Recipe-MPR: A Test Collection for Evaluating Multi-aspect Preference-based Natural Language Retrieval

Haochen Zhang*
University of Toronto
Toronto, Canada
hcz.zhang@mail.utoronto.ca

Anton Korikov*
University of Toronto
Toronto, Canada
anton.korikov@mail.utoronto.ca

Parsa Farinneya
University of Toronto
Toronto, Canada
parsa.farinneya@mail.utoronto.ca

Mohammad Mahdi Abdollah Pour
University of Toronto
Toronto, Canada
m.abdollahpour@mail.utoronto.ca

Manasa Bharadwaj
LG Electronics, Toronto AI Lab
Toronto, Canada
manasa.bharadwaj@lge.com

Ali Pesaranghader
LG Electronics, Toronto AI Lab
Toronto, Canada
ali.pesaranghader@lge.com

Xi Yu Huang
University of Toronto
Toronto, Canada
xiyu.huang@mail.utoronto.ca

Yi Xin Lok
University of Toronto
Toronto, Canada
y.lok@mail.utoronto.ca

Zhaoqi Wang
University of Toronto
Toronto, Canada
mr.wang@mail.utoronto.ca

Nathan Jones
University of Toronto
Toronto, Canada
nathan.jones@mail.utoronto.ca

Scott Sanner†
University of Toronto
Toronto, Canada
ssanner@mie.utoronto.ca

ABSTRACT

The rise of interactive recommendation assistants has led to a novel domain of natural language (NL) recommendation that would benefit from improved multi-aspect reasoning to retrieve relevant items based on NL statements of preference. Such preference statements often involve multiple aspects, e.g., "I would like **meat lasagna** but I'm **watching my weight**". Unfortunately, progress in this domain is slowed by the lack of annotated data. To address this gap, we curate a novel dataset[1] which captures logical reasoning over multi-aspect, NL preference-based queries and a set of multiple-choice, multi-aspect item descriptions. We focus on the recipe domain in which multi-aspect preferences are often encountered due to the complexity of the human diet. The goal of publishing our dataset is to provide a benchmark for joint progress in three key areas: 1) structured, multi-aspect NL reasoning with a variety of properties (e.g., level of specificity, presence of negation, and the need for commonsense, analogical, and/or temporal inference), 2) the ability of recommender systems to respond to NL preference utterances, and 3) explainable NL recommendation facilitated by aspect extraction and reasoning. We perform experiments using a variety of methods (sparse and dense retrieval, zero- and few-shot reasoning with large language models) in two settings: a *monolithic* setting which uses the full query and an *aspect-based* setting which isolates individual query aspects and aggregates the results. GPT-3 results in much stronger performance than other methods with 73% zero-shot accuracy and 83% few-shot accuracy in the monolithic setting. Aspect-based GPT-3, which faciliates structured explanations, also shows promise with 68% zero-shot accuracy. These results establish baselines for future research into explainable recommendations via multi-aspect preference-based NL reasoning.

CCS CONCEPTS

• **Information systems** → **Test collections**; • **Applied computing** → **Document searching**.

KEYWORDS

multi-aspect preference retrieval, natural language reasoning, recipe retrieval, benchmark dataset

ACM Reference Format:
Haochen Zhang, Anton Korikov, Parsa Farinneya, Mohammad Mahdi Abdollah Pour, Manasa Bharadwaj, Ali Pesaranghader, Xi Yu Huang, Yi Xin Lok, Zhaoqi Wang, Nathan Jones, and Scott Sanner. 2023. Recipe-MPR: A Test Collection for Evaluating Multi-aspect Preference-based Natural Language Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '23), July 23–27, 2023, Taipei, Taiwan.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3539618.3591880

*Both authors contributed equally to this research.
†Affiliate to Vector Institute of Artificial Intelligence, Toronto
[1]https://github.com/D3Mlab/Recipe-MPR

SIGIR '23, July 23–27, 2023, Taipei, Taiwan.

ACM ISBN 978-1-4503-9408-6/23/07…$15.00
https://doi.org/10.1145/3539618.3591880

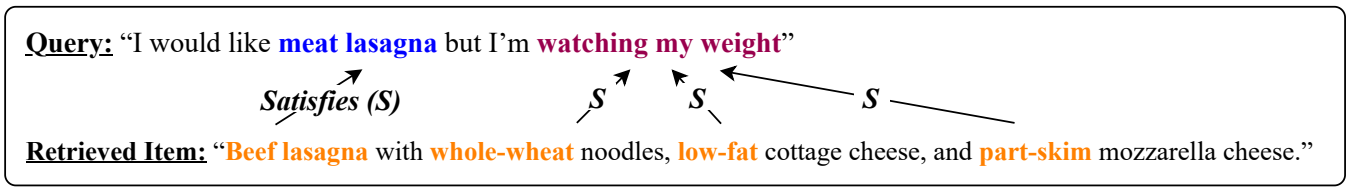


Figure 1: Preference Satisfaction: Each preference aspect (blue/purple span) is satisfied by at least one item aspect (orange span).

Table 1: Examples of data entries in Recipe-MPR.

Example 1	
Query	I would like meat lasagna but I'm watching my weight
Properties	☑ Specific ☑ Commonsense ☐ Negated ☐ Analogical ☐ Temporal
Options	☐ Vegetarian lasagna with mushrooms, mixed vegetables, textured vegetable protein, and meat replacement ☐ Forgot the Meat Lasagna with onions, mushrooms and spinach ☑ Beef lasagna with whole-wheat noodles, low-fat cottage cheese, and part-skim mozzarella cheese ☐ Cheesy lasagna with Italian sausage, mushrooms, and 8 types of cheese ☐ Meat loaf containing vegetables such as potatoes, onions, corn, carrots, and cabbage
Aspects	beef lasagna → meat lasagna whole-wheat, low-fat, part-skim → watching my weight
Example 2	
Query	I want chicken that has a kick to it
Properties	☐ Specific ☑ Commonsense ☐ Negated ☐ Analogical ☐ Temporal
Options	☐ Cheese chicken made with chicken legs, eggs, cheese, and breading ☑ Chicken wings made with hot chili sauce, butter, and Worcestershire sauce ☐ Hard-cooked egg with a sriracha kick ☐ Easy chicken legs made with Italian salad dressing ☐ Shrimp With a Kick - made with garlic, olive oil, and fresh cilantro
Aspects	Chicken → chicken hot chili sauce → kick

1 INTRODUCTION

Interactive, natural language (NL) AI assistants are developing rapidly, both in terms of their performance on a variety of real world tasks [5, 32, 39, 40] and the scale of their deployment (e.g., ChatGPT[2]). Paramount to the development of effective NL assistants is the ability to provide precise feedback on task performance [14, 29, 42], motivating the need for task-specific datasets. Amidst the recent progress on generative, dialogue-orientated tasks [11, 14, 28], the integration of information retrieval abilities into conversational assistants remains an open and pressing challenge [1, 10]. Among many retrieval-based tasks, a primary domain is conversational recommendation (ConvRec) [9]: the recommendation of items based on a user's NL description of her preferences.

Studies of human-to-human recommendation interactions [23] have shown that NL preference statements often involve multiple *preference aspects*: facets (or parts) of a preference that require independent reasoning. Furthermore, an open challenge remains the integration of structured reasoning abilities into conversational AI agents [21, 22, 41, 44]. For these reasons, we introduce a novel dataset which models the task of satisfying each individual preference aspect in a preference statement based on attributes inferred from a set of item descriptions, following the form of a multiple-choice question answering task [35].

Specifically, we study user statements (which in our problem setting take the role of *queries*) where preference aspects are identifiable by spans in the utterance, such as the bold spans in the query *"I would like **meat lasagna** but I'm **watching my weight**"*. Similarly, we consider a set of item descriptions in which certain spans identify *item aspects*: qualities of an item which require independent reasoning during retrieval. An example of such a description with bold spans identifying item aspects relevant to the above query is "***Beef lasagna** with **whole-wheat** noodles, **low-fat** cottage cheese, and **part-skim** mozzarella cheese*". In our dataset, this lasagna recipe is relevant to retrieve for the above preference statement because of *satisfaction* relations between item aspects and preference aspects. Specifically, as illustrated in Figure 1, we say that "***beef lasagna***" satisfies "***meat lasagna***", and "***whole-wheat***", "***low-fat***", and "***part-skim***" each satisfy "***watching my weight***".

In this model of a ConvRec task, *an item is deemed relevant to a query if each preference aspect in the query is satisfied by at least one item aspect in the item description.* For this task, we introduce a dataset consisting of NL queries with labeled preference aspects, a set of candidate NL item descriptions, and an item description identified as relevant for each query with labeled item aspects. We focus on the recipe domain, which often involves multi-aspect preferences due to the complexity of the human diet and personal food preferences. In addition, our dataset also labels whether a preference expression uses any of several reasoning strategies such as analogical, negated, and commonsense reasoning. For instance, the query "Can I get a breakfast that's easy to eat on-the-go, like a wrap?" is identified as using analogical and commonsense reasoning. We call our dataset the Recipe Multi-aspect Preference-based Retrieval dataset (Recipe-MPR), for which two examples of complete data entries are given in Table 1.

While there are alternative ways to benchmark ConvRec performance, there are several reasons this formulation and dataset are worth studying. Specifically, our dataset evaluates a system's ability to infer multi-aspect preference satisfaction from NL, which we conjecture is a key structural element of ConvRec. In addition, the explicit *item aspect*-satisfies-*preference aspect* structure of our data facilitates explanations, such as the correctness explanation in Figure 1. Furthermore, the relationship between our problem definition, the satisfiability problem (SAT) and textual entailment [24] provide opportunities for future research into the integration of NL reasoning and symbolic reasoning in ConvRec (see Section 4).

[2]https://openai.com/blog/chatgpt/

Finally, while most ConvRec datasets focus on multi-turn interactions, our data isolates a single interaction step, focusing on cases when a recommendation can be made in a single turn.

In addition to providing new data and theoretical discussion, we present numerical experiments evaluating several baselines on our dataset. As well as investigating a monolithic setting where models are given the full query as an input, we also explore a basic form of aspect-level reasoning by modifying the input to sequentially isolate individual aspects and then aggregate the results, These aspect-based baselines are a step towards recommendations that are more explainable and verifiable at an aspect level. We find that sparse retrieval methods (OWC, TF-IDF [37], BM25 [34]) have very poor performance (less than 23% accuracy) due to their reliance on exact term matches. We test multiple large language models (LLMs) (BERT [6], TAS-B [13], OPT [45], GPT-2 [30], GPT-3 [27]) in few-shot and zero-shot settings and find monolithic GPT-3 provides the best results with few-shot and zero-shot accuracies of 83.4% and 72.6%, respectively. Despite performing slightly worse, our best explicit aspect-level result (zero-shot GPT-3) is promising with 67.6% accuracy, which is close to full-query zero-shot GPT-3. Improving the accuracy of explicit aspect-level reasoning is an obvious future research direction. However, our dataset can also further develop such explicit NL aspect-based reasoning by supporting research into areas such as aspect extraction or the joint optimization for explainability and recommendation performance (see Section 4).

2 RELATED WORK

ConvRec Datasets

Existing ConvRec datasets can be broadly categorized as synthetic or human-generated [9], with the majority focusing on multi-turn dialogues. Synthetic datasets such as ConvRec [15], TG-Redial [46], and COOKIE [8] contain simulated or partially-simulated dialogues derived from user-item data and conversation templates. While synthetic data can be produced in large volumes, it is typically of lower quality than human-generated data.

Several non-synthetic datasets such as ReDial [20], MovieSent [38], and CCPE-M [31] contain annotated NL data from human dialogues. ReDial is a multi-turn dialogue dataset of crowdworker movie recommendation interactions with certain utterances that refer to specific movies annotated with "*liked*", "*didn't like*", or "*didn't say*" tags. CCPE-M is a similar dataset of crowdsourced dialogues with certain utterances annotated as containing entities, entity descriptions and entity preferences. MovieSent is an extension of this dataset to include additional entity (movie) information based on RottenTomatoes[3] and sentiment labels for user utterances.

While containing valuable data, these datasets primarily focus on eliciting human preference through dialogue. In contrast, our work studies a setting where we are given a clear NL preference statement. Furthermore, much of the annotation in the above datasets concerns recording users' responses toward a specific item (e.g., "liked"). In contrast, we focus on annotating the *reasons why* a recommendation is valid by annotating preference aspects, item aspects, and satisfaction relations between them. Our preference statements and item descriptions are also explicitly *multi-aspect*, which is not necessarily true for existing datasets.

[3]https://www.rottentomatoes.com/

Rationale-Labeled Datasets

Though it deals with a different set of domains, tasks, and labeling approaches, work on the **E**valuating **R**ationales **A**nd **S**imple **E**nglish **R**easoning (ERASER) [7] datasets is highly relevant to ours. ERASER is a collection of seven datasets that explores the use of labeled spans as rationales for various NL tasks. It includes the Commonsense Explanations (CoS-E) corpus [32], where rationales are spans of a multiple choice question (MCQ) that support a correct answer, such as the bold text in the question *"Where do you find the* ***most amount of leaves****?"* for a correct answer "*Forest*". ERASER includes two more question-answering corpora, MultiRC [17] and BoolQ [5], and a sentiment analysis corpus, Movie Review [43], each with similarly annotated rationales.

Also included is the **e**xplanation-augmented **S**tanford **N**atural **L**anguage **I**nference (e-SNLI) corpus [4], where rationales are labeled for the task of inferring one of three principal NL inference (NLI) relations: entailment, contradiction, or neutral. Of these three relations, the connection between preference satisfaction and textual entailment, which specifies that a hypothesis is true if a premise is true, will be discussed further in Section 4. An example of rationales for entailment in e-SNLI are the bold spans in the premise "*A man in an orange vest* ***leans over a pickup truck***." and hypothesis "*A man is* ***touching a truck***."

In addition to collecting data, the authors of ERASER investigate how NLP methods can extract rationales during tasks and how the impact of these rationales on predictions can be measured. Based on Lei *et al.* [19], they consider two-step "hard" extraction where an encoder identifies rationalizing spans and then an independent decoder uses these spans as inputs for predictions. In addition, they consider "soft" extraction which assigns a continuous importance score to tokens using feature-importance explainability methods (gradients, attention, LIME [33]). They propose two measures for assessing the significance of rationales for prediction. *Comprehensiveness* is the change in confidence for the correct prediction when rationales are deleted from the input (expected to be a loss for valid rationales), and *sufficiency* is the change in this confidence when everything except the rationales is deleted from the input.

Though ERASER investigates similar ideas, our work explores a new domain, task structure, and annotation method. Specifically, we focus on annotating and studying the satisfaction of preference aspects by item aspects in ConvRec contexts. Furthermore, our data reflects a precise multi-aspect satisfaction structure, of which the implications for hybrid NL/symbolic reasoning are discussed further in Section 4.

Multi-aspect Retrieval

Work has been done in the information retrieval field that considers multiple aspects – specifically, Kong *et al.*[18] consider multiple aspects when calculating relevance scores in dense retrieval. However, their work uses a proprietary dataset where queries and documents contain a fixed number of aspects from known categories. Similarly, the label aggregation method of Kang *et al.* [16] has some similarities to our task, but assumes there are a fixed number of known categories; an unrealistic assumption for NL preference expressions in ConvRec settings.

3 THE RECIPE-MPR DATASET

3.1 Overview

To model and benchmark progress on the ConvRec task, we introduce a manually-curated, publicly released dataset, Recipe-MPR.[4] Our dataset contains a set of NL preference statements (queries) and for each query, as shown in Table 1, a set of NL item descriptions where one item is marked as a relevant recommendation. Furthermore, we identify spans in the query describing *preference aspects* and spans in the recommended item description (*item aspects*) which satisfy these preference aspects. Motivated by the need to study multi-aspect preferences [23], all queries in our dataset contain more than one preference aspect, and an item recommendation must result in each preference aspect being satisfied by at least one item aspect. These aspect satisfaction labels aim to explicitly benchmark a fundamental element of ConvRec: the NL inference of multiple item-preference satisfaction relations. We also annotate whether a preference statement uses one of several reasoning strategies such as analogical or temporal reasoning (see Section 3.2). In terms of domain, we focus on recipe recommendation since this is an area where multi-aspect preferences are typical due to the complexity of the human diet.

Our dataset consists of 500 entries constructed with the help of recipe information available in FoodKG [12] and Recipe1M+ [25]. Two example entries are shown in Table 1. The i'th entry $x_i = \{q_i, \mathcal{P}_i, O_i, a_i, \mathcal{I}_i, \mathcal{E}_i\}$ of our dataset $\mathcal{D}$ includes a query q_i (a preference statement), a set of five options $O_i = \{o_i^1, \cdots, o_i^5\}$ (item descriptions), and a unique answer index $a_i \in [1, 5]$. Each query contains at least two spans (e.g., the coloured spans in Table 1) making up a set of preference aspects $\mathcal{P}_i$, and each recommended item contains at least two spans making up a set of item aspects $\mathcal{I}_i$. The satisfaction relations are defined by the set of directed edges $\mathcal{E}_i$ which contains (j, p) if item aspect $j \in \mathcal{I}_i$ satisfies preference aspect $p \in \mathcal{P}_i$. While the underlying goal of introducing our dataset is progress towards retrieval-based ConvRec, the task we model can also be interpreted as multiple choice question-answering since only five possible options are given per query.

The query text, text description of options, and provided annotations are all manually curated by five data curators. All data curators were researchers on the project and co-authors of this paper. Each was asked not to provide any personally identifying information.

3.2 Query Generation and Property Annotation

We aimed to manually generate queries that were: a) natural, to simulate conversational language, and b) multi-aspect, to reflect the often compound nature of human preference queries. Each data curator was asked to generate 100 varied queries that did not overlap in content. Each query was also labelled according to whether it used one or more of the five following reasoning strategies:

(1) **Specific:** mentions a certain dish or recipe name, e.g., "*spaghetti carbonara*".
(2) **Commonsense:** requires commonsense reasoning, e.g., inferring "*I'm watching my weight*" to mean "*I want a low calorie meal*".
(3) **Negated:** contains contradiction or denial, using terms like "*but*", "*but not*", "*without*", "*doesn't*", etc.
(4) **Analogical:** uses metaphors or similes to express preferences using a comparison, e.g., "*like McDonald's*".
(5) **Temporal:** contains explicit references to time such as a time of day, or terms concerning the passage of time like "*slow*", "*fast*", "*lasting*", etc.

Table 2: Summary of preference reasoning strategies in Recipe-MPR.

Property	Specific	Commonsense	Negated	Analogical	Temporal
#Queries	151	268	109	30	32

The number of queries using each reasoning strategy is summarized in Table 2.

3.3 Option Generation

For each query q_i, an annotator was tasked to provide a set of five options (item descriptions) O_i, in which one option is a relevant recommendation. Recipe information was obtained from the FoodKG database, with the corresponding recipe ID being recorded for each. The knowledge graph foundation from FoodKG provides opportunities for future work that further explores the integration of discrete and NL reasoning. For each option $o_i^l \in O_i$, the curators were asked to write a brief text description for the corresponding recipe according to its name, ingredients, and nutritional information. Sometimes, additional recipe details such as the cooking method or the estimated time, were included if needed to differentiate options. The requirements for option generation provided to the data curators were:

(1) The *incorrect* options should be *hard negatives* (i.e., near misses). Hard negative options are defined as options that are close to the correct answer, but differing by at least one aspect. This requirement is motivated by the need to reflect a real-world setting, where many items may satisfy some but not all of the preference aspects. Examples of queries along with incorrect, hard negative, options are shown in Table 3.
(2) There should only be one answer that can be considered a relevant recommendation (i.e. satisfy all preference aspects) among the five options. Certain recipes may appear more than once in the dataset as a correct answer or as an option.[5]
(3) The text descriptions that are manually written do not have to contain full recipe details, but need to include enough information to discern the correct option from the wrong ones.
(4) The text descriptions should remain factual and avoid any human inference from the given recipe information.
(5) The text description for the correct answer should avoid direct word-matching with the query as much as possible.

[4] https://github.com/D3Mlab/Recipe-MPR

[5] 75% of recipes are unique.

Table 3: Examples of incorrect options. Incorrect options were designed to be hard negatives.

Query	Incorrect Option
Can I have a recipe for clam chowder that isn't too fat?	Clam chowder made with heavy whipping cream
Can I have a shrimp pasta recipe for someone with low spice tolerance?	Spaghetti noodles and shrimp containing red pepper flakes
I want to make a paella but I'm short on time	Paella made with Spanish rice mix, vegetables, and chicken sausage in a slow-cooker

3.4 Annotating Aspect Satisfaction

For the entries they generated, data curators were responsible for labeling preference aspects $\mathcal{P}_i$ and item aspects $\mathcal{I}_i$ as spans in the query and recommended item description, respectively. They also identified the preference satisfaction relations $\mathcal{E}_i$, for instance "***beef lasagna***" satisfies "***meat lasagna***" in the first entry of Table 1. In some cases, an <INFERRED> tag was used when it was not possible to identify an explicit span in the recommended item description that satisfied a preference aspect. For example, the preference aspect "***without basil***" is satisfied by the <INFERRED> tag if the item description does not include basil in the ingredients. Approximately 11% of item aspects are the <INFERRED> tag. In total, our dataset contains 1,140 preference aspects and 1,298 item aspects, with averages of 2.3 and 2.6 aspects per query and per item description, respectively. Figure 2 shows the top frequency words in the query aspects across the whole dataset.

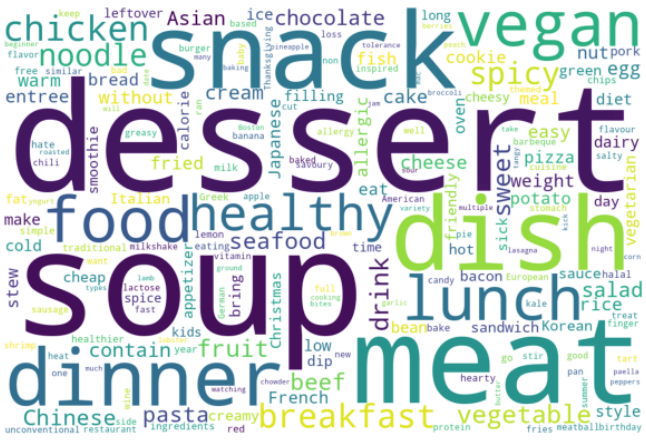


Figure 2: Most frequent words used in query aspects.

3.5 Data Validation

Following the data generation and annotation stages, the dataset was validated by curators through two rounds of validation. In each round, the queries and associated labels, options, and descriptions were validated by someone other than the original curator. Thus, each data sample was validated by two other curators other than the original curator. The data validation was to ensure:

- Queries, options, and text descriptions follow the guidelines established above.
- Labels for query reasoning strategies are used consistently and correctly.
- A single correct answer can be identified without ambiguity and without looking at the ground truth label.
- The descriptions for any recipe used multiple times are consistent.
- There are no duplicate queries.

In the second validation round, the second validator helped resolve conflicts and oversee that suggested changes were made. Changes were made only if both validators agreed on the modifications.

4 PREDICTED IMPACT

Our corpus presents structured NL data that models a recommendation task using explicit labels for preference aspects and the item aspects that satisfy them in valid recommendations. It thus supports a variety of directions for future work. By providing rationales for recommendations, this data supports work on explainable NLP [7] in a novel, multi-aspect setting. While explainable recommendation has the potential to become a sophisticated academic topic, it is also practically relevant: a recent study of human-human recommendation conversations classified 38% of utterances as explanations [23]. In addition to explainability research, by annotating individual preference satisfaction relations, this data supports future work on integrating discrete symbolic reasoning into NLP.

Our data supports multiple opportunities to explore explainable recommendation as part of the emerging field of explainable NLP, including aspect (rationale) extraction, aspect evaluation, and joint optimization for recommendation performance and explainability. As discussed in Related Work, two aspect extraction methods that can be investigated with our data are: 1) "hard" (discrete) token selection techniques implemented with an independent encoder and decoder [19], and 2) "soft" selection methods assigning a continuous score to each token implemented with a feature-importance technique (e.g. LIME [33], gradients, attention) [7]. Furthermore, it is worth exploring how rationale extraction can be adapted to account for the known multi-aspect structure in our data (i.e., that preference aspects must be independently satisfied by at least one item aspect). To the best of our knowledge, such structured, multiple-rationale extraction has not yet been explored in explainable NLP.

While a baseline to evaluate the coherency of extracted aspects with human-labeled aspects is Interval Over Union (IOU, a measure of span overlap) [7], our corpus also motivates the investigation of new coherency metrics specific to multi-aspect settings. Similarly, novel, multi-aspect notions of rationale sufficiency and comprehensiveness [7] will be valuable. In addition to work on extracting and evaluating aspects, our dataset also allows for research on the joint optimization of models for recommendation performance and coherency of extracted aspects with labeled aspects. Such bi-objective optimization could connect to past recommendation work which used latent factor models to jointly optimize recommendation quality and review text topic distributions [26].

In addition to providing a benchmark for explainable ConvRec, our corpus also facilitates the exploration of how discrete symbolic reasoning can be integrated into NL recommendation. The attentive reader may have already noticed that our model of a ConvRec task can be expressed as a SAT problem [36]. Specifically, let $\mathcal{E}_i(j, p)$

map to TRUE when item aspect $j \in \mathcal{I}_i$ satisfies preference aspect $p \in \mathcal{P}_i$. An item with aspects $\mathcal{I}_i = \{j_1, ..., j_n\}$ is relevant to a query q_i with properties $\mathcal{P}_i = \{p_1, ..., p_m\}$ when the following SAT problem evaluates to TRUE:

$$\Big(\mathcal{E}_i(j_1, p_1) \vee \mathcal{E}_i(j_n, p_1)\Big) \wedge \cdots \wedge \Big(\mathcal{E}_i(j_1, p_m) \vee \mathcal{E}_i(j_n, p_m)\Big).$$

The ability to express our data in this formal way may be useful for further investigation into how symbolic reasoning may be combined with NL inference in recommendation. There are also clear similarities between our task of inferring satisfaction relations between queries and item descriptions and the well-studied task of textual entailment [24] (see Section 2) which has the potential to be leveraged for explainable recommendation. While our experiments begin to use the NL inference abilities of LLMs to disjointly reason over individual aspects in this paper, we envision that much more sophisticated experimentation with our corpus is possible in the future.

5 EXPERIMENTAL METHODS

To assess the difficulty of our dataset and establish a starting point for future work, we evaluate several baseline models on our corpus. Full code to reproduce these experiments is included in our dataset repository.[4] Our baseline methods include sparse retrieval (OWC, TF-IDF, BM25) [34, 37], dense retrieval using LLM embeddings (BERT, TAS-B, GPT-3) [6, 13, 27], and zero- and few-shot reasoning with LLMs (OPT, GPT-2, GPT-3) [3, 30, 45]. In addition to measuring performance on the full corpus, we examine differences across the five reasoning strategies. Furthermore, we begin to explore the differences between aspect-level reasoning and query-level reasoning for these baselines. Specifically, our experiments include two settings: *monolithic*, where the input is the full query, and *aspect-based*, where the model makes separate predictions using each individual preference aspect as a separate input, after which these predictions are aggregated.

5.1 Baseline Models

- **Sparse.** Sparse methods represent queries and options as sparse vectors. We consider Overlapping Word Count (OWC) which ranks the options based on the number of terms overlapping with the query, TF-IDF [37], and BM25 [34] as sparse methods. Prior to applying these baselines, queries, and options are preprocessed with stopword removal and lemmatization using the Natural Language Toolkit [2].
- **Dense.** These neural methods represent queries and options as continuous embedding vectors to provide dense, lower-dimensional semantic representations. Dot product similarities of embedded queries and options are then used for matching. Specifically, we use pre-trained BERT[6] [6], TAS-B[7] [13] which is a fine-tuned version of BERT, and GPT-3 embeddings[8] [27]
- **Zero-Shot.** Such methods use pre-trained LLMs which are not explicitly trained on this task, specifically pre-trained GPT-2[9] [30], OPT-1.3B[10] [45] and GPT-3 DaVinci[11] [3]. GPT-2 and OPT are used to rank options based on the log-likelihood that the query precedes the option. Since the GPT-3 API does not currently permit log-likelihood scoring of prespecified completions, GPT-3 is given the full list of options in the prompt and asked to choose the best option in the monolithic setting or provide scores for each option in the aspect-based setting (see Sections 5.2 and 5.3).
- **Few-Shot.** Few-shot methods extend zero-shot LLM methods by concatenating a fixed number of correct query-answer samples onto each input query. This is done to provide the LLM with added context on the task.

5.2 Monolithic Setting

In monolithic experiments, the full query is given as an input, where all preference aspects are provided in the initial NL context. In this setting, the model does not rely on any external knowledge of the underlying problem structure: it does not know *a priori* what the preference aspects are, nor that no preference aspect can be left unsatisfied. For the few-shot methods, the example template concatenated onto the input for GPT-2 and OPT was: "*input:* <sample query>, *output:* <sample correct option description>," since these LLMs evaluated the likelihood of one query-option pair at a time. Since GPT-3 prompts included the full list of options, the example template for GPT-3 was: "*Query:* <sample query>, *Options:* <sample option list>, *Option:* <sample correct option description>". Full prompt details are available in the code documentation.[4]

5.3 Aspect-based Setting

We also investigate simple methods for explicit aspect-level reasoning, aiming to establish baselines for future work on explainable and verifiable NL recommendation. Specifically, we sequentially provide one preference aspect at a time as an input to a model, after which we aggregate the output scores. For instance, for the first entry in Table 1, the model would first use "***meat lasagna***" as an input, then "***watching my weight***" as an input, and finally aggregate the results. This approach uses knowledge of the preference aspects and the problem structure (i.e. that all preference aspects must be satisfied) to force language models to reason about preference aspects disjointly, followed by an aggregation step. While this is a simple baseline, it is a step towards the study of how language models can be guided to perform more advanced forms of discrete reasoning with explicit aspects.

For a query q_i with $|\mathcal{P}_i| = N_i$ preference aspects, an individual aspect $j \in \mathcal{P}_i$ given to a model as an input results in option scores $\{s^1_{i,j}, ..., s^5_{i,j}\}$ where $s^l_{i,j}$ is the score for the l'th option o^l_i, and a higher score indicates a model is more confident an option is correct.[12] To produce a single score S^l_i for each option, outputs are aggregated aspect-wise using one of the following functions:

- **Min**: $\min_{j \in \{1,...,N_i\}} s^l_{i,j}$
- **Max**: $\max_{j \in \{1,...,N_i\}} s^l_{i,j}$

[6] BERT-110M: https://huggingface.co/bert-base-uncased
[7] TAS-B: https://huggingface.co/sebastian-hofstaetter/distilbert-dot-tas_b-b256-msmarco
[8] text-embedding-ada-002: https://platform.openai.com/docs/api-reference/embeddings/create
[9] GPT-2: https://huggingface.co/gpt2
[10] OPT-1.3B: https://huggingface.co/docs/transformers/model_doc/opt
[11] text-davinci-003: https://platform.openai.com/docs/models/gpt-3-5
[12] We refer to a general *output score* since scores have different meanings for different models, for instance TF-IDF score versus log-likelihood.

Table 4: Monolithic (full query) setting % accuracy ± 95% CIs.

Sparse	OWC	17.6 ± 1.6
	TFIDF	20.8 ± 2.7
	BM25	19.0 ± 4.2
Dense	BERT	18.6 ± 2.3
	TAS-B	31.2 ± 3.0
	GPT-3	54.0 ± 2.4
Zero-Shot	OPT	30.8 ± 2.9
	GPT-2	27.0 ± 5.6
	GPT-3	72.6 ± 3.7
Few-Shot	OPT	31.0 ± 4.3
	GPT-2	24.6 ± 5.5
	GPT-3	**83.4 ± 2.5**

- **Amean** (arithmetic mean): $\frac{1}{N_i}\sum_{j=1}^{N_i} s_{i,j}^{l}$
- **Gmean** (geometric mean): $\sqrt[N_i]{\prod_{j=1}^{N_i} s_{i,j}^{l}}$

One of the goals of studying these aggregation functions was to investigate whether aggregation by **min** or **Gmean** (which are strongly affected by small elements) would be enough to capture the requirement that all preference aspects must be satisfied. The rationale for emphasizing the smallest scores is that if a model correctly inferred that at least one preference aspect j was unsatisfied in an option o_i^l by producing a low score $s_{i,j}^l$, these aggregation functions would correctly assign a low total score for option o_i^l.

Since log-likelihoods could not be used for option scores for GPT-3, it was explicitly prompted to provide scores for each option. For the few-shot example templates, the GPT-2 and OPT examples followed the same format as in the monolithic case with the aspect replacing the query. For GPT-3, to help the model output scores in text, the few-shot examples used scores of 0 for all incorrect options and 1 for the correct option (see code[4] for format details). However, the few-shot approach in the aspect-based setting is limited by the fact that our dataset does not include aspect labels for incorrect options, even though the hard negative options are typically positive for at least one aspect. Thus, while the few-shot examples identify aspects in the correct option, they do not properly identify aspects in all options.

5.4 Experimental Details

The baseline methods were evaluated on the full dataset via 5-fold cross-validation over five randomized, independent 400/100 train/test splits. Accuracy was used as the metric for all experiments. For few-shot methods, five examples were randomly selected from the training set to use as part of the prompt, and performance was evaluated on the remaining 100 test samples. For all other methods (zero-shot, dense, sparse), performance was evaluated directly on the test set since these methods do not use the training data.

A second round of experiments investigated the effects of reasoning strategies. Since some strategies had very few examples (e.g., only 30 analogical queries) and query strategies are multi-label, the per-strategy performance is evaluated on a single fold of 5/495 train/test split, where the 5 training samples are used for the few-shot prompt. Since only one fold is used, no confidence intervals (CIs) are computed for the reasoning strategy experiments.

6 EXPERIMENTAL RESULTS

6.1 Monolithic Setting

The results for the accuracy of all baseline methods with 95% confidence intervals on the monolithic setting for the full corpus are shown in Table 4 and Figure 3a. Dense and zero-shot methods outperform sparse methods, which is expected since sparse methods focus on lexical overlap and often fail to capture semantic similarity. Our dataset explicitly avoided lexical overlap between the correct answer and the query, while allowing for exact term matches in incorrect options (designed to be hard negatives). The performance of the sparse methods was near-random selection (20%).

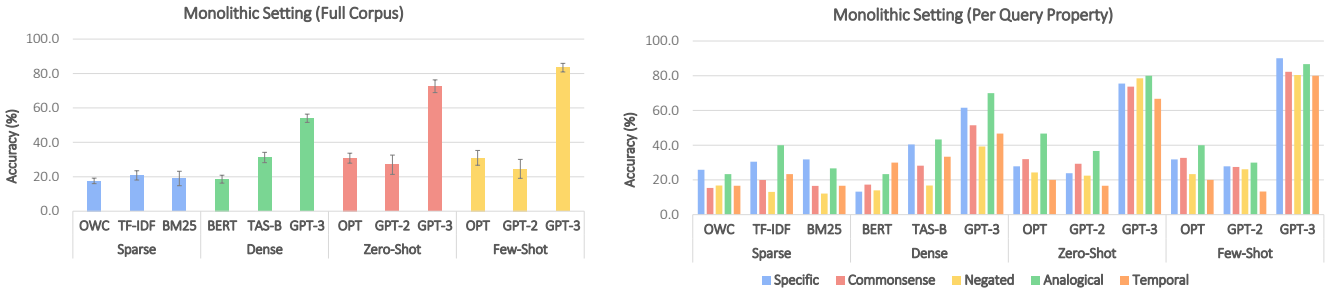
GPT-3 in the few-shot setting achieved the best overall performance, including compared to any aspect-based method (see Section 6.2), with 83.4% accuracy. This strong result is remarkably higher than sparse retrieval, and we interpret it as a validation of the quality of our data. Zero-shot GPT-3 gave the next best result with 72.6% accuracy, indicating that while GPT-3 benefited from examples, it also achieved strong performance without them. Dense retrieval using GPT-3 gave the third best result in the monolithic setting with 54.0% accuracy, suggesting that comparing the embedding similarity between options fails to capture part of the reasoning required, though it is still able to solve over half of the problems.

Three of the other LLMs, GPT-2, OPT, and TAS-B, all achieved similar performance to each other near 30% accuracy, while BERT, the oldest LLM tested, achieved 18.6% accuracy. These results show a clear increase in performance for more advanced generations of LLMs, indicating that our dataset succeeds in benchmarking improvements in LLM reasoning abilities. Interestingly, OPT and GPT-2 did not achieve higher accuracy in the few-shot setting. We conjecture that, unlike GPT-3, these models are not able to make necessary inferences about the problem structure from the examples.

In addition, we generate results for each reasoning strategy, shown in Table 5 and Figure 3b. All methods achieve above-average performance on Analogical queries relative to their performance across all strategies, and the highest overall performance was for the Specific category for GPT-3 at 90.1% accuracy. These results suggest that queries that use Analogical and Specific reasoning may be favorable to queries that use other strategies. In contrast, the worst performance of all zero-shot and few-shot methods was on the Temporal category, suggesting that LLMs struggle to make inferences that require temporal reasoning in our dataset.

6.2 Aspect-based Setting

Table 6 and Figure 4a show results from the aspect-based setting, where individual preference aspects are used as inputs and the output scores are then aggregated. Zero-shot GPT-3 with **Gmean** aggregation (the best aggregation function for this method) achieves 67.6% accuracy, which is comparable (within CI range) to monolithic zero-shot GPT-3, suggesting that the multi-aspect satisfaction structure of the problem is successfully captured by this explicit aspect-based reasoning method. However, while the use of few-shot examples led to performance improvement for GPT-3 in the monolithic setting, it led to a performance decrease for LLMs in the aspect-based setting. The likely cause is that, since our dataset



(a) Monolithic Setting: Full Corpus

(b) Monolithic Setting: Per Query Property

Figure 3: Experimental Results for Monolithic Setting

Table 5: Per reasoning strategy accuracy (%) for the monolithic setting.

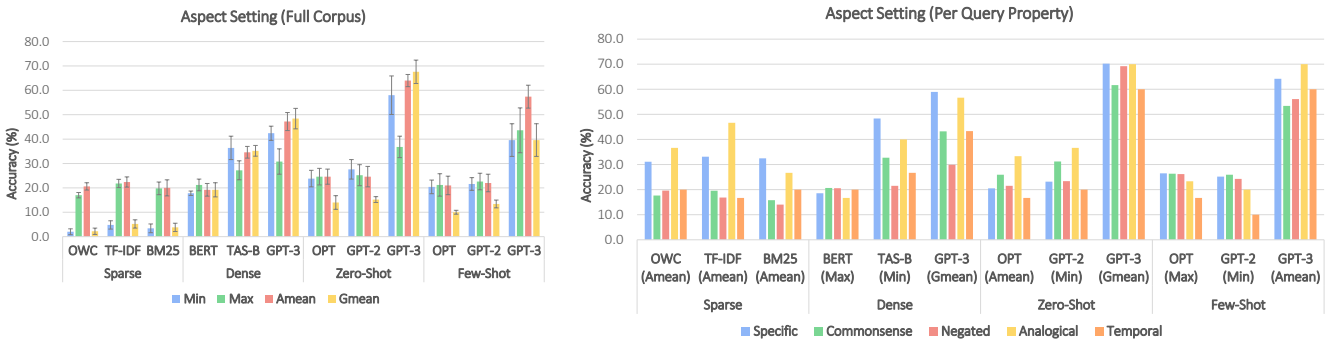
		Specific	Commonsense	Negated	Analogical	Temporal
Sparse	OWC	25.8	15.4	16.8	23.3	16.7
	TF-IDF	30.5	19.9	13.1	40.0	23.3
	BM25	31.8	16.5	12.2	26.7	16.7
Dense	BERT	13.3	17.3	14.0	23.3	30.0
	TAS-B	40.4	28.2	16.8	43.3	33.3
	GPT-3	61.6	51.5	39.3	70.0	46.7
Zero-Shot	OPT	27.8	32.0	24.3	46.7	20.0
	GPT-2	23.8	29.3	22.4	36.7	16.7
	GPT-3	75.5	73.7	78.5	80.0	66.7
Few-Shot	OPT	31.8	32.7	23.4	40.0	20.0
	GPT-2	27.8	27.4	26.2	30.0	13.3
	GPT-3	**90.1**	**82.3**	**80.4**	**86.7**	**80.0**

Table 6: Aspect-based setting accuracy (%) ± 95% CIs. The best aggregation function for each method is indicated in bold.

		Min	Max	Amean	Gmean
Sparse	OWC	2.0 ± 1.2	17.0 ± 1.1	**20.6 ± 1.5**	2.2 ± 1.3
	TF-IDF	4.8 ± 1.7	21.8 ± 1.7	**22.4 ± 2.1**	5.2 ±1.7
	BM25	3.4 ± 1.8	19.8 ± 2.6	**20.0 ± 3.3**	3.8 ± 1.7
Dense	BERT	17.8 ± 0.9	**21.2 ± 2.4**	19.2 ± 2.6	19.2 ± 2.9
	TAS-B	**36.4 ± 4.8**	27.2 ± 3.9	34.6 ± 2.4	35.2 ± 2.2
	GPT-3	42.4 ± 2.9	30.8 ± 5.2	47.2 ± 3.7	**48.4 ± 4.2**
Zero-Shot	OPT	23.8 ± 3.4	24.6 ± 3.4	**24.6 ± 3.1**	14.0 ± 2.8
	GPT-2	**27.6 ± 4.0**	25.2 ± 4.3	24.6 ± 4.2	15.2 ± 1.2
	GPT-3	58.0 ± 7.9	36.8 ± 4.4	64.0 ± 2.5	**67.6 ± 4.8**
Few-Shot	OPT	20.4 ± 2.8	**21.2 ± 4.6**	21.0 ± 3.8	10.0 ± 0.8
	GPT-2	21.6 ± 2.6	**22.6 ± 3.4**	22.0 ± 3.6	13.4 ± 1.6
	GPT-3	39.6 ± 6.7	43.6 ± 9.2	**57.4 ± 4.7**	39.6 ± 6.7

includes aspect labels only for the correct options and not the incorrect options (see Section 5.3), the few-shot examples misguided the LLM by identifying aspect satisfaction in the correct option only. Though the monolithic and best aspect-based zero-shot GPT-3 results were comparable, the best performance on the dataset was achieved by few-shot monolithic GPT-3 due to its ability to benefit from examples.

For most sparse, dense, and zero-shot methods (except OWC and OPT), aspect-based performance with the best aggregation function was comparable to its monolithic counterpart (within CI ranges). However, since not all aggregation functions led to good performance, it is useful to speculate about possible limitations of our simple aspect-based approach. One possible limitation is that isolating a single aspect degrades the ability to benefit from the NL context in which the aspects occur. Another is that even if high scores are output for all options which satisfy individual aspects, simple aggregation functions may not be suitable for combining



(a) Aspect-Based Setting: Full Corpus

(b) Aspect-Based Setting: Per Query Property

Figure 4: Experimental Results for Aspect Setting

Table 7: Per reasoning strategy accuracy (%) for the aspect-based setting.

		Specific	Commonsense	Negated	Analogical	Temporal
Sparse	OWC (Amean)	31.1	17.7	19.6	36.7	20.0
	TF-IDF (Amean)	33.1	19.6	16.8	46.7	16.7
	BM25 (Amean)	32.5	15.8	14.0	26.7	20.0
Dense	BERT (Max)	18.5	20.7	20.6	16.7	20.0
	TAS-B (Min)	48.3	32.7	21.5	40.0	26.7
	GPT-3 (Gmean)	58.9	43.2	29.9	56.7	43.3
Zero-Shot	OPT (Amean)	20.5	25.9	21.5	33.3	16.7
	GPT-2 (Min)	23.2	31.2	23.4	36.7	20.0
	GPT-3 (Gmean)	**70.2**	**61.7**	**69.2**	**70.0**	**60.0**
Few-Shot	OPT (Max)	26.5	26.3	26.2	23.3	16.7
	GPT-2 (Min)	25.2	25.9	24.3	20.0	10.0
	GPT-3 (Amean)	64.2	53.4	56.1	**70.0**	**60.0**

these scores. Lastly, forcing the model to make multiple inferences per query exposes it to more points of failure.

We also investigate the effects of reasoning strategies on aspect-level reasoning. Per-strategy experiments were performed using the best aggregation function for each baseline, with results shown in Table 7 and Figure 4b. As in the monolithic setting, the best aspect-based method (zero-shot GPT-3 with **Gmean**) achieves the strongest results on Specific and Analogical queries with 70.2% and 70.0% accuracy, respectively, and its worst performance on the Temporal queries with 60.0% accuracy. Thus, prompts that are Specific and/or Analogical may be good choices for both monolithic and aspect-level reasoning settings, while performance on Temporal prompts should be a direction for future work.

7 CONCLUSION

Aiming to advance research into ConvRec, we introduce a novel manually-curated dataset of multi-aspect NL preference statements and NL item descriptions of both ground-truth true positive matches and hard negative mismatches. We specifically focus on the multiple-choice task of retrieving items that correctly match multi-aspect preferences stated in an NL query. Also, to provide explanations for recommendations, we explicitly annotate preference aspects, item aspects, and satisfaction relations between the two. As part of the dataset, we have released code to reproduce results for a diverse set of baselines in both a standard full-query (monolithic) setting and an aspect-based setting, the latter of which forces reasoning over isolated query aspects and aggregates the results. While the best results came from GPT-3 in the monolithic setting, our aspect-based GPT-3 baselines also performed well with a zero-shot accuracy near that of the monolithic setting (68% vs 73%, respectively). Overall, our dataset and baselines establish a foundation for further research into explicit multi-aspect NL reasoning, including research directions such as aspect-specific few-shot methods, aspect extraction and evaluation, matching multiaspect NL queries to FoodKG knowledge graph entities backing each option, and joint optimization for explainability and recommendation performance.

8 ACKNOWLEDGMENTS

We would like to thank Zhenwei Tang and Touqir Sajed for their valuable assistance. This work was supported by LG Electronics, Toronto AI Lab Grant Ref #2022-1473.

REFERENCES

[1] Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Armin Toroghi, Anton Korikov, Ali Pesaranghader, Touqir Sajed, Manasa Bharadwaj, Borislav Mavrin, and Scott Sanner. 2023. Self-supervised Contrastive BERT Fine-tuning for Fusion-Based Reviewed-Item Retrieval. In *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I*. Springer, 3–17.

[2] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[4] Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems* 31 (2018).

[5] Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044* (2019).

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[7] Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. ERASER: A benchmark to evaluate rationalized NLP models. *arXiv preprint arXiv:1911.03429* (2019).

[8] Zuohui Fu, Yikun Xian, Yaxin Zhu, Yongfeng Zhang, and Gerard de Melo. 2020. COOKIE: A dataset for conversational recommendation over knowledge graphs in e-commerce. *arXiv preprint arXiv:2008.09237* (2020).

[9] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126.

[10] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176* (2022).

[11] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415* (2019).

[12] Steven Haussmann, Oshani Seneviratne, Yu Chen, Yarden Ne'eman, James Codella, Ching-Hua Chen, Deborah L McGuinness, and Mohammed J Zaki. 2019. FoodKG: a semantics-driven knowledge graph for food recommendation. In *International Semantic Web Conference*. Springer, 146–162.

[13] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 113–122.

[14] Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2019. Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456* (2019).

[15] Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503* (2017).

[16] Changsung Kang, Xuanhui Wang, Yi Chang, and Belle Tseng. 2012. Learning to rank with multi-aspect relevance for vertical search. In *Proceedings of the fifth ACM international conference on Web search and data mining*. 453–462.

[17] Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 252–262. https://doi.org/10.18653/v1/N18-1023

[18] Weize Kong, Swaraj Khadanga, Cheng Li, Shaleen Kumar Gupta, Mingyang Zhang, Wensong Xu, and Michael Bendersky. 2022. Multi-Aspect Dense Retrieval. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 3178–3186.

[19] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155* (2016).

[20] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018)*.

[21] Shuokai Li, Ruobing Xie, Yongchun Zhu, Fuzhen Zhuang, Zhenwei Tang, Wayne Xin Zhao, and Qing He. 2022. Self-Supervised learning for Conversational Recommendation. *Information Processing & Management* 59, 6 (2022), 103067.

[22] Shuokai Li, Yongchun Zhu, Ruobing Xie, Zhenwei Tang, Zhao Zhang, Fuzhen Zhuang, Qing He, and Hui Xiong. 2023. Customized Conversational Recommender Systems. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II*. Springer, 740–756.

[23] Shengnan Lyu, Arpit Rana, Scott Sanner, and Mohamed Reda Bouadjenek. 2021. A workflow analysis of context-driven conversational recommendation. In *Proceedings of the Web Conference 2021*. 866–877.

[24] Bill MacCartney and Christopher D. Manning. 2008. Modeling Semantic Containment and Exclusion in Natural Language Inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*. Coling 2008 Organizing Committee, Manchester, UK, 521–528. https://aclanthology.org/C08-1066

[25] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence* 43, 1 (2019), 187–203.

[26] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*. 165–172.

[27] Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005* (2022).

[28] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155* (2022).

[29] Ethan Perez, Siddharth Karamcheti, Rob Fergus, Jason Weston, Douwe Kiela, and Kyunghyun Cho. 2019. Finding generalizable evidence by learning to convince Q&A models. *arXiv preprint arXiv:1909.05863* (2019).

[30] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[31] Filip Radlinski, Krisztian Balog, Bill Byrne, and Karthik Krishnamoorthi. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*. 353–360.

[32] Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain yourself! Leveraging language models for commonsense reasoning. *arXiv preprint arXiv:1906.02361* (2019).

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[34] Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval* 3, 4 (2009), 333–389.

[35] Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *Comput. Surveys* 55, 10 (2023), 1–45.

[36] Stuart J Russell. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.

[37] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press Cambridge.

[38] Sergey Volokhin, Joyce Ho, Oleg Rokhlenko, and Eugene Agichtein. 2021. You sound like someone who watches drama movies: Towards predicting movie preferences from conversational interactions. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3091–3096.

[39] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems* 32 (2019).

[40] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461* (2018).

[41] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903* (2022).

[42] Sanghyun Yi, Rahul Goel, Chandra Khatri, Alessandra Cervone, Tagyoung Chung, Behnam Hedayatnia, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tur. 2019. Towards coherent and engaging spoken dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015* (2019).

[43] Omar Zaidan and Jason Eisner. 2008. Modeling annotators: A generative approach to learning from annotator rationales. In *Proceedings of the 2008 conference on Empirical methods in natural language processing*. 31–40.

[44] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. 2022. On the paradox of learning to reason from data. *arXiv preprint arXiv:2205.11502* (2022).

[45] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. OPT:

Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068* (2022).

[46] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards topic-guided conversational recommender system. *arXiv preprint arXiv:2010.04125* (2020).