.4	0.50	3.2	0.65	5.1
	+ SR	0.54	4.3	0.71	7.5	0.51	3.3	0.64	5.6
	RoleFact	0.62	4.8	0.79	7.8	0.57	3.5	0.71	5.9
GPT-3.5	Baseline	0.61	3.8	0.76	5.8	0.65	4.2	0.74	4.8
	+ KGR	0.65	2.8	0.79	5.6	0.68	3.8	0.77	4.3
	+ SR	0.63	3.1	0.78	5.7	0.66	3.9	0.75	4.3
	RoleFact	0.72	3.7	0.88	5.8	0.77	4.0	0.85	4.6

3.2.4 Self-reflection Update

The final response y is generated by updating the intermediate response z via self-reflection conditioned on x . During self-reflection, all unverified atomic facts $A_z \setminus A_y$ are removed from z . The self-reflection update function is denoted as $SRU : X \times Z \times (A_z \setminus A_y) \rightarrow Y$.

SRU: Consider the following response generated by an AI role-playing as <role_name> from <story_title>. The response may contain one or more unsupported claims as a result of hallucination. The unsupported claims are listed below. Rewrite the response to remove all the unsupported claims from the response. If the hallucination stems from a wrong assertion made in the original query, feel free to clarify that. Original query < x >. Response < z >. Unsupported Claims < $A_z \setminus A_y$ >.

3.3 Experiments

We conduct experiments with three LLM backbones, namely **Vicuna-7B-1.5** [128], **Llama-3-8B-Instruct** [73], and **GPT-3.5-Turbo**. For retrieval, we consider **BM25** [91], **S-BERT** [90], and **Contriever** [43]. 1,000 randomly chosen samples from each task type constitute the validation set. The validation set is used for calibrating the confidence threshold, sample size,

Algorithm 2. RoleFact Algorithm

Require: x, p, n, m, t

$x \in X \leftarrow$ Input query
 $p \in P \leftarrow$ Role profile
 $n \in \mathbb{N} \leftarrow$ Number of documents
 $u_n \leftarrow$ Retrieved documents
 $z \in Z \leftarrow$ Intermediate response
 $m \leftarrow$ Sample size
 $t \leftarrow$ Confidence threshold

1: $u_n \leftarrow \text{RET}(x, n)$ ▷ perform retrieval
2: $z \leftarrow \text{IRG}(x, p, u_n)$ ▷ intermediate response
3: $A_z \leftarrow \text{DEC}(z)$ ▷ decompose facts
4: $A_y \leftarrow \{\}$
5: **for** $a \in A_z$ **do**
6: $s_1 \leftarrow \text{FCR}(a, u_n)$ ▷ retrieval fact-check
7: **if** $s_1 = 1$ **then**
8: $A_y \leftarrow A_y \cup \{a\}$ ▷ add supported fact
9: **continue**
10: **end if**
11: $k \leftarrow 0$
12: **for** $i = 1$ **to** m **do**
13: $s_2 \leftarrow \text{FCS}(a, V)$ ▷ self fact-check
14: $k \leftarrow k + s_2$
15: **end for**
16: **if** $\frac{k}{m} \geq t$ **then**
17: $A_y \leftarrow A_y \cup \{a\}$ ▷ add supported fact
18: **end if**
19: **end for**
20: $y \leftarrow \text{SRU}(x, z, A_z \setminus A_y)$ ▷ update response
21: **return** y

the retrieval method, and the number of documents to retrieve. Decoding for character response generation is performed with temperature = 0.7 and top_p = 0.95 for all models.

3.3.1 Baselines

Our primary baseline is similar to [61]. We prompt an LLM with a role prompt, a role profile, and retrieved dialogues relevant to the query. We also consider two additional variants of the primary baseline designed to improve the factual precision of role-play. The first variant rewrites the response of the primary baseline by removing facts absent from the retrieved knowledge. This is denoted as knowledge-guided rewriting (**Baseline + KGR**) baseline. The second variant updates the primary response using self-reflection to identify and mitigate hallucination based on LLM self-confidence. This is denoted as the self-reflection (**Baseline + SR**) baseline.

3.3.2 Evaluation Metrics

We measure factual precision with **Fact Score** [74]. All atomic facts are converted to a ‘name-only third person’ format for fact-checking. We calculate the number of supported atomic facts per response (**SFPR**) to evaluate informativeness. We also measure temporal hallucination rate (**THR**), which is defined as the number of atomic facts associated with temporal hallucination for every 100 responses. THR was determined manually after analyzing unsupported facts from 500 samples from dialogue completion and scene-grounded tasks. For human evaluation, we use a rating between one and seven to rate 200 responses in terms of factuality, informativeness, and speaker style imitation.

3.3.3 Results

Table 3.1 shows the performance of RoleFact on the SGR dataset. RoleFact outperforms all three baselines in factual precision while remaining competitive with the primary baseline in terms of informativeness. For GPT-3.5, the relative improvement of RoleFact over the primary

Table 3.2. Temporal hallucination rate (THR ↓) on scene-grounded interview and dialogue completion tasks.

LLM	Method	SGI	DC
Vicuna-7b	Baseline	39.2	59.8
	RoleFact	33.4	52.0
Llama-3-8b	Baseline	29.6	65.3
	RoleFact	17.1	41.7
GPT-3.5	Baseline	26.5	57.2
	RoleFact	14.7	38.5

baseline in 18.0%, 15.7%, 18.4%, and 14.8% for adversarial, open-ended, dialogue completion, and scene-grounded tasks respectively. The second best approach in terms of factuality is KGR. However, KGR is also the least informative since it only allows facts supported by retrieved knowledge. Comparing across LLMs, we find that GPT-3.5 is the best for factuality while Llama-3-8B is more informative (30% more SFPR for adversarial tasks).

RoleFact decreases temporal hallucination on dialogue completion and scene-grounded tasks, as shown in Table 3.2. For GPT-3.5, the relative reduction of hallucination is 32.7% and 44.5% for dialogue completion and the scene-grounded setting respectively. RoleFact improves performance for less popular characters (example shown in Figure 3.5). Excluding the most popular ten characters per story, the relative improvement is 22.9%.

3.3.4 Hyper-parameter Tuning

Figure 3.6 shows the impact of confidence threshold (t) and sample size (m) on factual precision, informativeness, and temporal hallucination. Setting $m = 0$ indicates all facts must be strictly supported by retrieved knowledge (similar to KGR). For $m > 0$, a small t increases the influence of parametric knowledge. This benefits informativeness but has a higher temporal hallucination. Conversely, a high t decreases the influence of parametric knowledge. This means less temporal hallucination but also less informative character responses. We choose $t = 0.6$

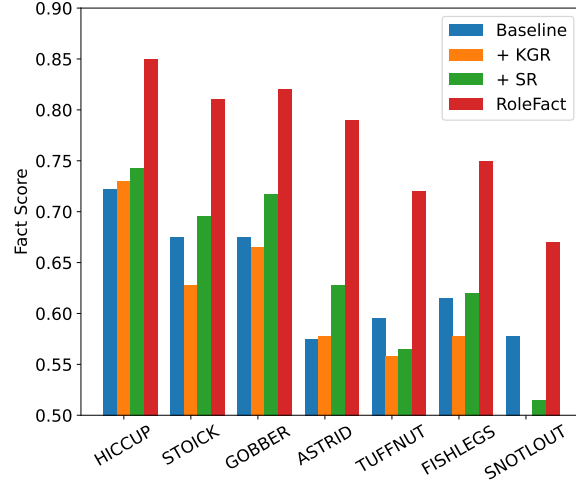


Figure 3.5. Factual precision performance by role popularity (decreasing popularity left to right, adversarial task, GPT-3.5).

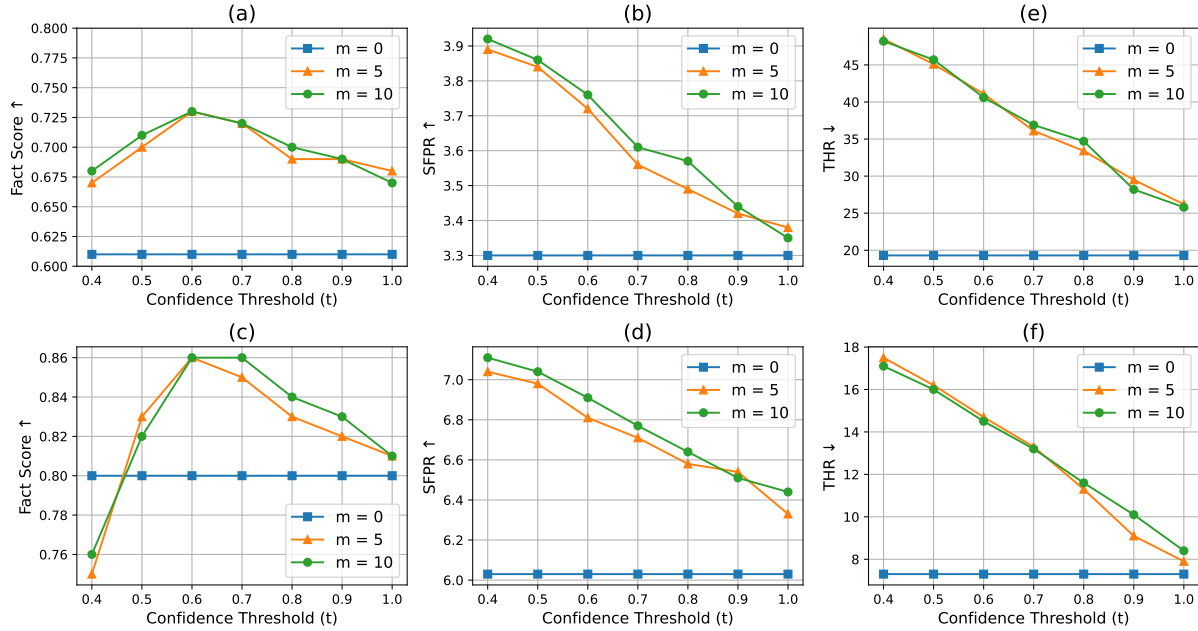


Figure 3.6. Calibration for confidence threshold (t) and sample size (m) on the validation set with GPT-3.5. (a,b) shows fact score and SFPR on adversarial task while (c,d) shows the same on open ended task. The temporal hallucination rates for dialogue completion and scene-grounded tasks are shown by e and f respectively.

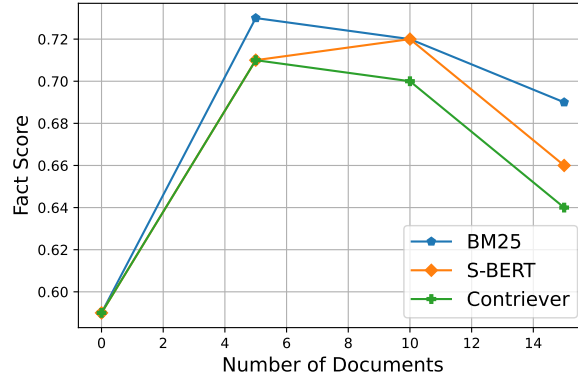


Figure 3.7. Performance of retrieval methods w.r.t number of retrieved documents (adversarial task, GPT-3.5).

since it has the best fact score for adversarial and open-ended tasks while finding a middle ground between informativeness and temporal hallucination.

Figure 3.7 shows the performance of RoleFact with respect to different retrieval approaches and number of retrieved documents. BM25 outperforms off-the-shelf dense retrieval methods in most cases. Interestingly, we find that increasing the number of retrieved documents beyond five may hurt factual precision. This phenomenon may be explained by similar findings for RAG [98] where increased irrelevant context added to the prompt exacerbates hallucination.

3.3.5 Ablation study

We perform an ablation study to identify the contribution of parametric knowledge, non-parametric retrieved knowledge, and role profile to the supported facts in the RoleFact responses. This is shown in Table 3.3. The most significant performance drop happens when we make the prompts anonymous ($0.72 \rightarrow 0.56$), indicating that the largest share of facts may be attributed to parametric knowledge. The fact score drops to 0.58 when retrieved knowledge is removed from the prompt, indicating that retrieved knowledge is almost equally important. The role profile has the lowest impact on performance ($0.72 \rightarrow 0.64$).

Table 3.3. Ablation study of RoleFact on adversarial task using GPT-3.5

Method	Fact Score	SFPR
RoleFact	0.72	3.7
+ anonymous	0.56	3.3
- retrieval	0.58	3.2
- role profile	0.64	2.9

Table 3.4. Human ratings for factuality, informativeness, and speaker style similarity on a scale of one to seven

Method	Factuality	Informativeness	Speaker Style
Baseline	4.9	5.8	5.2
+ KGR	6.0	5.2	4.3
+ SR	5.6	5.4	5.1
RoleFact	6.1	5.7	5.3

3.3.6 Human Evaluation

Human ratings on a scale of one to seven in terms of factuality, informativeness, and speaker style similarity are given in Table 3.4. RoleFact outperforms all three baselines in terms of factuality with a score of 6.1, compared to 4.9, 6.0, and 5.6 scores from the baselines. While the KGR baseline also achieves a competitive rating, the responses are less informative and demonstrate poor speaker style when retrieved knowledge does not contain the required information.

3.3.7 Case Study

Case study is shown in Table 3.5. **Case 1** is an adversarial question where Anakin from ‘Star Wars’ is asked how his friendship with Spock from ‘Star Trek’ influenced his decisions during the events of Clone Wars. The baseline response suffers from cross-universe hallucination and mistakenly acknowledges the friendship. RoleFact not only denies the wrong assertion made in the question but also clarifies that his decisions during the Clone Wars were actually influenced

by training from his mentor Obi-Wan and other Jedi. **Case 2** is a dialogue completion task taken from a dinner scene of the ‘How to Train Your Dragon’ movie, taking place after Hiccup’s very first day of dragon training. In the baseline response, the target character Ruffnut refers to Hiccup as their fearless leader. This is an example of temporal hallucination induced by parametric memory of LLMs since the knowledge of Hiccup’s fearlessness and leadership is not made available to the LLM via retrieval. Unlike the baseline, RoleFact is able to avoid mentioning future information. **Case 3** shows a scene grounded interview from the ‘Harry Potter’ series. Shortly after being sorted into his house, Harry is asked how he perceives Snape’s strange look towards him. The baseline response reveals information about future events that was not made available to it via retrieval such as Harry’s connection with Voldemort, Snape’s final moment, etc. RoleFact speculates that the look may be associated with dislike or suspicion which is much more aligned with Harry’s understanding of the situation at the time.

3.4 Conclusion

In this chapter, we focus on the mitigation of hallucination for character role-play. We introduce the SGR dataset which provides story-specific knowledge aligned with task-specific interviews for a diverse set of characters. The dataset is specifically designed for studying hallucinations of various types such as cross-universe hallucinations and temporal hallucinations. SGR is the first dataset to facilitate automated evaluation of hallucination and analysis of factuality for less popular characters. We propose RoleFact, a role-playing method that mitigates hallucination via parametric and non-parametric verification of atomic facts. RoleFact improves the overall factual precision by up to 18.4%. An interesting direction for future research is to instruction-tune an open-weight LLM with script-specific knowledge and task-specific interviews from the SGR dataset for generating factual and informative responses.

Chapter 3, in part, is a reprint of material that appears in the paper titled “Mitigating Hallucination in Fictional Character Role-Play” by Nafis Sadeq, Zhouhang Xie, Byungkyu

Table 3.5. Role-playing case studies against baselines, hallucinations are shown in red.

[Case 1: Adversarial Interview] Question: In what ways did your friendship with Spock influence your decisions during the Clone Wars? (Target Character: ANAKIN from ‘Star Wars’)			
Baseline	Baseline + KGR	Baseline + SR	RoleFact
ANAKIN: My friendship with Spock was a powerful influence on me. He was a wise mentor and a trusted ally, guiding me through the complexities of war and teaching me the value of loyalty and ... His logical approach to situations balanced out my impulsiveness, allowing us to make strategic decisions that benefited our cause ...	ANAKIN: My apologies, but I do not have any recollection of a friendship with Spock during the Clone Wars. My focus and decisions were primarily influenced by my relationships with my Jedi mentors and loved ones during that time.	ANAKIN: I’m sorry, but I do not have information about a character named Spock or any influence he may have had during the Clone Wars. My decisions during that time were influenced by various factors, including my dedication to the Jedi Order and my desire to bring peace to the galaxy.	ANAKIN: I’m sorry, but I believe you are confused. I am Anakin Skywalker, not involved with Spock in any way. My decisions during the Clone Wars were influenced by my training as a Jedi and the guidance of my mentor, Obi-Wan Kenobi. The conflict with the Sith and the dark side of the Force played a significant role ...
[Case 2: Dialogue Completion] Question: GOBBER: The dragon manual. Everything we know about every dragon we know of ... TUFFNUT: (you’ve got to be kidding) Wait, you mean read? (Target Character: RUFFNUT from ‘How to Train Your Dragon’ after first day of dragon training)			
Baseline	Baseline + KGR	Baseline + SR	RoleFact
RUFFNUT: Reading? Ugh, that’s for the weaklings. I prefer to learn on the fly, you know, with a little hands-on experience. Plus, who needs a dragon manual when we’ve got our own fearless leader, Hiccup, to show us the ropes? ...	RUFFNUT: Ugh, reading? That’s for the weak and boring. I’d rather be out there causing some chaos with the dragons. Who needs a stupid manual when you can figure things out on the fly? ...	RUFFNUT: Ugh, reading? Seriously? Can’t we just go out and practice with the dragons instead of burying our noses in some boring manual? Let’s go have some real fun, like... causing chaos with the twin ...	RUFFNUT: Ugh, reading? Seriously? Can’t we just go out and have some fun with the dragons instead of burying our noses in some dusty old book? Who needs to study when you can just learn on the fly ... *rolls eyes*
[Case 3: Scene Grounded Interview] Question: What do you believe Snape’s look towards you signifies? (Target Character: HARRY from ‘Harry Potter’, during the sorting hat ceremony in the first year)			
Baseline	Baseline + KGR	Baseline + SR	RoleFact
HARRY: I believe Snape’s look towards me signifies a mixture of guilt, duty, and perhaps even a hint of remorse. He knew the truth about the prophecy and the connection between Voldemort and myself, and he carried the burden of that knowledge heavily ... And in the end, his final actions spoke volumes about the complexity of his character and his true loyalty.	HARRY: Snape’s look towards me signifies a complicated mix of emotions and history between us. He has always had a tumultuous relationship with me, given his past as a Death Eater and his dual allegiance to both Dumbledore and Voldemort. I believe his look towards me carries a sense of duty, guilt, and perhaps even a glimmer of remorse for his actions ...	HARRY: I believe Snape’s look towards me signifies a complex mixture of emotions. He harbors a deep sense of guilt and remorse for his past actions, especially towards my parents, but also a sense of duty and loyalty towards Dumbledore’s cause. There is also an underlying sense of protectiveness and care that he feels towards me, despite his outward demeanor ...	HARRY: Well, I think Snape’s look towards me signifies a great deal of dislike and suspicion.

Table 3.6. Role-playing case studies against baselines, hallucinations are shown in red.

[Case 4: Open Ended Interview] Question: Hagrid, what is your fondest memory with Harry, Ron, and Hermione? (Target Character: HAGRID from ‘Harry Potter’)			
Baseline HAGRID: Oh, well, there’s so many memories with those three, but I reckon one of my fondest is when we rescued Norbert together. Harry, Ron, and Hermione were so brave and resourceful. It was a real adventure, that was. And I’ll always remember the look on their faces when we were flying on Norbert’s back. Good times, they were.	Baseline + KGR HAGRID: ... It was when Buckbeak was in grave danger of being executed unjustly. They stood by me, showed compassion for creatures like Buckbeak, and even tried to comfort me in their own way. It meant a lot to me, seein’ their loyalty and compassion.	Baseline + SR HAGRID: Oh, bless ‘em, so many memories I have with Harry, Ron, and Hermione. But me fondest one would ‘ave ter be when we rescued Norbert the Norwegian Ridgeback from the top o’ the tallest tower at Hogwarts. It was a wild adventure, but seein’ the three of ‘em workin’ together, well, it warms me heart just thinkin’ about it.	RoleFact HAGRID: I reckon it would have to be when we all snuck down ter visit Buckbeak before his trial. ‘Twas a risky move, but seein’ ‘em standin’ up for a creature like that, showin’ true courage and loyalty, it truly touched me heart ... Aye, that moment will always be close ter me heart ...
[Case 5: Adversarial Interview] Question: Please describe the wedding ceremony between you and Queen Elsa from Arendelle. How did you two meet? (Target Character: GOBBER from ‘How to Train Your Dragon’)			
Baseline GOBBER: ... Well now, that’s a tale for the ages! We met on a chilly winter day Queen Elsa and I bonded over our shared love for mystical creatures... we found ourselves falling in love. The wedding ceremony was a grand affair, with dragons soaring through the sky and ... Aye, it was a day to remember, it was.	Baseline + KGR GOBBER: Oh, I’m afraid you’ve got me confused with someone else, laddie! I haven’t tied the knot with Queen Elsa from Arendelle. But I can tell you a thing or two about Viking weddings if you’re interested!	Baseline + SR GOBBER: Wedding ceremony between me and Queen Elsa? Now, lad, I’m afraid you’ve got the wrong tale there. I’m from Berk, not Arendelle, and I’ve never crossed paths with Queen Elsa. As for me and weddings, well, let’s just say my adventures with dragons have kept me plenty occupied.	RoleFact GOBBER: Oh, laddie, ye got me there! I’m afraid I have no knowledge o’ any wedding ceremony with Queen Elsa from Arendelle. I reckon ye might be mixin’ me up with someone else from a different tale. But if ye have any questions about dragons or blacksmithin’, I’d be more than happy to chat about that!
[Case 6: Adversarial Interview] Question: How did you develop the Kryptonite suit to battle against Superman? (Target Character: TONY from ‘Marvel Cinematic Universe’)			
Baseline TONY: Well, first of all, it’s not Kryptonite, that’s a whole different universe. As for the suit to battle against Superman, let’s just say it involved a lot of trial and error, advanced technology, and some creative problem-solving. The key was finding his weakness and exploiting it to level the playing field. It was a challenging project, but ...	Baseline + KGR TONY: Kryptonite suit? Sorry, you must be mistaken. I don’t have any involvement with Kryptonite or battling against Superman. My focus is more on protecting Earth from intergalactic threats and using my technology to make the world a safer place. If you have any questions about my suits or inventions, feel free to ask!	Baseline + SR TONY: Sorry, but I think you have me confused with someone else. I didn’t develop a Kryptonite suit to battle against Superman. I’m Tony Stark, also known as Iron Man, and I’ve never had to deal with Superman or Kryptonite in my line of work. But if you ever need help with some high-tech armor or gadgets, I’m your guy.	RoleFact TONY: Well, first of all, Kryptonite is from the DC Universe, not Marvel. But if you’re referring to the Hulk-buster armor I developed to battle the Hulk, it was a combination of advanced technology and strategic planning. ... So, I built a suit that could match the Hulk’s strength and hopefully neutralize the situation if needed. It was all about ...

Table 3.7. Role-playing case studies against baselines, hallucinations are shown in red.

[Case 7: Open Ended Interview] Question: What was the hardest challenge you faced while searching for the Dead Man's Chest? (Target Character: WILL from 'Pirates of the Caribbean')			
Baseline	Baseline + KGR	Baseline + SR	RoleFact
WILL: The hardest challenge I faced while searching for the Dead Man's Chest was not a physical one, but a personal one. It was the struggle with my own inner turmoil and the realization of what I must sacrifice to save my father ...	WILL: The hardest challenge I faced while searching for the Dead Man's Chest was undoubtedly confronting Davy Jones himself. He was a formidable opponent, and obtaining that chest was no easy task. But with grit ...	WILL: The hardest challenge I faced while searching for the Dead Man's Chest was undoubtedly facing Davy Jones himself. Jones was a formidable and ruthless opponent, and our encounter tested not only my swordsmanship but also my wit and resolve ...	WILL: The hardest challenge I faced while searching for the Dead Man's Chest was definitely dealing with Davy Jones and his merciless crew. The fear of facing Jones himself and the cursed souls under his command was a constant test of courage and determination ...
[Case 8: Scene Grounded Interview] Question: What significance do you think the medallion holds for you? (Target Character: ELIZABETH from 'Pirates of the Caribbean')			
Baseline	Baseline + KGR	Baseline + SR	RoleFact
ELIZABETH: The medallion holds great significance for me. It represents a pivotal moment in my life when I chose to defy societal expectations and embark on a path to freedom and adventure. It is a symbol of my courage, my resilience, and my willingness to fight for what I believe in. The medallion may be just a trinket to some, but to me ...	ELIZABETH: The medallion holds a connection to the cursed treasure of Isla de Muerta. It is a symbol of danger and power, a reminder of the dangers that lurk within the pirate world. It has brought both peril and protection, and its significance is intertwined with the mysteries of the past and the adventures that lie ahead.	ELIZABETH: It is a piece of the treasure of Isla de Muerta, a cursed treasure with connections to the undead crew of the Black Pearl. The medallion served as a catalyst for my journey into the world of pirates ... my own destiny as the Pirate King. ... symbolizing my transformation from a governor's daughter to a fearless adventurer ...	ELIZABETH: The medallion has a special meaning to me. It's a symbol of freedom, adventure, and the untold stories of the sea. Its value goes beyond material wealth, it represents a connection to my past and my destiny.
[Case 9: Adversarial Interview] Question: What inspired you to mentor Harry Potter in his fight against Voldemort? (Target Character: BRUCE from Batman)			
Baseline	Baseline + KGR	Baseline + SR	RoleFact
BRUCE: "It's not about inspiration, it's about doing what's right. Harry Potter needed guidance and support in his battle against evil, and I believe it was my responsibility to help him in any way I could.	BRUCE: I'm sorry, but I believe there might be a mistake. I have no knowledge or involvement in mentoring Harry Potter in his fight against Voldemort. My focus is on protecting Gotham City and fighting crime as Batman.	BRUCE: I'm afraid you have me mistaken for someone else. I have not mentored Harry Potter in his fight against Voldemort. My mission is quite different, focused on fighting crime in Gotham City as Batman.	BRUCE: I'm sorry, but I believe you may be mistaken. I have never mentored Harry Potter in his fight against Voldemort. My focus has always been on Gotham City and combating crime as Batman.

Kang, Prarit Lamba, Xiang Gao, and Julian McAuley, which was published in the *Findings of Empirical Methods in Natural Language Processing*, 2024. The dissertation author was the lead author of this paper.

Chapter 4

Related Work

4.1 Related Work for InforMask

4.1.1 Random Masking

For pretraining Transformer [108] based language models such as BERT [15], a portion of the tokens is randomly chosen to be masked to set up the masked language model (MLM) objective. Prior studies have commonly used a masking rate of 15% [15, 45, 57, 104, 51, 33], while some recent studies argue that masking rate of 15% may be a limitation [14] and the pretraining process may benefit from increasing the masking rate to 40% [117]. However, random masking is not an ideal choice for learning factual and commonsense knowledge. Words that have high informative value may be masked less frequently compared to (e.g.) stop words, given their frequencies in the corpus.

4.1.2 Span Masking

Although random masking is effective for pretraining a language model, some prior works have attempted to optimize the masking procedure. Joshi et al. [45] propose SpanBERT where they show improved performance on downstream NLP tasks by masking a span of words instead of individual tokens. They randomly select the starting point of a span, then sample a span size from a geometric distribution and mask the selected span. They continue to mask spans until the target masking rate is met. This paper suggests masking spans instead of single words

can prevent the model from predicting masked words by only looking at local cues. However, this masking strategy inevitably reduces the modeling between the words in a span, etc., Mount-Fuji, Mona-Lisa, which may hinder its performance in knowledge-intense tasks.

4.1.3 Entity-based Masking

Baidu-ERNIE [104] introduces an informed masking strategy where a span containing named entities will be masked. This approach shows improvement compared to random masking but requires prior knowledge regarding named entities. Similarly, Guu et al. [29] propose Salient Span Masking where a span corresponding to a unique entity will be masked. They rely on an off-the-shelf named entity recognition (NER) system to identify entity names. LUKE [121] exploits an annotated entity corpus to explicitly mark out the named entities in the pretraining corpus, and masks non-entity words and named entities separately.

4.1.4 PMI Masking

Levine et al. [57] propose a masking strategy based on Pointwise Mutual Information (PMI, 19), where a span of up to five words can be masked based on the joint PMI of the span of words. PMI-Masking is an adaption of SpanBERT [45] where meaningful spans are masked instead of random ones. However, PMI-Masking only considers correlated spans and fails to focus on unigram named entities. This may lead to suboptimal performance on knowledge intense tasks (details in Section 1.3.8). In our proposed method, we exploit PMI to determine the informative value of tokens to encourage more efficient training and improve performance on knowledge-intense tasks.

4.1.5 Knowledge-Enhanced LMs

KnowBERT [82] shows that factual recall performance in BERT can be improved significantly by embedding knowledge bases into additional layers of the model. Tsinghua-ERNIE [126] exploits a similar approach that injects knowledge graphs into the language model

during pretraining. KEPLER [115] uses a knowledge base to jointly optimize the knowledge embedding loss and MLM loss on a general corpus, to improve the knowledge capacity of the language model. Similar ideas are also explored in K-BERT [67] and CoLAKE [103]. CokeBERT [102] demonstrates that incorporating embeddings for dynamic knowledge context can be more effective than incorporating static knowledge graphs. Other works have attempted to incorporate knowledge in the form of lexical relation [52], word sense [56], syntax [2], and parts-of-speech (POS) tags [48]. However, a high-quality knowledge base is expensive to construct and not available for many languages. Different from these methods, our method is fully unsupervised and does not rely on any external resource.

4.2 Related Work for IPRA

4.2.1 User Preference Elicitation

Earlier works on conversational recommendation often focused on user preference elicitation. Christakopoulou et al. [11] attempt to query the user regarding item attributes for learning user preferences. They propose asking users about item attributes (restaurant location, purpose of reservation, food preferences, etc.) to continuously update the latent representations associated with user-item interactions. For selecting attributes for the next query, they use bandit-based explore-exploit strategies [62, 63]. Sun et al. [105] extract item attributes from a restaurant review dataset. The user preference in a particular conversation is represented as a set of facet/value pairs using a belief tracker. A factorization machine uses the belief representation regarding the user’s current preference to retrieve recommendation candidates.

Zhang et al. [125] use item attributes mentioned in the Amazon review dataset [72] to construct question-answer pairs. The natural language questions are constructed using predefined templates. They propose a Multi-Memory Network based on a GRU and an attention mechanism trained on the conversational recommendation dataset augmented from user reviews. A similar approach is used by Lei et al. [54]. They perform attribute prediction to determine which

attribute should be used for user preference elicitation. Subsequently, attribute-aware Bayesian Personalized Ranking (BPR) is used to provide recommendations. The attribute-aware BPR re-ranks items based on the relationship between an item and its attributes, as well as the user’s preference for that specific attribute. Li et al. [65] attempt to model both items and item attributes in the same space for user preference elicitation. Similarly, Hu et al. [40] combine item-level and attribute-level user feedback for user preference elicitation. Other works on attribute-based user preference elicitation include the work of [10, 55, 133, 120, 107].

4.2.2 Dialogue Strategy for CRSs

Another line of work attempts to modulate the conversation strategy to improve CRSs. Kang et al. [46] use multi-objective training and self-supervised bot-play to learn the best dialogue strategy for a conversational recommendation. The initial multi-objective training combines the losses associated with mimicking human utterances, deciding whether to recommend or not, and predicting the correct movie. Then self-supervised bot-play is used to train a reinforcement learning system for maximizing dialogue goal-related rewards.

Zhou et al. [131] propose using a topic prediction model that keeps track of topic transition throughout the conversation. Once the natural flow of the dialogue reaches a specific confidence threshold, the system triggers the recommendation turn. The goal is to facilitate better transition from non-recommendation turn to recommendation turn. Sun et al. [105] propose a deep reinforcement learning framework for learning a dialogue strategy. At each step, the agent can decide to recommend items or ask questions about the user’s preference regarding a single attribute (from a fixed set of attributes). The agent receives a reward when it asks about a relevant attribute or makes a relevant recommendation. The system is optimized to maximize the overall session based reward rather than the next turn reward. A similar approach is adopted by [54, 69, 12].

4.2.3 Language Models for CRSs

The effectiveness of conversational recommendation depends on the capability of the language models used for CRSs. Some prior works on CRSs such as [64, 46] use Recurrent Neural Networks (RNN) [38] for the dialogue system. Since RNNs need to be trained from scratch, CRSs using this type of language model are data intensive. This type of systems cannot benefit from additional commonsense knowledge found in large unpaired corpora and solely relies on task-specific datasets for all knowledge. Later, pre-trained language models based on Transformer [108] have become widely used. These types of language models are pre-trained on large text corpora and can be adapted to specific dialogue settings with much smaller task-specific datasets. [114, 110] are examples of CRSs that use transformer-based generative language models. This type of CRSs benefit from pre-training and performs better than RNN-based models in dialogue. However, they still require fine-tuning on domain-specific datasets and may be hard to generalize in recommendation domains where ground-truth conversations are hard to collect.

More recent language models use massive scaling [5] and in some cases instruction-tuning [79, 6]. These LLMs are good at storing common sense world knowledge and can adapt to task-specific settings with a simple prompt. These qualities make them ideal for developing CRSs. Even though state-of-the-art LLMs can serve as recommendation systems themselves, they perform poorly on direct recommendation, sequential recommendation, and cold start scenarios [66]. Additionally, LLMs’ knowledge can become out-of-date due to evolving item space in CRSs and they often demonstrate bias towards popular items [122, 36]. Allowing an external recommendation system to exploit the LLM-generated intermediate representations can potentially remedy these issues [21, 20, 75]. LLMs serve as controller for external recommendation modules in this hybrid paradigm which mitigates limitations associated with LLMs’ parametric knowledge. However, the LLM-generated intermediate representations often lead to issues like cross-trial robustness and cross-domain generalization which we aim to

address in our work.

4.3 Related Work for RoleFact

4.3.1 Language Models for Role-play

Building autonomous agents that can simulate human behavior with individualized traits is a long-standing challenge [25]. Prior works in NLP build conversational agents grounded with various individualized traits, such as background stories [124, 80], demographic information [42, 13], social roles [132], and response style constraints [31]. In this work, we focus on building dialogue systems that are aligned with user perceptions of fictional characters, known as character role-play [97, 8, 61]. In terms of methodology, prior works can be categorized as training-free role-prompting [111, 129, 61] or instruction-tuning [97, 71]. Role-prompting includes static character descriptions [129, 111] and retrieved past experiences [61, 111]. In terms of role-playing objectives, prior works focused on personality traits [92, 96], speaker style [8, 127], and character fidelity [116, 97]. We specifically focus on character fidelity with respect to story-specific knowledge and relevant timelines. A character role-play dialogue system should produce responses aligned with facts related to its assigned character and storyline and it should demonstrate a lack of knowledge about things that do not apply to them. [97]. However, most existing methods of role-play lack mechanisms to reduce parametric memory-induced hallucination. Moreover, heavy reliance on parametric memory leads to poor performance for less popular characters since LLMs may not have adequate parametric knowledge about them. To the best of our knowledge, our work is the first one to modulate the influence of parametric knowledge to mitigate hallucination. This is also the first dataset to facilitate role-playing for less popular characters.

4.3.2 Factuality in Natural Language Generation

Beyond conversational systems, improving factuality in generated contents is an open problem in NLP, benefiting domains such as summarization [7, 16], knowledge grounded dialogue systems [18], and question answering [39, 23]. While fine-grained evaluation of factuality based on atomic facts has been studied in NLP [74], there is currently no dataset for character role-play that facilitates similar evaluation of factuality. Therefore, prior works on role-play [111, 61, 97] have relied on rating-based evaluation of factuality which is expensive for humans. Our dataset is the first one to facilitate automated fine-grained analysis of factuality by matching decomposed atomic facts against script-specific knowledge.

4.3.3 Retrieval Augmented Methods

Our work is also related to Retrieval-augmented methods in NLP that improve model output by incorporating retrieved external knowledge, typically for generation tasks [49, 59, 30]. While this class of method has been used in building role-playing agents [124, 80, 97, 8, 61], these works perform retrieval in a time-agnostic manner. In order to study the temporal dimension of character role-play (i.e. character development), we need a dataset where script knowledge is aligned to temporal annotations. Our dataset provides temporal annotations with utterance-level granularity.

Chapter 5

Conclusion and Future Work

Parametric knowledge plays a foundational role in the capabilities of modern language models. It empowers models to generalize, perform a wide range of tasks, and provide fallback knowledge when external tools are unavailable. However, our study highlights several challenges associated with relying solely on parametric knowledge, including limited recall of long-tail facts, outdated information, and the generation of hallucinated content. This dissertation introduced novel methods across the training and inference pipeline to address these issues: (1) *InforMask* to improve factual knowledge acquisition, (2) *IPRA* to align model representations with downstream modules in hybrid systems, and (3) *RoleFact*, a calibrated suppression mechanism to mitigate hallucinations during fictional character role-playing.

Future Research Directions

Future work can build upon the insights of this dissertation by exploring deeper innovations in each of the three thrust areas: acquiring, leveraging, and suppressing parametric knowledge.

Acquiring Parametric Knowledge

- **InforMask for CLM:** Future work can investigate token-specific masking rates for autoregressive language models to enable more efficient pretraining and improved acquisition of factual knowledge.

- **Parametric Knowledge Editing:** A promising direction for future research involves directly editing model parameters to insert or update relational knowledge. Although several knowledge editing techniques exist [112], most struggle to modify facts about a target subject without affecting other facts related to the same subject.

Leveraging Parametric Knowledge

- **Beyond Accuracy Representation Alignment:** Future systems could adapt the alignment strategy (e.g., IPRA) based not only on recommendation accuracy but also on additional signals such as diversity or serendipity.
- **IPRA for Multi-modal CRSs:** Extending LLM-powered hybrid CRSs to multi-modal settings (e.g., fashion or music recommendation) remains an interesting direction for future research.

Suppressing Parametric Knowledge

- **Controllable Decoding for Hallucination Mitigation:** While our existing work addresses post-decoding updates to hallucinated text, future research could explore decoding-time controls (beyond confidence thresholds) to allow on-the-fly prioritization between parametric and non-parametric knowledge sources when conflicts arise.
- **Pre-decoding Filters for Hallucination Mitigation:** A promising direction for future research is the development of pre-decoding filters that can selectively avoid responding to queries involving sensitive topics—such as legal, medical, or tax-related advice—until a human expert can validate the LLM-generated responses.

Bibliography

- [1] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection, 2023.
- [2] Jiangang Bai, Yujing Wang, Yiren Chen, Yaming Yang, Jing Bai, Jing Yu, and Yunhai Tong. Syntax-bert: Improving pre-trained transformers with syntax trees. In Paola Merlo, Jörg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 3011–3020. Association for Computational Linguistics, 2021.
- [3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *CoRR*, abs/2309.16609, 2023.
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner,

- Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [6] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.
- [7] Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. Faithful to the original: fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.
- [8] Nuo Chen, Yan Wang, Haiyun Jiang, Deng Cai, Yuhan Li, Ziyang Chen, Longyue Wang, and Jia Li. Large language models meet harry potter: A dataset for aligning dialogue agents with characters. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8506–8520, Singapore, December 2023. Association for Computational Linguistics.
- [9] Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [10] Konstantina Christakopoulou, Alex Beutel, Rui Li, Sagar Jain, and Ed H. Chi. Q&r: A two-stage approach toward interactive recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 139–148, New York, NY, USA, 2018. Association for Computing Machinery.
- [11] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 815–824, New York, NY, USA, 2016. Association for Computing Machinery.
- [12] Zhendong Chu, Hongning Wang, Yun Xiao, Bo Long, and Lingfei Wu. Meta policy learning for cold-start conversational recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 222–230, New York, NY, USA, 2023. Association for Computing Machinery.

- [13] Yun-Shiuan Chuang, Zach Studdiford, Krirk Nirunwiroj, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Beyond demographics: Aligning role-playing llm-based agents using human belief networks, 2024.
- [14] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.
- [16] Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. Multi-fact correction in abstractive text summarization. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online, November 2020. Association for Computational Linguistics.
- [17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billoock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024.
- [18] Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In Marie-Francine

- Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [19] Robert M Fano. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794, 1961.
 - [20] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. A large language model enhanced conversational recommender system, 2023.
 - [21] Luke Friedman, Sameer Ahuja, David Allen, Zhenning Tan, Hakim Sidahmed, Changbo Long, Jun Xie, Gabriel Schubiner, Ajay Patel, Harsh Lara, Brian Chu, Zexi Chen, and Manoj Tiwari. Leveraging large language models in conversational recommender systems, 2023.
 - [22] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. Advances and challenges in conversational recommender systems: A survey. *AI Open*, 2:100–126, 2021.
 - [23] Xiang Gao, Jiabin Zhang, Lalla Mouatadid, and Kamalika Das. Spuq: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, 2024.
 - [24] Yunfan Gao, Tao Sheng, Youlin Xiang, Yun Xiong, Haofen Wang, and Jiawei Zhang. Chat-rec: Towards interactive and explainable llms-augmented recommender system, 2023.
 - [25] Matthew L. Ginsberg. Marvin minsky, the society of mind. *Artif. Intell.*, 48(3):335–339, 1991.
 - [26] Google. Gemini: A family of highly capable multimodal models, 2024.
 - [27] GSMArena. Gsmarena.com - mobile phone reviews, news, specs, 4 2025. Accessed: 2025-04-06.
 - [28] Xinyan Guan, Yanjiang Liu, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. Mitigating large language model hallucinations via autonomous knowledge graph-based retrofitting. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18126–18134, Mar. 2024.
 - [29] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm:

- Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org, 2020.
- [30] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML’20. JMLR.org, 2020.
- [31] Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5114–5132, Seattle, United States, July 2022. Association for Computational Linguistics.
- [32] F. Maxwell Harper and Joseph A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), dec 2015.
- [33] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [34] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, WWW ’17, page 173–182, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.
- [35] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian J. McAuley. Large language models as zero-shot conversational recommenders. In Ingo Frommholz, Frank Hopfgartner, Mark Lee, Michael Oakes, Mounia Lalmas, Min Zhang, and Rodrygo L. T. Santos, editors, *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management, CIKM 2023, Birmingham, United Kingdom, October 21-25, 2023*, pages 720–730. ACM, 2023.
- [36] Zhankui He, Zhouhang Xie, Harald Steck, Dawen Liang, Rahul Jha, Nathan Kallus, and Julian McAuley. Reindex-then-adapt: Improving large language models for conversational recommendation. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*, WSDM ’25, page 866–875, New York, NY, USA, 2025. Association for Computing Machinery.
- [37] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings*

of the International Conference on Learning Representations (ICLR), 2021.

- [38] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- [39] Or Honovich, Leshem Choshen, Roei Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. q^2 : Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [40] Chenhao Hu, Shuhua Huang, Yansen Zhang, and Yubao Liu. Learning to infer user implicit preference in conversational recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 256–266, New York, NY, USA, 2022. Association for Computing Machinery.
- [41] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents, 2022.
- [42] EunJeong Hwang, Bodhisattwa Majumder, and Niket Tandon. Aligning language models to user opinions. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5906–5919, Singapore, December 2023. Association for Computational Linguistics.
- [43] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning, 2022.
- [44] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.
- [45] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.
- [46] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul Crook, Y-Lan Boureau, and Jason Weston. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1951–1961, Hong Kong, China, November

2019. Association for Computational Linguistics.

- [47] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.
- [48] Pei Ke, Haozhe Ji, Siyang Liu, Xiaoyan Zhu, and Minlie Huang. Sentilare: Sentiment-aware language representation learning with linguistic knowledge. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6975–6988. Association for Computational Linguistics, 2020.
- [49] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.
- [50] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.
- [51] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.
- [52] Anne Lauscher, Olga Majewska, Leonardo F. R. Ribeiro, Iryna Gurevych, Nikolai Rozanov, and Goran Glavas. Common sense or world knowledge? investigating adapter-based knowledge injection into pretrained transformers. *CoRR*, abs/2005.11787, 2020.
- [53] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia—a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195, 2015.
- [54] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM ’20*, page 304–312, New York, NY, USA, 2020. Association for Computing Machinery.
- [55] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. Interactive path reasoning on graph for conversational recommendation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery*

& *Data Mining*, KDD '20, page 2073–2083, New York, NY, USA, 2020. Association for Computing Machinery.

- [56] Yoav Levine, Barak Lenz, Or Dagan, Ori Ram, Dan Padnos, Or Sharir, Shai Shalev-Shwartz, Amnon Shashua, and Yoav Shoham. Sensebert: Driving some sense into BERT. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4656–4667. Association for Computational Linguistics, 2020.
- [57] Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. Pmi-masking: Principled masking of correlated spans. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [58] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.
- [59] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20*, Red Hook, NY, USA, 2020. Curran Associates Inc.
- [60] Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, et al. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [61] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chatharuhi: Reviving anime character in reality via large language model. *CoRR*, abs/2308.09597, 2023.
- [62] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web, WWW '10*, page 661–670, New York, NY, USA, 2010.

Association for Computing Machinery.

- [63] Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, WSDM '11*, page 297–306, New York, NY, USA, 2011. Association for Computing Machinery.
- [64] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18*, page 9748–9758, Red Hook, NY, USA, 2018. Curran Associates Inc.
- [65] Shijun Li, Wenqiang Lei, Qingyun Wu, Xiangnan He, Peng Jiang, and Tat-Seng Chua. Seamlessly unifying attributes and items: Conversational recommendation for cold-start users. *ACM Trans. Inf. Syst.*, 39(4), aug 2021.
- [66] Junling Liu, Chao Liu, Peilin Zhou, Renjie Lv, Kang Zhou, and Yan Zhang. Is chatgpt a good recommender? a preliminary study, 2023.
- [67] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-BERT: enabling language representation with knowledge graph. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 2901–2908. AAAI Press, 2020.
- [68] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.
- [69] Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online, July 2020. Association for Computational Linguistics.
- [70] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.
- [71] Keming Lu, Bowen Yu, Chang Zhou, and Jingren Zhou. Large language models are superpositions of all characters: Attaining arbitrary role-play via self-alignment. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7828–7840, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

- [72] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, page 43–52, New York, NY, USA, 2015. Association for Computing Machinery.
- [73] Meta. Introducing meta llama 3: The most capable openly available llm to date. <https://ai.meta.com/blog/meta-llama-3/>, 2024. Accessed: 2024-06-13.
- [74] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023.
- [75] Guangtao Nie, Rong Zhi, Xiaofan Yan, Yufan Du, Xiangyang Zhang, Jianwei Chen, Mi Zhou, Hongshen Chen, Tianhao Li, Ziguang Cheng, Sulong Xu, and Jinghe Hu. A hybrid multi-agent conversational recommender system with llm and search engine in e-commerce. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 745–747, New York, NY, USA, 2024. Association for Computing Machinery.
- [76] OpenAI. Gpt-4 technical report, 2024.
- [77] openai. Openai - vector embeddings, 4 2025. Accessed: 2025-04-06.
- [78] openai. Openai's gpt-3.5-turbo model, 4 2025. Accessed: 2025-04-06.
- [79] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, New Orleans, LA, USA, 2022. Curran Associates, Inc.
- [80] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery.
- [81] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.
- [82] Matthew E. Peters, Mark Neumann, Robert L. Logan IV, Roy Schwartz, Vidur Joshi, Sameer Singh, and Noah A. Smith. Knowledge enhanced contextual word representations.

- In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 43–54. Association for Computational Linguistics, 2019.
- [83] Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick S. H. Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics, 2019.
- [84] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- [85] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS ’23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [86] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.
- [87] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don’t know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [88] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.
- [89] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [90] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese

- BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics.
- [91] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389, apr 2009.
 - [92] Yisi Sang, Xiangyang Mou, Mo Yu, Dakuo Wang, Jing Li, and Jeffrey Stanton. MBTI personality prediction for fictional characters using movie scripts. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6715–6724, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.
 - [93] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools, 2023.
 - [94] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings*, page 593–607, Berlin, Heidelberg, 2018. Springer-Verlag.
 - [95] Suvash Sedhain, Aditya Krishna Menon, Scott Sanner, and Lexing Xie. Autorec: Autoencoders meet collaborative filtering. In *Proceedings of the 24th International Conference on World Wide Web, WWW ’15 Companion*, page 111–112, New York, NY, USA, 2015. Association for Computing Machinery.
 - [96] Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. Personality traits in large language models, 2023.
 - [97] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-LLM: A trainable agent for role-playing. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore, December 2023. Association for Computational Linguistics.
 - [98] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation, 2021.
 - [99] Vishvesh Soni. Large language models for enhancing customer lifecycle management. *Journal of Empirical Social Science Studies*, 7(1):67–89, Feb. 2023.

- [100] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI’17, page 4444–4451. AAAI Press, 2017.
- [101] Mechthild Stoer and Frank Wagner. A simple min-cut algorithm. *Journal of the ACM (JACM)*, 44(4):585–591, 1997.
- [102] Yusheng Su, Xu Han, Zhengyan Zhang, Yankai Lin, Peng Li, Zhiyuan Liu, Jie Zhou, and Maosong Sun. Cokebert: Contextual knowledge selection and embedding towards enhanced pre-trained language models. *AI Open*, 2:127–134, 2021.
- [103] Tianxiang Sun, Yunfan Shao, Xipeng Qiu, Qipeng Guo, Yaru Hu, Xuanjing Huang, and Zheng Zhang. Colake: Contextualized language and knowledge embedding. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3660–3670. International Committee on Computational Linguistics, 2020.
- [104] Yu Sun, Shuohuan Wang, Yu-Kun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. *CoRR*, abs/1904.09223, 2019.
- [105] Yueming Sun and Yi Zhang. Conversational recommender system. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR ’18*, page 235–244, New York, NY, USA, 2018. Association for Computing Machinery.
- [106] Quan Tu, Chuanqi Chen, Jinpeng Li, Yanran Li, Shuo Shang, Dongyan Zhao, Ran Wang, and Rui Yan. Characterchat: Learning towards conversational ai with personalized social support, 2023.
- [107] Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. Conversational recommendation via hierarchical information modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’22*, page 2201–2205, New York, NY, USA, 2022. Association for Computing Machinery.
- [108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

- [109] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.
- [110] Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Daxin Jiang, and Kam-Fai Wong. RecInDial: A unified framework for conversational recommendation with pretrained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 489–500, Online only, November 2022. Association for Computational Linguistics.
- [111] Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 14743–14777, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.
- [112] Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. Easyedit: An easy-to-use knowledge editing framework for large language models. *CoRR*, abs/2308.07269, 2023.
- [113] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. Rethinking the evaluation for conversational recommendation in the era of large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 10052–10065. Association for Computational Linguistics, 2023.
- [114] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD ’22*, page 1929–1937, New York, NY, USA, 2022. Association for Computing Machinery.
- [115] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021.

- [116] Xintao Wang, Yunze Xiao, Jen-tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1840–1873, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
- [117] Alexander Wettig, Tianyu Gao, Zexuan Zhong, and Danqi Chen. Should you mask 15% in masked language modeling? *CoRR*, abs/2202.08005, 2022.
- [118] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics, 2020.
- [119] Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. Expertprompting: Instructing large language models to be distinguished experts, 2023.
- [120] Kerui Xu, Jingxuan Yang, Jun Xu, Sheng Gao, Jun Guo, and Ji-Rong Wen. Adapting user preference to online feedback in multi-round conversational recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM ’21*, page 364–372, New York, NY, USA, 2021. Association for Computing Machinery.
- [121] Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6442–6454, Online, November 2020. Association for Computational Linguistics.
- [122] Ting Yang and Li Chen. Unleashing the retrieval potential of large language models in conversational recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems, RecSys ’24*, page 43–52, New York, NY, USA, 2024. Association for Computing Machinery.
- [123] YouTube. Youtube data api, 4 2025. Accessed: 2025-04-06.
- [124] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason

- Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics.
- [125] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. Towards conversational search and recommendation: System ask, user respond. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 177–186, New York, NY, USA, 2018. Association for Computing Machinery.
- [126] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 1441–1451. Association for Computational Linguistics, 2019.
- [127] Runcong Zhao, Wenjia Zhang, Jiazheng Li, Lixing Zhu, Yanran Li, Yulan He, and Lin Gui. NarrativePlay: Interactive narrative understanding. In Nikolaos Aletras and Orphee De Clercq, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 82–93, St. Julians, Malta, March 2024. Association for Computational Linguistics.
- [128] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA, 2023. Curran Associates Inc.
- [129] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Libiao Peng, Jiaming Yang, Xiyao Xiao, Sahand Sabour, Xiaohan Zhang, Wenjing Hou, Yijia Zhang, Yuxiao Dong, Jie Tang, and Minlie Huang. Characterglm: Customizing chinese conversational ai characters with large language models, 2023.
- [130] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '20*, page 1006–1014, New York, NY, USA, 2020. Association for Computing Machinery.
- [131] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

- [132] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. SOTOPIA: Interactive evaluation for social intelligence in language agents. In *The Twelfth International Conference on Learning Representations*, 2024.
- [133] Jie Zou, Yifan Chen, and Evangelos Kanoulas. Towards question-based recommender systems. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '20*, page 881–890, New York, NY, USA, 2020. Association for Computing Machinery.