Debate Prompting for Query-driven Contrastive Explanation for Recommendations

by

George-Kirollos Yousry Guirguis Youssef Saad

A thesis submitted in conformity with the requirements
for the degree of Master of Applied Science

Department of Mechanical & Industrial Engineering
University of Toronto

Debate Prompting for Query-driven Contrastive Explanation for Recommendations

George-Kirollos Yousry Guirguis Youssef Saad
Master of Applied Science
Department of Mechanical & Industrial Engineering
University of Toronto
2025

Abstract

The rapid expansion of natural language processing (NLP) has increased interest in providing grounded explanations for recommendations, particularly in contrastive, query-driven contexts. This work investigates methods for generating explanations that use both user reviews and objective data sources. Inspired by existing literature and discourse theory, we explore various design choices to enhance the quality of these explanations. We introduce a debate-style prompting approach to generate contrastive explanations for recommendations. We find that debate-style prompting outperformed STRUM-LLM-like variants for contrastive summarization on two datasets, and for another dataset, either matches or beats the baselines on all criteria. We also examine whether the level of discourse aggressiveness influences explanation quality, finding no significant improvement. Additionally, we assess the reliability of large language models (LLMs) for pairwise Win Rate evaluation through a user study, showing that LLM-based pairwise evaluations align well with human judgment and can thus be considered trustworthy.

Acknowledgements

I would like to thank my supervisor, Dr. Scott Sanner, my family, and my friends for their support throughout the development of this thesis.

Contents

List of Tables

List of Figures

List of Listings

Chapter 1

Introduction

1.1 Introduction

In our daily lives, making decisions and comparing options are fundamental activities. Whether it's deciding between travel destinations, comparing products on Amazon, or choosing a restaurant for dinner, we are constantly faced with tough choices amid an overwhelming amount of information [23,24]. The abundance of rich information sources, such as user-generated reviews on platforms like TripAdvisor, Yelp, and Amazon, as well as objective data from resources like Wikipedia, WikiVoyage, or travel guides, offers a wealth of knowledge [64]. However, parsing through all of these sources to extract the information that is most relevant to our needs is often a difficult and exhausting process.

With the advancement of large language models (LLMs) and sophisticated search techniques, there are now powerful methods to extract and summarize relevant content efficiently [5, 10, 11, 69]. These models help transform the way we approach information, making it possible to condense vast amounts of text into more digestible insights. Despite these advancements, a challenge still remains: how can we further optimize this information for a specific query and identify what truly matters in a comparison?

This question is particularly significant when we consider comparative decision-making. People do not only want the best recommendation based on a query—they also want to understand why it is the best, and they want these comparisons to be grounded in real data and insights [41, 47, 55]. Consider a scenario where someone is comparing hotels for an upcoming trip, with the following query:

"I want a place with a sofa, a desk, bathrobes, private bathrooms, and family rooms."

A system might respond with a list of hotels, but that doesn't give the insights users want. Reviews on TripAdvisor provide subjective user experiences about hotels that may be of interest; however, they often come in a large abundance, making it difficult for users to extract meaningful insights. A system that can summarize these reviews, instead of just showing them all, is needed to help users understand the key points without being overwhelmed by the volume of information [25, 26]. So we propose a debate-style prompting methodology, using reviews or other rich data to give better comparative outputs. The input of this system would be a query and the system would source relevant entities and content from reviews and objective data sources. Upon completion of the pipeline, the output would be as seen in Figure 1.1.

Query: I want a place with a sofa, a desk, bathrobes, private bathrooms, and family rooms

	The Ivy At Verity	Town Inn Suites
desk	(1) Combines luxury and practicality with large rooms, beautifully decorated with seating and working setups [26] (2) Attention to detail ensures relaxed ambiance with a writing desk for guest comfort [39] (3) Desks are part of a luxuriously comfortable furnishing setup, though can distract if seeking basic functionality [36]	(1) Offers a straightforward setup with a desk and necessary amenities, focusing on function [1] (2) Rooms are spacious and somewhat functional, meeting essential needs without opulence [17] (3) Encourages work-life balance with nicely arranged living room, catering to pragmatic guests [28]
sofa	(1) Offers a luxurious seating arrangement with upholstered chairs, overstuffed armchair, and ottoman for comfort [1] (2) Features a cozy environment with seating areas to recline and relax, including a sunny deck [24] (3) Decor includes lovely upholstered seating, exposed brick walls, and beautiful wallpapers enhancing visual appeal [39]	(1) Focuses on practicality with spacious setups including a large kitchen and living space [22] (2) Rooms are clean and quiet, providing reliability and consistency in the guest experience [10] (3) Spacious living room with facilities like dishwasher, promoting a functional homely atmosphere [34].
bathrobes	(1) Described as plush and soft, offering a luxurious experience [44] (2) Accompanied by Bulgari amenities adding to a luxurious feel [49] (3) Heated floors in the bathrooms complement the plush bathrobes, enhancing comfort [41]	(1) Lacks specific mentions of bathrobes, focusing on practicality over luxury [1] (2) Rooms emphasize essential functionality with descriptions like 'comfortable, clean, and cozy apartment' [42] (3) Guarantees essentials with services like daily housekeeping, ensuring hassle-free convenience [50]

Figure 1.1: Example of good comparative output for the query

This research aims to address this gap by optimizing comparative decision-making, leveraging rich data sources and LLM capabilities to provide clearer, more grounded comparisons.

1.2 Research Challenges

In this work, we aim to address several key research challenges related to the generation and evaluation of query-driven contrastive explanations. Existing methods often focus on single entity summarization, are not query-driven, or are not sufficiently contrastive [2, 3, 23, 24, 31]. These challenges are formalized into the following research questions (RQs):

- **RQ1: Does debate-style prompting improve query-driven contrastive explanations?**
  Discourse theory is explicitly contrastive and shares many of the same principles of good contrastive explanation [14, 22, 46]. We hypothesize that debate, which is grounded in discourse theory, and has been specifically established to compare and contrast two opposing positions or options, should naturally perform well in this area. Furthermore, since LLMs have likely observed human debates in their training data, we conjecture that they already have a natural aptitude for debate-style reasoning. Therefore, we aim to investigate whether using debate-style prompting, where the model takes contrasting positions, leads to improved quality of query-driven contrastive explanations compared to existing prompting methods. The evaluation will be based on clearly defined criteria in the literature for contrastive explanation. The goal is to determine if debate-style prompting effectively enhances contrast and relevance, among other metrics, as defined by existing literature.

- **RQ2: Does the aggressiveness in debate-style prompting impact the quality of contrastive explanations?**
  We conjecture that a more assertive debate style may lead to stronger debates and therefore, we explore how the level of assertiveness or aggressiveness in the debate impacts the resulting

explanations. Specifically, we examine whether being nicer or assertively confrontational in approach yields better contrast by highlighting the pros and cons of each entity more effectively, or if it reduces the quality of the explanations.

- **RQ3: Can we trust pairwise LLM Win Rate evaluation of query-driven contrastive explanations?**
  LLMs have been widely used for automated evaluation of natural language outputs in a variety of settings, such as AlpacaEval [18, 19] and G-Eval [39]. Hence, we believe that they should provide reliable automatic evaluations that would otherwise be prohibitively expense to run with our scale of experimentation. Nonetheless, to check this hypothesis that automated LLM evaluation aligns with human judgment, we aim to carry out a user study to see whether LLMs can accurately assess the quality of two contrasting explanations in a manner that aligns with human judgment.

1.3 Summary of Contributions

In this work, we provide the following contributions to address the research challenges:

- **RQ1: Introduction of a novel debate-style prompting methodology for contrastive explanation.**
  We introduce a debate-style prompting architecture that allows the model to take contrasting positions on a given query. This methodology is designed to improve the quality of query-driven contrastive explanations by fostering a deeper exploration of different perspectives and enhancing contrast, as evaluated by well-established criteria in the literature. The experimental results demonstrate that debate-style prompting outperforms existing baselines on key criteria such as contrast, diversity, and usefulness, with win rates significantly higher for debate-style prompting compared to traditional summarization methods.

- **RQ2: Evaluation of the impact of aggressiveness in debate-style prompting for contrastive explanation.**
  We analyze the effect of varying levels of aggressiveness in debate-style prompting on the quality of contrastive explanations. We design an experiment to compare standard, aggressive, and nice variations of debate-style prompts. Therefore, we explore how the level of assertiveness or aggressiveness in the debate impacts the resulting explanations. The results indicate that the aggressive prompt generally does not outperform the standard version on all criteria on most datasets, with a slight advantage for aggressiveness on the Hotels dataset. These findings suggest that prompt aggressiveness does not significantly influence the effectiveness of contrastive explanations.

- **RQ3: Validation of the reliability of LLM-based Win Rate evaluation for contrastive explanations.**
  We validate the reliability of pairwise LLM-based evaluations of query-driven contrastive explanations through a user study. We compare LLM evaluations with human judgments using Cohen's Kappa and Fleiss' Kappa as metrics for agreement. The results show that LLM evaluations align well with human evaluations, often outperforming the agreement of human

evaluators amongst themselves. This demonstrates that LLM-based evaluation can be a reliable approach for assessing the quality of contrastive explanations.

1.4 Outline

The thesis is comprised of the following chapters:

Chapter 2: Background This chapter reviews the existing literature and approaches relevant to query-driven contrastive summarization. We highlight the strengths and limitations of current methods and identify areas for improvement.

Chapter 3: Prompting Methodology Here, we discuss the overall architecture and key design decisions of our approach. We compare our methodology to existing baselines, emphasizing the unique aspects and innovations in our techniques.

Chapter 4: Datasets This chapter introduces the three datasets used in our experiments: the TravelDest dataset, which contains objective information from WikiVoyage, and two datasets featuring subjective reviews scraped from TripAdvisor.

Chapter 5: Experiments and Evaluation We describe the experimental designs used to address the three research questions. This chapter provides an in-depth analysis of the results and insights gained from these experiments.

Chapter 6: Conclusion Finally, we summarize the contributions and findings of the thesis, and suggest potential directions for future research.

Chapter 2

Background

2.1 LLMs & Prompt Engineering

The introduction of Transformer architectures by Vaswani et al. [58] marked a pivotal shift in natural language processing (NLP), heralding the emergence of large language models (LLMs) that have since redefined the field [58]. Models such as OpenAI's Generative Pre-trained Transformer (GPT) series [6] and Meta's LLaMA [56] exemplify this new paradigm, employing Transformers to achieve state-of-the-art performance across a broad spectrum of NLP tasks.

LLMs, characterized by their ability to process and generate human-like text, are trained on massive datasets through unsupervised learning. These models utilize the self-attention mechanism introduced by Transformers, enabling them to capture intricate linguistic dependencies and produce highly coherent text [12]. The pre-training phase allows these models to develop a contextual understanding of language, which is then refined through fine-tuning for specific applications, making them effective for a diverse array of tasks, from machine translation to conversational AI [48].

2.1.1 Transformers

As previously introduced, Transformers have being the core piece behind the introduction of LLMs [58]. The core of the Transformer model is the attention mechanism, which provides a way to measure how words in a text sequence relate to each other. It provides an advantage over previous language modeling works, in that the attention mechanism is inherently parallelized and thus handles sequences in parallel rather than sequentially which long-short term memory models (LSTMs) do [28].

The Transformer architecture consists of an encoder-decoder framework, both of which are built upon stacks of multiple identical layers. The encoder contains six identical layers, each comprised of two primary components: a multi-head self-attention mechanism, which allows each position in the input to attend to every other position, and a position-wise fully connected feed-forward network. A residual connection wraps around each of these sub-layers, followed by layer normalization.

2.1.1.1 Decoder & Prompting

The decoder mirrors the encoder's architecture with six identical layers, but with an additional third sub-layer for multi-head attention over the encoder's output [58]. This allows the decoder to incorporate context from the encoder during generation. The decoder also includes a modified multi-headed self-attention mechanism, which uses a masking mechanism that ensures that each position in the sequence only depends on the preceding positions, thereby maintaining causality in the generated outputs. Residual connections and layer normalization are employed in a similar manner as in the encoder.

During inference, the decoder generates the output sequence autoregressively, meaning it produces one token at a time [58]. At each step, the decoder receives the previously generated tokens, where they pass through the masked multi-headed self-attention. Then, the decoder attends to the encoder's output through the encoder-decoder attention mechanism. Following this, the output is passed through a feed-forward network and the next token is predicted. The output representation is projected onto the vocabulary space, typically followed by a softmax function to obtain a probability distribution over possible next tokens. The token with the highest probability is selected as the next token in the sequence. This process repeats until a special end-of-sequence token is generated or a predefined maximum length is reached.

For the case of text generation, we often use decoder-only models, such as GPT-4 [44] and LLaMA-3 [17, 58]. In these models, since there is no encoder, there is also no cross-attention mechanism. A high level overview of prompting a decoder-only LLM architecture is provided in Figure 2.1.

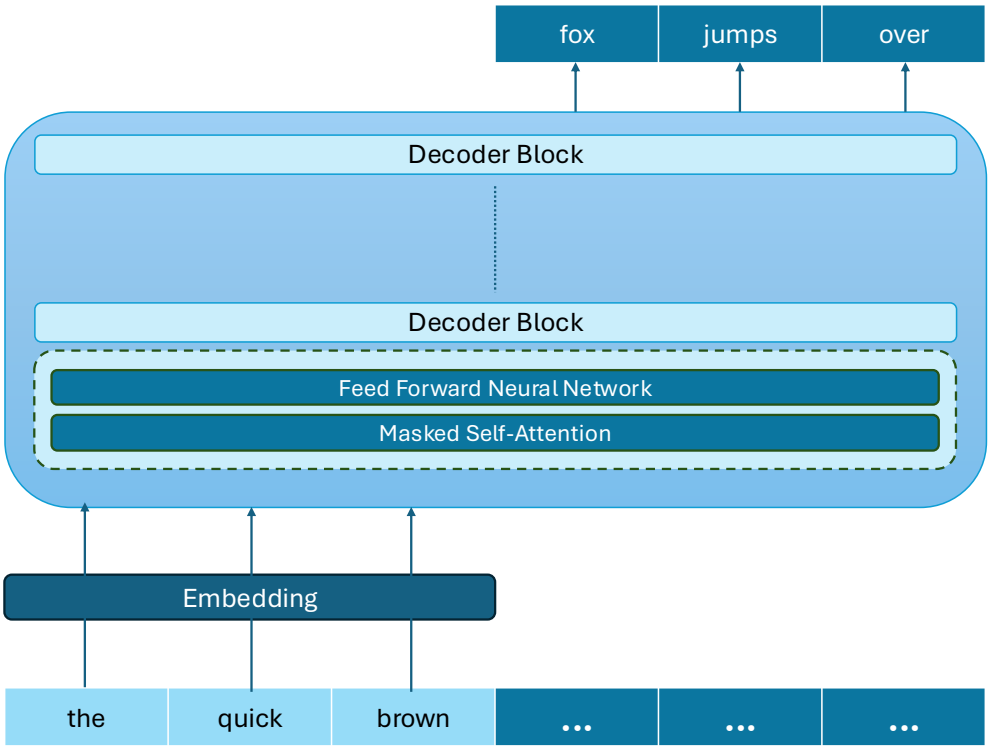

Figure 2.1: Prompting Decoder-only Architecture

2.1.2 Prompt Engineering

Prompt engineering plays a crucial role in optimizing the performance of LLMs by strategically designing prompts to influence model outputs, encouraging reasoning, reducing errors, and enhancing reliability [50]. This section summarizes key approaches in prompt engineering, discussing their common goals, methodologies, and limitations.

2.1.2.1 Zero-, One- & Few-shot Learning

With the introduction of GPT-3, three general forms of prompting are introduced: zero-, one- and few-shot [6, 38].

Zero-shot learning involves providing no examples to an LLM and only providing natural language instructions [6, 38]. An example of a zero-shot prompt is as follows:

> *Translate this text from English to French: how are you today?*

One-shot learning involves providing the instructions and a single example to the LLM [6, 38].

Few-shot learning is a method of applying in-context learning (ICL) to allow the LLM to learn to complete complex tasks just from examples provided in the context [15, 38]. It usually performs better than zero- and one-shot learning, especially with high parameter count models like GPT-3 [6]. An example of a few-shot prompt is as follows:

> *Translate the text from English to French.*
>
> *English: My name is Rob.*
> *French: Je m'appelle Rob.*
>
> *English: The weather is nice.*
> *French: Il fait beau.*
>
> *English: It is raining.*
> *French: Il pleut.*
>
> *English: How are you?*
> *French:*

2.1.2.2 Structured Reasoning Techniques

Prompt engineering techniques that enhance model reasoning often employ structured approaches to guide the LLM through complex problem-solving tasks. Techniques such as Chain-of-Thought (CoT), Tree-of-Thought (ToT), and Graph-of-Thought (GoT) prompting all build on the principle of breaking down problems into smaller, more manageable steps [63, 66, 68]. These methods enable LLMs to produce more structured and coherent outputs.

Chain-of-Thought (CoT) prompting is a few-shot prompting methodology that provides examples in the prompt where a problem is decomposed into intermediate steps that are solved before reaching a final answer [63]. An example of CoT is as follows:

> **Q:** Rob has 5 oranges. He buys half a crate's worth of oranges. A crate has 12 oranges. How many oranges does he have?
>
> **A:** Rob starts with 5 oranges. Half a crate of 12 oranges is 6 oranges. 5+6 = 11. So Rob has 11 oranges.

Variations like Automatic CoT and LogiCoT further refine this approach by either automating the reasoning steps or providing explicit logical validation [72, 73].

Chain-of-Symbol (CoS) and other symbol based techniques offer an alternative by employing condensed symbols, which can improve spatial reasoning and human interpretability, albeit with scalability challenges and integration challenges with other reasoning techniques [30].

Chain-of-Table prompting addresses the challenges of reasoning with tables by using a step-by-step tabular reasoning approach [62]. This method allows LLMs to perform logical reasoning on tables using SQL/DataFrame operations, which improves model performance on benchmark tabular datasets.

Overall, structured reasoning techniques aim to enhance the logical coherence and reliability of the model's responses.

2.1.2.3 Fact-Verification and Retrieval Techniques

Another major area of prompt engineering involves techniques that focus on ensuring factual correctness and reducing the risk of generating misleading information, known as hallucination. Methods such as Retrieval-Augmented Generation (RAG), ReAct prompting, and Chain-of-Verification (CoVe) incorporate various strategies to ground the generated text in verifiable facts [13, 34, 67, 70]. RAG integrates retrieval from external knowledge bases to add context, while ReAct combines reasoning with real-time interactions, such as querying external APIs, to verify information. CoVe, on the other hand, encourages explicit verification steps within the model's reasoning process.

Retreival-Augmented Generations (RAG) is a methodology where a retriever is used to find relevant text documents based on a given query. These retrieved documents are then provided as extra context along with the query to a generator, which creates a response or output [34]. Typically, the retrieval methodology used here is dense retrieval, which uses the similarity of neural embeddings of passages and a query to rank relevant passages for the given query [32]. Dense retrieval is outlined in more detail in Section 2.2.1.2. A diagram showing the architecture of a typical RAG system is provided in Figure 2.2.

Chain-of-Note (CoN) evaluates document relevancy and filters out unreliable content to ensure only pertinent information is used. Chain-of-Knowledge (CoK) takes this a step further by dynamically adapting information from both internal and external sources to ground complex reasoning in reliable content [35].

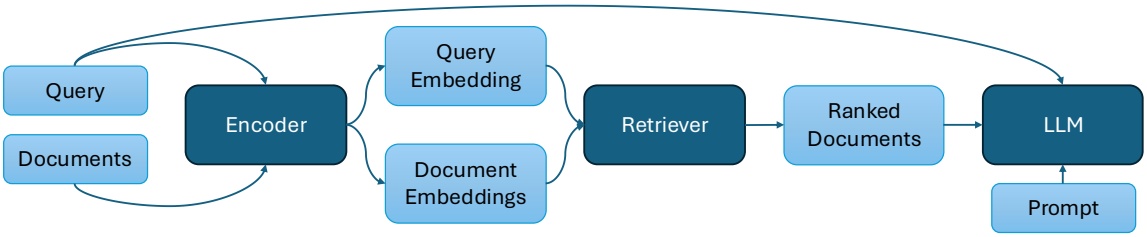

Figure 2.2: RAG Architecture

These fact-verification and retrieval techniques help improve the factual reliability of LLM outputs.

2.1.2.4 Challenges and Limitations in Contrastive Summarization

Despite the significant advances brought about by existing prompt engineering methods, their limitations become apparent when dealing with more nuanced reasoning tasks, such as contrastive explanations or debate-like scenarios. Current techniques excel at guiding LLMs to produce coherent, step-by-step responses or verify factual claims; however, they lack mechanisms for self-contrasting multiple perspectives or synthesizing arguments in a debate format. Methods such as CoT and RAG provide a solid foundation for systematic reasoning, yet they do not explicitly support the development of arguments that involve understanding opposing views or evaluating competing claims. In this thesis, we aim to extend prompting-based methodologies to address such challenges.

2.2 Natural Language Explanations

Text-based explanation in NLP involves generating explanations that help users understand content by providing relevant details in a more accessible form. In the context of recommendation, we have the following example within the domain of travel destinations:

> **Query:** nice relaxing beach vacation
> **Recommendation:** Cancun
> **Explanation:** Cancun is well-known for it's full service beach resorts and relaxed atmosphere.

Broadly, text-based explanation systems for recommendation rely on one of three techniques: extractive, abstractive, or hybrid approaches [4]. Traditional approaches primarily relied on extractive methods, which selected key segments directly from the source text. However, as NLP technologies have evolved, there has been a shift towards abstractive and contrastive summarization approaches, particularly with the development of LLMs. These newer models incorporate advanced architectures, enabling richer and more adaptable summaries that can cater to diverse user needs and queries. This review discusses major works in the domain, categorizing them into information retrieval (IR) and LLM-based approaches, and provides an analysis of their methodologies, applications, and limitations.

Information retrieval (IR) and large language model (LLM) approaches represent two major methods for generating explanations:

- **Information Retrieval (IR) Approaches**: These methods use sparse or dense retrieval techniques to identify relevant information from a dataset based on user queries. Sparse retrieval focuses on exact matches, while dense retrieval captures semantic relevance, each with its own advantages and limitations.
- **LLM-based Approaches**: These methods involve prompting LLMs to generate explanations that are either extractive, abstractive, or a hybrid, offering more adaptable and user-specific summaries compared to traditional IR methods. LLMs enable more sophisticated reasoning and richer explanations, although they may also introduce challenges related to hallucination or inconsistency.

2.2.1 Information Retrieval (IR)-based Approaches

Information retrieval-based approaches for generating explanations can be divided into sparse and dense retrieval methods, and further examined based on the data sources used for explanations. These approaches are primarily used for query-driven explanations, where the goal is to provide relevant information based on user queries by retrieving content from large collections of documents.

2.2.1.1 Sparse Retrieval

Sparse retrieval methods rely on traditional term-matching algorithms such as TF-IDF or BM25. These methods are effective for retrieving documents or segments that contain exact matches to user queries, making them suitable for scenarios where precision is critical [49,51,53]. Another prominent method utilizes universal sentence encoders and TF-IDF to produce aspect-oriented summaries of news articles, based on manually specified intents [1]. The main advantage of sparse retrieval is its interpretability, as it allows for clear tracing of retrieved results back to the original query terms. However, these methods often struggle with capturing semantic similarities, especially in cases where synonyms or more abstract relationships are involved.

2.2.1.2 Dense Retrieval

Dense retrieval methods utilize embeddings generated by models like BERT to capture semantic meaning, allowing for better retrieval of content that may not explicitly match the query terms but is semantically relevant [1, 32, 53]. Dense retrieval has the advantage of being able to capture nuanced relationships and contextual similarities between user queries and content, providing more comprehensive results. However, it can be less interpretable compared to sparse retrieval, as the embeddings and similarity scores are not directly linked to the original query terms.

2.2.1.3 Data Sources for Explanations

Information retrieval methods for generating explanations can draw from different data sources, such as item description snippets or user reviews. Each data source offers distinct advantages and challenges.

Informational & Factual Content Information and factual content, often drawn from product descriptions, news articles, or other objective content, provides factual and concise information [1,64].

These data sources are beneficial for providing objective and consistent explanations, but they are limited in their expressiveness and may lack details that users find helpful in understanding real-world usage.

User-Generated Reviews User reviews offer rich, expressive content that captures experiential details often missing from more objective data sources [3, 33, 59]. They provide insights into user opinions, preferences, and experiences, which can make explanations more relatable and informative. However, reviews introduce challenges such as subjectivity and disagreement among users, which can complicate the generation of consistent and reliable explanations.

Hybrid Additionally, a hybrid approach can be used to retrieve information from the web, which may include a combination of informational/factual content and user-generated reviews [23, 24, 53].

2.2.2 LLM-based Approaches

Recent advancements in natural language processing have led to the use of LLMs for generating explanations.

2.2.2.1 Grouped Explanation Methods

LLM-based explanation methods can be grouped into extractive, abstractive, and hybrid techniques, each with specific strengths and limitations. Extractive methods, such as Vector Quantized Variational Autoencoders (VQ-VAE), STRUM, and STRUM-LLM, rely on extracting relevant segments from the source content to provide explanations [3, 23, 24]. VQ-VAE encodes sentences into discrete latent codes and clusters them to generate summaries, while STRUM and STRUM-LLM focus on extracting data along specified attributes, providing structured, contrastive summaries across different entities. These approaches work well for summarizing opinions efficiently but may not be as effective at providing contrastive, query-driven explanations, which limits their adaptability in more dynamic and user-specific scenarios.

Abstractive methods include Aspect-Controllable Opinion Summarization and Comparative Opinion Summarization via Collaborative Decoding, both of which use LLMs to generate summaries based on either user-specified aspects or token-level contrasts [2, 31]. These methods are more flexible in generating diverse explanations but may struggle with generalization, particularly when relying on synthetic datasets and lack attribute generation or extraction steps.

2.2.2.2 STRUM & STRUM-LLM

STRUM-LLM [24] and STRUM [23] are the closest existing methods for achieving LLM-based contrastive summarization on a set of aspects that apply to two entities. STRUM-LLM takes an extractive approach by extracting data along identified attributes, providing source attributions, and highlighting high-contrast, important attributes. This architecture uses multiple LLM components to handle extraction, merging, and evaluation of attributes and values. The architecture for this can be found in Figure 2.3. More specifically, each component uses an LLM and has the following purposes:

- **Aspect Extraction**: Extracts aspects and relevant values from provided sources, while maintaining an attribution to the original source.
- **Aspect Merge**: Merges similar aspects together into a single aspect to avoid redundancy.
- **Value Merge**: Takes each aspect and merges together similar values for the same aspect that is consistent with the majority opinion.
- **Contrastive Summarizer**: Identifies the most important and contrasting aspects and their respective values.
- **Usefulness**: This stage removes aspects that are found to be not useful and helps catch errors.

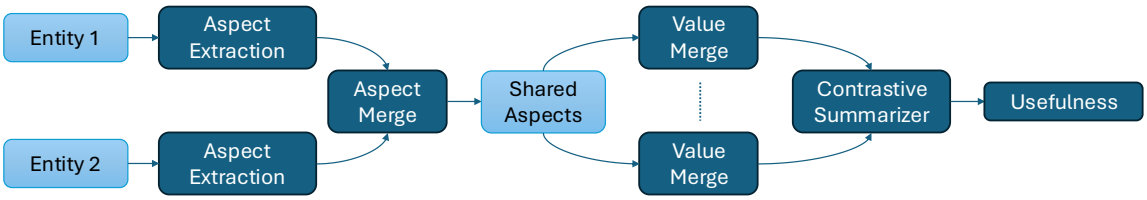

Figure 2.3: STRUM-LLM Architecture

While comprehensive, STRUM-LLM faces challenges in terms of scalability when applied to extremely large or diverse datasets, as the computational cost of using multiple LLM components can be prohibitive. Furthermore, its reliance on LLMs to judge relevant aspects without a user query can lead to biases that deviate from user intent. STRUM, on the other hand, uses entailment models and hierarchical clustering to extract, merge, and contrast aspects. However, the reliance on these models makes it computationally expensive and difficult to scale effectively with increased diversity in input data, which limits its practicality.

2.2.2.3 Summary

In summary, LLM-based explanation methods provide a range of tools for extractive and abstractive summarization, each with unique advantages and challenges. While extractive methods like STRUM-LLM and STRUM offer structured and attributed summarization, they are not query-driven and may not align with user intents. Additionally, STRUM and STRUM-LLM are purely extractive and are not optimized for contrastive summarization tailored for recommendations, which limits their effectiveness. Conversely, abstractive approaches such as Aspect-Controllable Opinion Summarization and Comparative Opinion Summarization via Collaborative Decoding provide more flexibility but can struggle with specificity and structure in comparative contexts.

2.2.3 Evaluation

Evaluating LLM-generated outputs is essential to ensure quality and alignment with user expectations. This section explores the key criteria used to evaluate these outputs, how LLMs can serve as evaluators, and the effectiveness of using pairwise comparisons for evaluation.

2.2.3.1 Criteria

Generally, users prefer having LLM-generated explanations for recommendations over baseline explanation generation methods [20, 41, 57] Specifically, users look for the following criteria:

Grounded/Correct: Explanations must be reliable and grounded in accurate information, often ensured by using extractions and citations. Correctness is a critical evaluative criterion that ensures reliability and factual accuracy, as highlighted in several studies [7, 20, 21, 26, 43, 47, 55].

Contrastive: Users value explanations that are contrastive, showcasing relative pros and cons [7, 43].

Relevancy: Explanations must be relevant to users' specific needs and contexts [7, 43, 57].

Diversity: Evaluative criteria also include diversity, ensuring that explanations are not repetitive and cover multiple facets of a given topic [7, 21, 43].

Usefulness: The quality of being useful is emphasized in evaluating explanations. The explanation must address user needs in a way that is helpful and informative [21, 41].

Effectiveness: Effective explanations are those that successfully communicate the intended information to help users make a decision. Higher perceived explanation quality has been found to correlate directly with increased user trust and transparency [26, 41].

2.2.3.2 Methodologies

Evaluation in the context of LLMs output and explanations requires effective methodologies that align well with human judgment.

LLMs as Evaluators: Advances in LLM evaluation methodologies highlight the value of GPT-4 and similar models in aligning with human evaluative judgments [39, 40, 65]. They have increasingly been used in the field of NLP to evaluate generated text [9, 54, 60, 74]. LLMs have demonstrated good alignment with human judgments, making them suitable for subjective assessment tasks.

Pairwise Evaluation: Particularly when making pairwise comparisons, LLMs have shown strong alignment with human judgments, making them suitable for complex evaluation tasks [40, 61]. We take the output from 2 methodologies and provide them to an LLM, asking it to select a winner. The wins over a series of experiments are summed to determine an overall Win Rate for each of the methodologies. The following is an example prompt for pairwise evaluation:

> **Query:** I want a restaurant downtown with a view of the water
>
> **Explanation A:** Canoe offers an exceptional dining experience, celebrated for having 'some of the best food we have ever had'.
>
> **Explanation B:** Canoe is great choice for a special occasion.

> *Your role is to evaluate Explanation A and Explanation B as being good explanations for restaurant recommendation for the criterion of relevancy. You should select either "A" or "B" as the winner. The summarizations provided should be relevant to each aspect and query provided.*

Liu et al. (2024) propose that pairwise preference as an evaluative approach is particularly effective for complex tasks where direct scoring and G-Eval [39] may not align as closely with human evaluators. This approach allows for more nuanced comparisons between different outputs, helping to determine which explanation or response is superior.

2.2.3.3 Reference-based Methods

There exists other evaluation methods that rely on n-gram co-occurrences with gold standard references, such as BLEU and ROUGE [8,37,45]. These methods often complement each other as BLEU is focused on precision, whereas ROUGE provides a better measure of recall. BERTScore builds upon these n-gram matching methodologies but rather than using exact word matches, it computes a BERT embedding similarity score (using the dot product of embeddings) between tokens in the model outputs, denoted as x, and tokens in the gold standard references, denoted as $\hat{x}$ [71]. The equations for recall and precision, respectively, using BERTScore are as follows:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \boldsymbol{x_i}^\intercal \hat{\boldsymbol{x}}_{\boldsymbol{j}}, P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_i \in \hat{x}} \max_{x_j \in x} \boldsymbol{x_i}^\intercal \hat{\boldsymbol{x}}_{\boldsymbol{j}} \tag{2.1}$$

Furthermore, these reference-based methods require the creation of gold standard references which is often not available and expensive to produce. They also often fail to capture the nuanced quality of generated content and are not as effective for evaluating the effectiveness of recommendations [8].

2.3 Discourse Theory

Discourse theory can help us understand and produce contrastive explanations. Understanding how people process persuasive messages is essential to developing effective discourse. Various theories in discourse analysis offer insights into how individuals engage with information, particularly when they are motivated to make informed decisions. Discourse theory can be viewed through several influential frameworks, both in persuasive contexts and in looking at how effective communication is structured.

2.3.1 Elaboration Likelihood Model (ELM)

Discourse theory in persuasive contexts can be analyzed through several influential frameworks that explore how effective communication is structured. One such approach is the Elaboration Likelihood Model (ELM), introduced by Petty and Cacioppo in 1986 [46]. ELM differentiates between two routes to persuasion:

- **Central Route**: involves careful deliberation, where individuals critically evaluate arguments and assess their merits in detail.

- **Peripheral Route**: relies on less substantive cues, such as the perceived credibility or attractiveness of a source, rather than the strength of the argument itself.

This model is particularly relevant in discourse where users have a personal stake, such as when responding to specific queries [14]. In such situations, individuals are more inclined to engage through the central route, thoroughly considering both positive and negative aspects of the arguments presented. This depth of elaboration often results in attitudes that are more stable, resistant to opposing arguments, and predictive of future behaviors consistent with those attitudes. Thus, persuasive discourse that fosters critical thinking and encourages users to deliberate on the pros and cons is likely to be more impactful.

2.3.2 Grice's Maxims

Another fundamental perspective in discourse theory is Grice's theory of conversation, which provides a framework for understanding how effective communication should ideally function [22]. Grice identified four conversational maxims:

- **Quantity**: advocates for being as informative as possible without being excessive.
- **Quality**: advocates for providing evidence-backed, true information.
- **Relation**: advocates for providing information relevant to the discussion at hand.
- **Manner**: advocates for providing organized, clear and succinct information.

By adhering to these principles, discourse can become more coherent and persuasive, encouraging deeper engagement and understanding among participants.

Chapter 3

Prompting Methodology

In this section, we outline the proposed debate-style prompting architecture methodology for query-based contrastive explanation. Our debate methodology takes inspiration from research in discourse theory as outlined in Section 2.3. This methodology builds on previous work as well, drawing inspiration from STRUM-LLM [24], and introduces a structured approach for effectively leveraging LLMs to generate contrastive explanations.

Problem Definition: We are given (a) a corpus of one or more reviews/objective data source extractions for each item and (b) a user-provided query. Our goal is to produce an aspect-based explanation containing exactly three short, extractive sentence descriptions per item and aspect, which compare and contrast the items (see Figure 1.1 in the Introduction for an example). A "good" explanation is one that is relevant to the query, well-grounded in the source reviews, concise in presentation, and contrastive – i.e., it clearly highlights the differences (and similarities) among the items. For details on how we evaluate these summaries, please see Section 2.2.3.

3.1 STRUM-LLM

STRUM-LLM is the closest existing work to our goal of achieving contrastive summarization across aspects between two entities [24]. However, STRUM-LLM is purely extractive, not query-driven, and, from our experimentation, appears to be less effective at generating contrastive outputs. Hence, we adopted a STRUM-LLM-like architecture with significant modifications, with debate being a key addition, tailoring it to query-driven recommendations and enabling more contrastive outputs that align with user needs. Each step inspired from the STRUM-LLM architecture in the process—aspect extraction, merging, and contrasting—was clearly defined, ensuring a logical and efficient flow of information between stages.

3.1.0.1 Key Distinctions

Several modification were required to adapt STRUM-LLM for query-driven contrastive summarization. These included adding a Filter stage, adding a query to all prompts, adding the Debate Component (covered in Section 3.3.2), and removing the Value Merge and Usefulness stages.

Filter Stage The Filter stage was introduced to streamline the information provided to the debate and contrast stages. Given that the initial aspect extraction and merging stages can yield a significant amount of data, the Filter stage narrows down the results to the most informative aspects and associated sentences. This ensures that the debate, and any other baselines it is compared to, receives identical information and is focused on the information that is most relevant and impactful.

Query-Driven Since our approach is query-driven, which differs from STRUM-LLM, we want each stage to reflect that. In the Aspect Extraction stage, we want the selected aspects to relate to the query. In all other stages, we want the LLM to have the context of the query to maintain relevancy, and thus it is included in every prompt. We want to ensure that the merged aspects in the Aspect Merge stage and the selected information in the Filter stage is relevant, as well as in the Debate Component and Contrastive Summarizer.

Removal of Stages We remove the Value Merge and Usefulness stages in our proposed methodology. Since our methodology is query-driven, the selected aspects in the Aspect Extraction phase are already constrained by this and the prompt limits the LLM to extracting phrases that are highly relevant to the aspect, which should also be relevant to the intent of the query. Since this limits the breadth of aspects and values, it renders an additional Value Merge stage redundant and including it has the possibility of removing information that may be useful at later stages. The addition of the Filter stage also limits the number of value phrases per aspect to the 10 most informative phrases, thus having a similar role to that of the Value Merge stage, making it redundant to have both. Additionally, the same filtering also renders the Usefulness stage redundant, since its purpose is to filter out aspects that are not useful and reduce errors; however, the Filter prompt already asks the LLM to extract the top 3 most informative aspects.

3.2 Prompting Architecture

The architecture begins with a given query and two relevant entities associated with that query. For each entity, we gather the 50 most relevant texts corresponding to that entity for the given query. This serves as the primary data for generating contrastive explanations. Then, we begin a series of prompting stages which feed into each other. The full architecture, alongside the STRUM-LLM architecture, is portrayed in Figure 3.1.

3.2.1 Aspect Extraction

In this stage, we pass in each entity individually and prompt the LLM to identify five aspects relevant to the query and the specific entity. In addition, the LLM extracts ten relevant sentences from the provided texts for each aspect. An example of this stage's input and output is provided in Figure 3.2. An example of the prompt provided to the LLM is provided in Listing 3.1. The variables are denoted using {{variable_name}}.

```
{{destination}}
{{sentences}}

Query: {{query}}

Given the following destination and numbered texts, generate diverse
 and elaborative aspect phrases that describe what the user might be
 looking for according to the intent of the query provided and the
information provided for the destination. Use the JSON format
provided.

Requirements:
- The aspect phrase must be elaborate, specific, descriptive and
detailed.
- You must include the aspect and list of relevant extracted phrases
 for the destination for that aspect.
- You must include a citation in a [#] format for the sentence that
supports the aspect phrase from the provided sentences. Follow the
same numbering as the provided sentences.
- The values must be entire, long phrases extracted exaclty from the
 provided sentences.
- You must include exactly 5 aspects.
- For each aspect, you must include at least 10 extracted phrases
and each extracted phrase must be highly relevant to the aspect.
- Prioritize relevancy in the extracted phrases over the number of
phrases.



Output format:
{
  "<aspect>": ["extracted phrase [sentence #]", extracted phrase [
  sentence #]", ...],
  "<aspect>": ["extracted phrase [sentence #]", extracted phrase [
  sentence #]", ...],
  ...
}
```

Listing 3.1: LLM Prompt for Aspect Extraction Stage

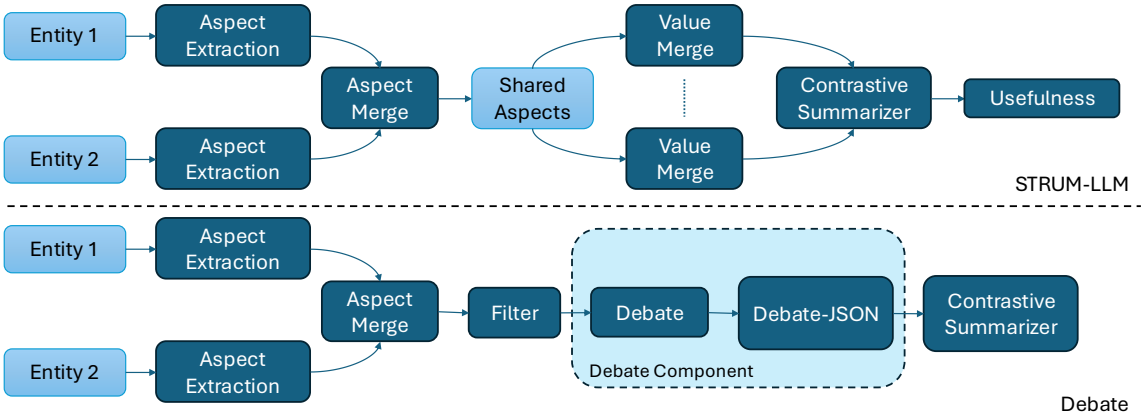

Figure 3.1: Debate Contrastive Explanation Architecture Alongside STRUM-LLM

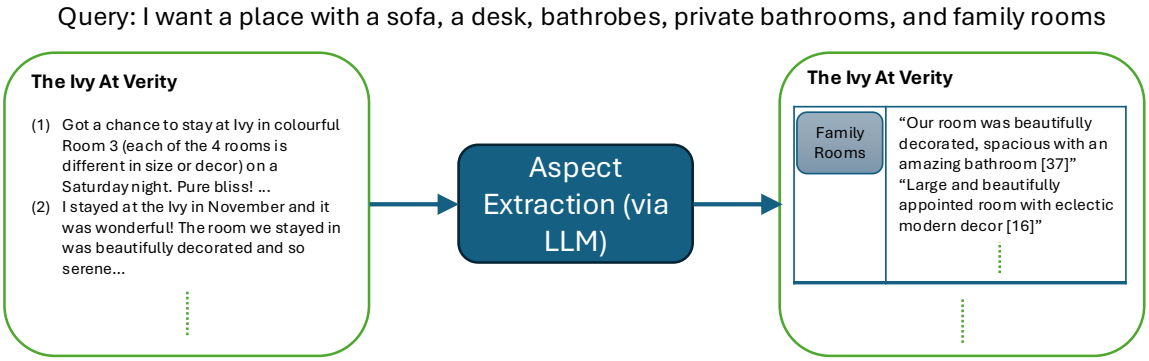

Figure 3.2: Example of Aspect Extraction Stage

3.2.2 Aspect Merge

Many of the aspects extracted for each entity are similar but not identical. The goal of this stage is to merge these similar aspects so they are treated as a single aspect in later stages. In this stage, the outputs from the Aspect Extraction stage for both entities are combined in a single prompt to the LLM. Similar aspects are merged, while those that are not merged remain unchanged for further stages. An example of the input and output of this stage is provided in Figure 3.3. An example of the prompt provided to the LLM is provided in Listing 3.2.

3.2.3 Filter

To refine the results and create a concise input for the debate stage, we introduce a Filter stage. This stage takes all aspects and relevant sentences for each entity and passes them through an LLM prompt to identify the top three most informative aspects and ten of the most informative relevant sentences for these aspects for each entity. An example of the input and output of this stage is provided in Figure 3.4. An example of the prompt provided to the LLM is provided in Listing 3.3.

```
Destination 1: {{dest1}}
Attributes 1: {{attributes1}}

Destination 2: {{dest2}}
Attributes 2: {{attributes2}}

Query: {{query}}

Merge any similar attributes from the attribute lists for each
destination. Return a JSON mapping the old attribute names exactly
to the new attribute names. Include the old attribute names from
both destinations in the output. Ensure the new attributes are
common to both destinations.

Output format:

{
  "{{dest1}}": {
    "oldAttr1": "newAttr1",
    "oldAttr2": "newAttr2",
    ...
  },
  "{{dest2}}": {
    "oldAttr3": "newAttr3",
    "oldAttr4": "newAttr4",
    ...
  }
}
```

Listing 3.2: LLM Prompt for Aspect Merge Stage

```
Destination 1: {{dest1}}
{{attributes1}}

Destination 2: {{dest2}}
{{attributes2}}

Query: {{query}}

Identify the top 3 most informative attributes. For each attribute,
identify exactly 10 of the most informative value pharases. You must
 have exactly 3 attributes per destination and exactly 10 value
phrases per attribute, no exceptions. Both destinations must have
the exact same 3 attributes. Follow the JSON output format provided
exactly.

Output format:
{
    "{{dest1}}": {
        "<attribute1_placeholder>": ["<value phrase 1> [<citation
        >]", "<value phrase 2> [<citation>]", ...],
        "<attribute2_placeholder>: ["<value phrase 1> [<citation>]",
         "<value phrase 2> [<citation>]", ...],
        ...
    },
    "{{dest2}}": {
        "<attribute1_placeholder>": ["<value phrase 1> [<citation
        >]", "<value phrase 2> [<citation>]", ...],
        "<attribute2_placeholder>: ["<value phrase 1> [<citation>]",
         "<value phrase 2> [<citation>]",
        ...
    }
}
```

Listing 3.3: LLM Prompt for Filter Stage

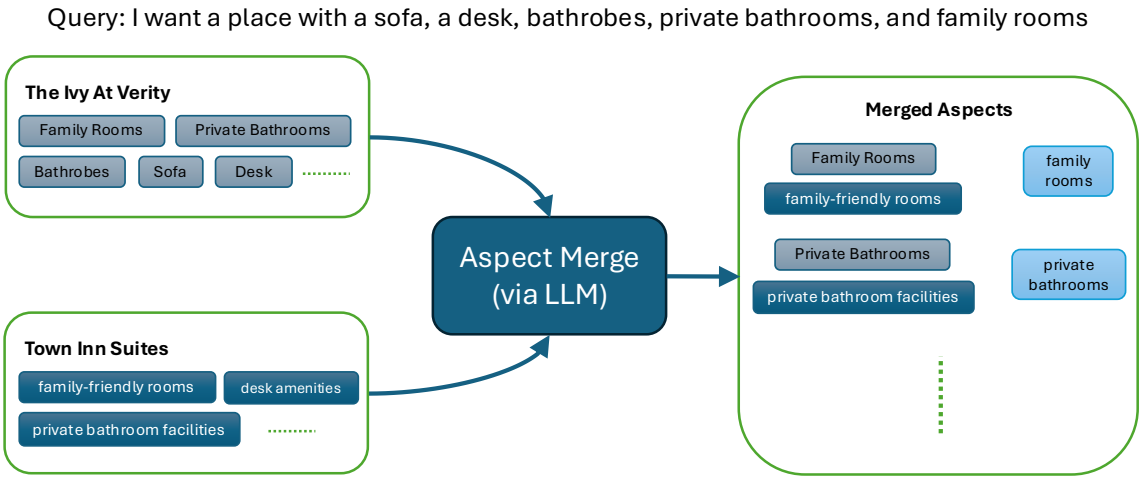

Figure 3.3: Example of Aspect Merge Stage

3.2.4 Debate

Motivated by research in discourse theory, much of the principles of ELM and Grice's Maxims, as outlined in Section 2.3, complement the criteria of good comparative explanations, as outlined in Section 2.2.3. Particularly, being contrastive and highlighting pros and cons is a highlight of both ELM and good comparative explanations [7, 43, 46]. As well, relevancy, diversity, usefulness, and groundedness are encouraged by Grice's Maxims as important for effective communication and also make for good comparative explanation [7, 20–22, 26, 41, 43, 57]. Therefore, given that LLMs like GPT-4 have seen debate and discourse in their training data, we conjecture that since they are already trained on these principles, we can leverage them to produce good contrastive explanations.

In the Debate stage, the LLM is prompted to simulate a debate between two individuals, Alice and Bob, each advocating for one of the two entities. Alice argues in favor of the first entity, while Bob supports the other. This debate is run independently for each aspect (so a total of 3 times for each query). Both individuals present arguments emphasizing the pros of their destination and the cons of the other. The debate prompt encourages the LLM to be extensive and detailed while incorporating exact phrases from the provided sentences with sentence number citations. An example of the debate input and prompt can be found in Figure 3.5 and the output can be found in Figure 3.6. The highlighted texts in the figure are examples of where the debate attempts to showcase contrast by emphasizing pros and cons. The underlined text showcases the use of direct quotes and citations from the provided textual data. An example of the prompt provided to the LLM is provided in Listing 3.4.

3.2.5 Debate-JSON

Following the debate, the LLM is instructed to provide a contrastive summary of the debate for the listed aspect in JSON format. Requirements for the output include avoiding direct mention of Alice or Bob and using destination names as keys. The summary must highlight the pros and cons of each destination, backed by quotes with sentence citations. This stage, as well as the Debate stage,

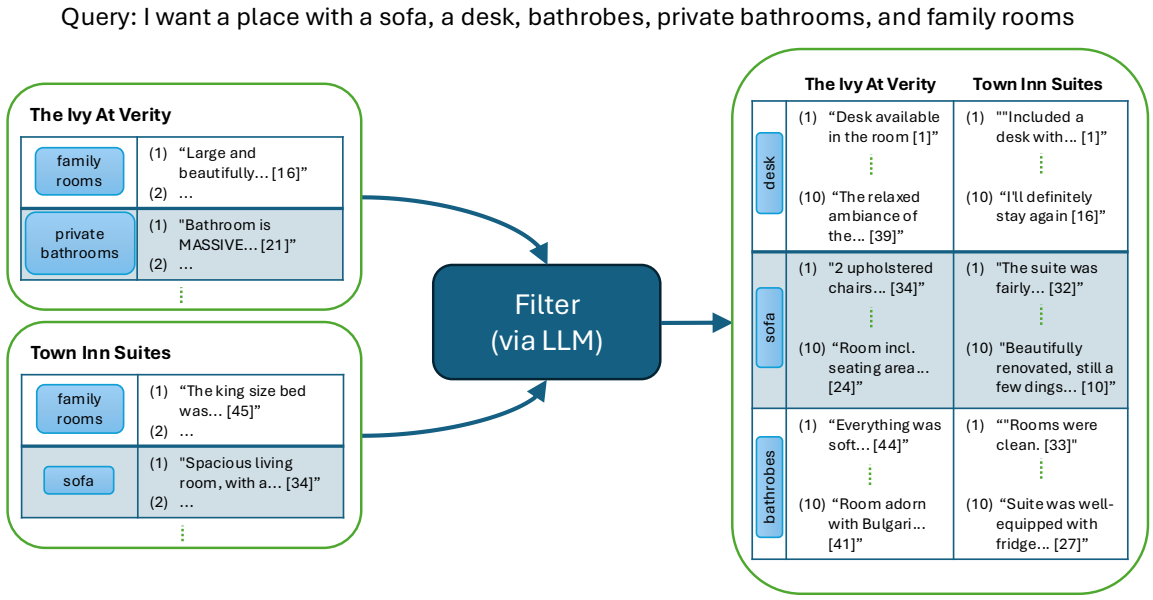

Figure 3.4: Example of Filter Stage

make up the Debate Component, which is the primary novelty from the baselines, and can be seen in Figure 3.1. An example of the output of this stage can be found in Figure 3.7. An example of the prompt provided to the LLM is provided in Listing 3.5.

3.2.6 Contrastive Summarizer

The final stage involves identifying the most contrasting and significant values between the two entities based on the debate and previous comparisons. The LLM returns exactly three aspects for each destination, with three bullet points each summarizing both the positive and negative information for each aspect. The aspects must be the same for both entities. The LLM is instructed to ensure that each bullet point is supported by a citation. An example of the prompt provided to the LLM is provided in Listing 3.6.

3.3 Design Decisions

In developing the methodology for generating contrastive explanations for natural language queries, several critical design decisions were made. These decisions shaped the architecture and flow of information throughout the stages, ultimately contributing to the desired quality of the output. The major design decisions are outlined below:

3.3.1 Language Model

The choice of language model is foundational to this methodology, particularly because the architecture relies on a series of interconnected prompts. Selecting the best language model is essential to ensure consistency and quality throughout each stage. We chose GPT-4o for its advanced ability

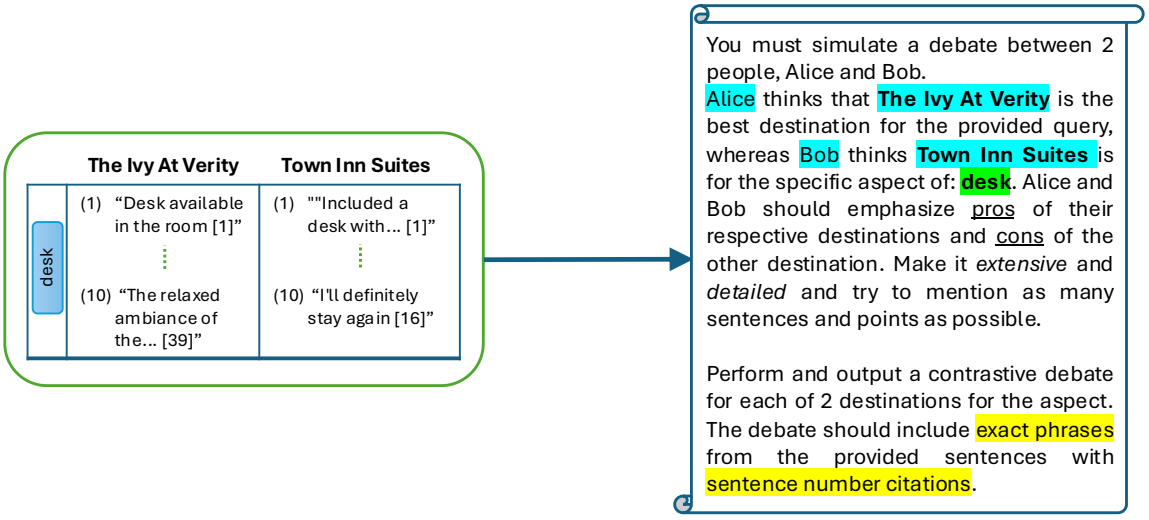

Figure 3.5: Example of input and prompt to Debate Stage

to generate coherent explanations and its good performance on reasoning benchmarks [44, 52]. This model's strengths make it well-suited for providing informative, human-like explanations that meet the requirements of our methodology.

3.3.2 Debate-style Prompting

To effectively contrast different entities, a debate-style prompting mechanism was employed. This approach simulates a discussion about two opposing perspectives, encouraging the LLM to argue in favor of one entity while simultaneously presenting the downsides of the other. This method aligns well with discourse theory principles outlined in Section 2.3, such as the Elaboration Likelihood Model (ELM), which emphasizes the importance of presenting pros and cons to facilitate deeper elaboration and effective contrast [46]. Additionally, many of these points align with Section 2.2.3.1 on what makes a good contrastive explanation [7, 43]. Grice's maxims also overlap with the characteristics of good explanations: the maxim of quantity aligns with diversity and usefulness, the maxim of quality with groundedness, the maxim of relation with relevancy, and the maxim of manner with clarity [21, 22, 27]. This debate-style approach is particularly effective for generating detailed and balanced contrastive explanations, as it emphasizes both positive and negative aspects in an engaging manner.

GPT-4o and other large language models have seen extensive examples of discourse in their training data, which includes debates, discussions, and structured arguments. We conjecture that this exposure has provided these models with an inherent ability to perform discourse effectively. As a result, GPT-4o can naturally engage in debate-style prompting and generate contrastive explanations, leveraging its familiarity with discourse structures to produce coherent and well-balanced arguments.

```
Query: {{query}}

Destination 1: {{dest1}}
{{sents1}}

Destination 2: {{dest2}}
{{sents2}}

You must simulate a debate between 2 people, Alice and Bob.
Alice thinks that {{dest1}} is the best destination for the provided
 query, whereas Bob thinks {{dest2}} is for the specific aspect of:
{{aspect}}. Alice and Bob should emphasize pros of their respective
destinations and cons of the other destination. Make it extensive
and detailed and try to mention as many sentences and points as
possible.

Perform and output a contrastive debate for each of 2 destinations
for the aspect. The debate should include exact phrases from the
provided sentences with sentence number citations.
```

Listing 3.4: LLM Prompt for Debate Stage

3.3.3 Concise Output Format

To facilitate easier analysis and comparison, the output format was designed to be concise yet informative. The LLM is asked to follow a specific output style with three aspects and three bullet points per aspect for each entity. This is the case for our debate approach and all baselines used, ensuring consistency for evaluation and promoting concision. This design decision was made to cater to both human & LLM downstream evaluation tasks. This is consistent as well with the literature that states that effective communication should be succinct [22].

```
Query: {{query}}
Aspect: {{aspect}}

Destination 1: {{dest1}}
{{sents1}}

Destination 2: {{dest2}}
{{sents2}}

Debate: {{debate}}

Based on the provided sentences and debate, provide a contrastive
comparison for each of 2 destinations for only the listed aspect in
JSON format.

Requirements are as follows:
- Do not mention Alice or Bob in the output.
- The keys should be the destination names, exactly as provided.
- The output should include summarization, backed by quotes with
exact phrases from the provided sentences with sentence number
citations.
- The output should be contrastive, specifically mentioning pros and
 cons of the destination.
- The phrasing of the output should be natural and more explanatory.
- You must include at least 5 points per aspect for each destination
.


Output format:
{
    "{{dest1}}": "<extracted phrases> [sentence #]",
    "{{dest2}}": "<extracted phrases> [sentence #]"
}
```

Listing 3.5: LLM Prompt for Debate-JSON Stage

```
Destination 1: {{dest1}}
{{attributes1}}

Destination 2: {{dest2}}
{{attributes2}}

Query: {{query}}

Identify the most contrasting and important values and return a JSON
 with these attributes and their values.

Requirements are as follows:
- You must return exactly 3 attributes for each destination.
- Each attribute must have exactly 3 bullet points, summarizing both
 the positives and negatives of the destination for that attribute.
- Each bullet point must be relevant to the attribute and must be
supported by a citation.
- The attributes should be identical for both destinations.
- Do not include meaningless attributes like null or N/A.

Output format:
{
    "{{dest1}}": {
        "<attribute1_placeholder>": ["<value phrase 1> [<citation
        >]", "<value phrase 2> [<citation>]", "<value phrase 3> [<
        citation>]"],
        "<attribute2_placeholder>: ["<value phrase 1> [<citation>]",
         "<value phrase 2> [<citation>]", "<value phrase 3> [<
        citation>]"],
        "<attribute3_placeholder>: ["<value phrase 1> [<citation>]",
         "<value phrase 2> [<citation>]", "<value phrase 3> [<
        citation>]"]
    },
    "{{dest2}}": {
        "<attribute1_placeholder>": ["<value phrase 1> [<citation
        >]", "<value phrase 2> [<citation>]", "<value phrase 3> [<
        citation>]"],
        "<attribute2_placeholder>: ["<value phrase 1> [<citation>]",
         "<value phrase 2> [<citation>]", "<value phrase 3> [<
        citation>]"],
        "<attribute3_placeholder>: ["<value phrase 1> [<citation>]",
         "<value phrase 2> [<citation>]", "<value phrase 3> [<
        citation>]"]
    }
}
```

Listing 3.6: LLM Prompt for Contrastive Summarizer Stage

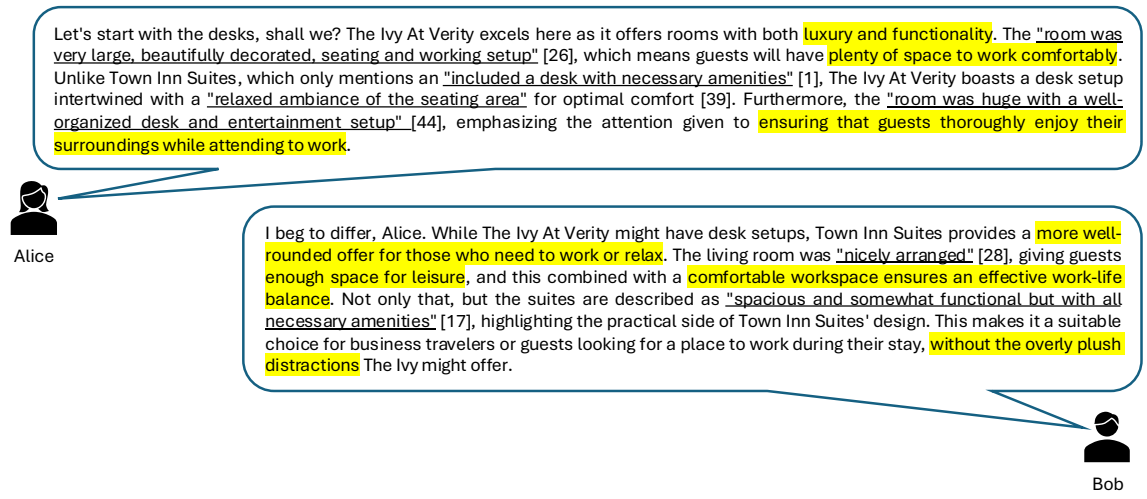

Figure 3.6: LLM Debate Example

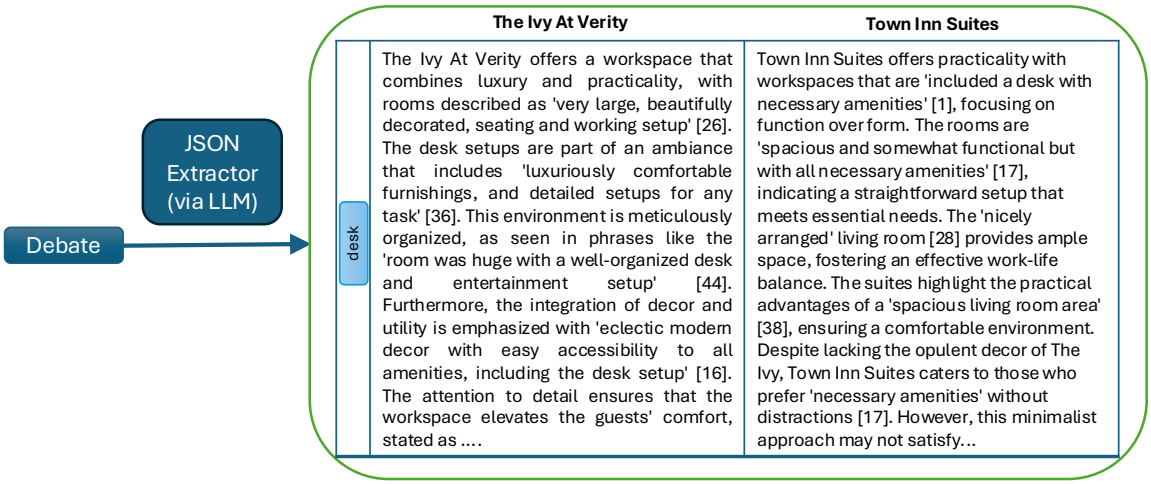

Figure 3.7: Example of output of Debate-JSON Stage

Chapter 4

Datasets

To evaluate the efficacy of contrastive explanations for natural language (NL) queries, we required datasets that contain comprehensive query-entity pairs with rich textual context. Specifically, the datasets must include a diverse set of queries, each associated with at least two relevant entities. Additionally, these entities must have detailed textual data that is pertinent to the given query, allowing for a thorough analysis of the contrastive explanations. This section describes the dataset requirements, the data collection process, and the specific datasets used in this study.

4.1 Query Requirements

The queries used in our evaluation play a critical role in assessing the quality of contrastive explanations. To ensure comprehensive evaluation, the queries need to be a combination of broad, specific and direct types. Broad queries allow for the inclusion of multiple entities that can provide different perspectives. An example of a broad query would be, "best cities for winter vacations". Specific and direct queries, on the other hand, help assess the model's capability to handle precise user needs. An example of a specific, direct query would be, "I want a restaurant with vegetarian options that's also kid-friendly". By using a diverse set of queries, we can evaluate the ability of LLMs to generate insightful, contrastive explanations that address different aspects of each entity.

4.2 Textual Data Requirements

Textual data is a crucial component of the datasets, as it provides the foundation for generating explanations. The textual data associated with each entity can be either subjective or objective, depending on the type of information being represented. Including both subjective and objective data helps in evaluating the model's capability to handle different types of information and provides a more comprehensive analysis of the explanations generated. The textual data must contain information that is relevant to aspects a user might care about in relation to a given query.

For example, the TravelDest dataset [64] contains textual data sourced from WikiVoyage, which offers rich, objective information about travel destinations, including cultural, adventure, nature, entertainment, and culinary aspects. This type of objective information ensures consistency and reliability in the analysis. On the other hand, the datasets for hotels and restaurants in Toronto

are based on reviews scraped from TripAdvisor, providing subjective, user-generated content. The inclusion of this subjective data allows us to analyze how well LLMs can handle more personal and diverse opinions in the context of contrastive explanations.

4.3 TravelDest Dataset

One of the core datasets employed for this evaluation is the TravelDest dataset, which contains 50 broad and indirect NL travel queries [64]. These queries cover various categories, including cultural, adventure, nature, entertainment, and culinary interests. Each query is associated with a set of 774 potential destination cities, accessible by major airlines, along with detailed textual descriptions for each destination, sourced from WikiVoyage (CC-licensed). These text descriptions offer ample data for deriving contrastive explanations, making the dataset suitable for our intended analysis.

The ground truth for the TravelDest dataset was established through human evaluation. Three independent annotators assessed the relevance of all 774 destination cities for each query, assigning scores on a scale from 1 to 5. Cities with an average score of at least 3 were deemed relevant, and these selections were subsequently verified by two additional travel experts to ensure consistency and reliability of the labels [64]. An example query, destinations, and some WikiVoyage data is provided for this dataset in Figure 4.1.

Query: Most romantic cities for a honeymoon

Abu Dhabi	São Paulo
"Le Méridien Abu Dhabi, Tourist Club Area, ☏ +971 2 6446666. Tell the taxi driver 'Lee Meridien' and he will not confuse it with Royal Meridien. Best amenity is the Meridien Village, an outdoor garden filled with restaurants and pubs, and on Thursday nights during the cooler months, a hangout for literally thousands of expats."	"Pirapora do Bom Jesus - Destination of a Catholic pilgrimage that is one of the oldest state's traditions."
"Shams Boutik (Reem Island, connected to Sun and Sky Towers). Su-Th 10AM–10PM, F Sa 10AM–midnight. A growing mall built around the community of Reem Island. It contains a growing number of good shops, including a supermarket, three restaurants, several fast food restaurants on the first floor, a café, a kids play area, a nail salon, a bookstore and more. Despite this, it is placed in an area that isn't usually busy, and is not very popular. (updated Jul 2017)"	"Some shopping malls that deserve special mention are Morumbi/Market Place (South Central - with more than 600 shops and dozens of restaurants), Eldorado (West - with an immense food court), Iguatemi (West - the oldest shopping mall of São Paulo, with very upscale profile), JK Iguatemi (West - the newest shopping mall for the wealthy Paulistanos), Cidade Jardim (West - famous for its internal gardens), Aricanduva (Far East - the city's largest and most famous working class shopping mall), and Frei Caneca (Downtown - the favorite of the LGBT public)."
"While walking in Abu Dhabi is not a problem for locals, tourists from colder climates will suffer from the heat and sun. The temperatures can exceed 45ºC in the summer."	"Santana de Parnaíba - City with a valuable Colonial historical center and strong religious traditions."
"UAE Pavilion. Sand dune-inspired exhibition centre designed by Norman Foster."	"You can find practically anything in São Paulo. Imported goods can be expensive, but look out for Brazilian-made bargains in all categories. Spend some time in one of the many 'shoppings' (as Brazilians call the shopping malls) and also look out for areas with shops catering for specific interests."
"The older bus service, operated by the Abu Dhabi Municipality, operates bus routes within city and to the other emirates. The routes within the city are very few. The buses are modern and air-conditioned. The services are as punctual as possible and operate more or less around the clock. The front few seats are reserved for women, men and families should move towards the back of the bus."	"And naturally, every safety recommendation that applies to big cities in general also applies to São Paulo:"
	"Just south of the city lies the Parque Estadual Serra do Mar (part of the Atlantic Forest South-East Reserves, a UNESCO World Heritage..."
...	...

Figure 4.1: Example of TravelDest Dataset

4.4 Toronto Hotels and Restaurants Datasets

In addition to TravelDest, we developed two additional datasets focusing on hotels and restaurants in Toronto. Natural language queries were manually created for each dataset, and reviews for

these entities were scraped from TripAdvisor. The use of reviews from TripAdvisor for extractive summarization tasks is seen in the literature as well [3]. The process to retrieve and identify relevant entities for the queries is outlined in Section 4.5. An example query, hotels, and some TripAdvisor review data is provided for this dataset in Figure 4.2.

Figure 4.2: Example of Hotels Dataset

These datasets also helped us experiment with subjective review data, as opposed to the more objective nature of the WikiVoyage data found in TravelDest. A summary of the three datasets is provided in Table 4.1.

Dataset Name	**Number of Queries**	**Number of Entities**	**Average Number of Reviews / Data Snippets Per Entity**	**Average Length of Review / Data Snippet (in characters)**	**Data Source**
TravelDest	50	774	163.31	264.53	WikiVoyage
Restaurants	26	43	94.51	441.96	TripAdvisor Reviews
Hotels	24	29	75.76	798.61	TripAdvisor Reviews

Table 4.1: Summary of Datasets

4.5 Dataset Processing

The processing pipeline involved several steps to extract and rank entities, generating the most relevant data subset for each query.

4.5.1 TravelDest Dataset

For the TravelDest dataset, we used an elaborative query reformulation (EQR) strategy, consistent with the original methodology proposed in the source paper [64]. We also used TAS-B embeddings, as described in the paper [29], to rank the entities.

4.5.2 Restaurants and Hotels Datasets

For the Restaurants and Hotels datasets, we did not use EQR. Instead, we used dense retrieval with OpenAI's text-embedding-3-small model[1] to encode the textual data and measure similarity between entities and queries using the cosine similarity of the embeddings.

4.5.3 Snippet Extraction

For each entity, we used the top-50 relevant data snippets to ensure balanced representation across entities. These snippets consisted of WikiVoyage entries for TravelDest or TripAdvisor reviews for hotels and restaurants.

4.5.4 Scoring and Selection

Cosine similarity was used to compute the similarity between query embeddings and each data snippet embedding. The arithmetic mean of the similarity scores of the top-50 snippets for each entity determined its relevance score. We selected the top two entities for each query based on these scores, and their associated top-50 snippets formed the basis for contrastive explanation generation.

4.5.5 Key Hyperparameters

The key hyperparameters were: (i) number of entities per query ($k = 2$), (ii) scoring method (average similarity of top data snippets), and (iii) number of data snippets per entity ($k = 50$). This standardized approach ensured consistency across all datasets for evaluating contrastive explanations.

[1] https://openai.com/index/new-embedding-models-and-api-updates/

Chapter 5

Experiments and Evaluation

This chapter presents the experiments and results for the following research questions:

- **RQ1**: Does debate-style prompting improve query-driven contrastive explanations?
- **RQ2**: Does the aggressiveness in debate-style prompting impact the quality of contrastive explanations?
- **RQ3**: Can we trust pairwise LLM Win Rate evaluation of query-driven contrastive explanations?

5.1 Baselines

The primary distinction between the proposed debate-style prompting architecture and the baselines lies in the absence of the Debate Component, as illustrated in Figure 3.1. The baselines used in this evaluation are designed to be similar to a STRUM-LLM system, which is outlined in Sections 2.2.2.2 and 3.1. This decision was made to isolate and demonstrate the specific effects of adding debate-style prompting to the closest architecture available in the current literature for contrastive summarization. However, these baselines have been modified to be query-driven, aligning them with the debate methodology described in 3.2. The baselines selected for this study are:

- **STRUM-LLM-like with Contrastive Summarizer**: This version incorporates a Contrastive Summarizer in the final stage, which explicitly instructs the model to "identify the most contrasting and important values." This is the same Contrastive Summarizer prompt used in the debate methodology outlined in Section 3.2. This baseline tests the hypothesis that contrast-focused prompts can enhance the summarization by emphasizing key differences. The prompt used for the Contrastive Summarizer is detailed in Listing 3.6.
- **STRUM-LLM-like with Simple Summarizer**: In this version, a Simple Summarizer replaces the Contrastive Summarizer in the final stage. This baseline serves as a control to evaluate the performance of a summarization approach without explicit contrastive instructions. The architecture of these baselines, alongside STRUM-LLM and our proposed methodology, is shown in Figure 5.1.

Components	STRUM-LLM	Baseline Simple	Baseline Contrastive	Debate
Query-Driven	×	✓	✓	✓
Aspect Extraction	✓	✓	✓	✓
Aspect Merge	✓	✓	✓	✓
Value Filtering	✓ (Value Merge)	✓ (Filter)	✓ (Filter)	✓ (Filter)
Debate	×	×	×	✓
Summarization	Contrastive	Simple	Contrastive	Contrastive

Table 5.1: Comparison of Methodologies

For further comparison, Table 5.1 outlines the primary differences between STRUM-LLM, the baselines, and the debate prompting methodologies.

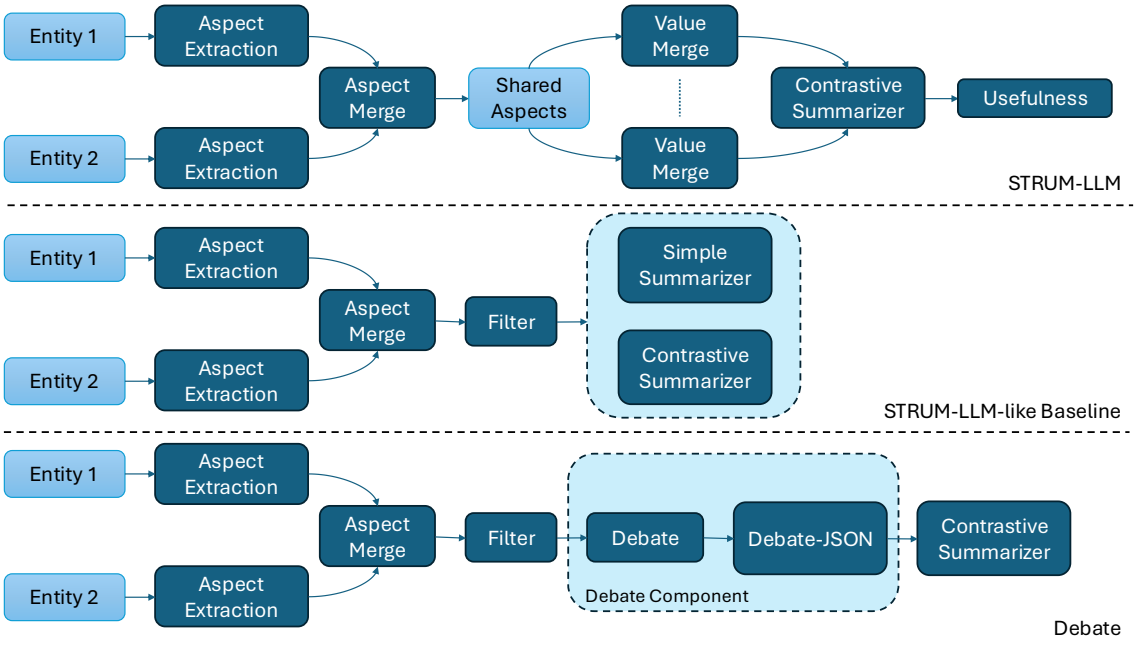

Figure 5.1: STRUM-LLM, Baselines, and Debate Architectures

In order to ensure a fair evaluation, the stages from Aspect Extraction up to and including the Filter stage were run once for each dataset, with identical results passing through the later stages for both of the baselines and the debate prompting methodology. This approach ensures that we are evaluating the effects of these later stages and not the differences in the earlier stages, since the output after the Filter stage is identical.

5.2 RQ1: Debate-style prompting methodology

RQ1: Does debate-style prompting improve query-driven contrastive explanations?

To answer this research question, we have designed a comprehensive evaluation using a pairwise

Win Rate comparison conducted by an LLM, GPT-4o. We aim to systematically assess whether debate prompting results in better contrastive explanations compared to existing baselines.

5.2.1 LLM Experimental Design

To evaluate the proposed debate prompting methodology, we employ a pairwise LLM evaluation approach. Recent studies have shown that LLMs, and particularly GPT-4, are effective at performing pairwise comparisons for evaluating the quality of explanations, as discussed in Section 2.2.3.2 [39,40]. Therefore, we have adopted this approach using GPT-4o. This method allows us to systematically compare outputs from the debate prompting against those from the baselines.

Specifically, for each query, we evaluate a single aspect individually (out of the three identified for that query). We present the output generated by the debate-style prompting as Explanation "A" (or "B") and one of the baselines' explanations as the other. We do this for both explanations in both directions (i.e., debate as A and baseline as B, and vice versa), and take an aggregated Win Rate over both directions. The LLM is also provided with the query and aspect for context, and it is allowed to choose Explanation A, B, or a tie as the winner. We count a direct win of A or B as 1 win, and a tie as 0.5.

The LLM is instructed to evaluate the two explanations based on the following set of criteria, using definitions based on established literature and covered in Section 2.2.3.1 [7, 21, 27, 43]:

- **Contrast**: The summarizations should differentiate between the two domains well, such as by including pros and cons and details, and help a user choose one domain instead of the other.
- **Relevancy**: The summarizations provided should be relevant to each aspect and query provided.
- **Diversity**: The summarizations should provide multiple different points in support of and against the domain for each aspect. Repetitive points should be penalized, and a variety of different points should be rewarded. Additional context that is not repetitive should be rewarded.
- **Usefulness**: The summarizations should provide useful information and be informative for a user to make a decision between the two domains.

The prompt used is provided in Listing 5.1. The "{{domain}}" variable represents the specific subject of comparison, such as travel destinations, hotels, or restaurants, depending on the dataset being evaluated. The LLM is also asked to provide an explanation for each judgment that it makes. Existing work, such as the work on Automatic CoT [72], shows that even just asking an LLM to explain its reasoning can lead to more correct outputs, hence this approach was used here.

5.2.2 Experimental Results

We ran these evaluations on the three datasets outlined in Chapter 4, which include datasets focusing on different domains such as travel destinations, hotels, and restaurants, with both objective data from WikiVoyage as found in the TravelDest dataset and subjective TripAdvisor review data as found in the other two.

```
Query: {{query}}

Explanation A:
{{a}}

Explanation B:
{{b}}

Your role is to evaluate Explanation A and Explanation B as being
good contrastive explanations for {{domain}} recommendation. The
provided criteria should be used and you should select either "A" or
 "B" as the winner for each criterion or "tie" if both explanations
are the same. You should provide explanations for each of your
choices.

Criteria:
contrast - The summarizations should differentiate between the two
{{domain}}s well, such as by including pros and cons and details,
and help a user choose one {{domain}} instead of the other.
relevancy - The summarizations provided should be relevant to each
aspect and query provided.
diversity - The summarizations should provide multiple different
points in support and against the {{domain}} for each aspect.
Repetitive points should be penalized and a variety of different
points should be rewarded. Additional context that is not repetitive
 should be rewarded.
usefulness - The summarizations should provide useful information
and be informative for a user to make a decision between the two {{
domain}}s.

Output in JSON format:
{
    "contrast": "A" or "B" or "tie",
    "contrast_explanation": <explanation>,
    "relevancy": "A" or "B" or "tie",
    "relevancy_explanation": <explanation>,
    "diversity": "A" or "B" or "tie",
    "diversity_explanation": <explanation>,
    "usefulness": "A" or "B" or "tie",
    "usefulness_explanation": <explanation>
}
```

Listing 5.1: LLM Prompt for Pairwise Win Rate Evaluation

Criterion	Debate vs. Baseline Contrastive Summarizer	Debate vs. Baseline Simple Summarizer
Contrast	**0.86** [0.80, 0.91]	**0.88** [0.83, 0.93]
Relevance	**0.58** [0.52, 0.64]	**0.58** [0.52, 0.64]
Diversity	**0.83** [0.77, 0.90]	**0.85** [0.79, 0.91]
Usefulness	**0.84** [0.78, 0.90]	**0.90** [0.85, 0.95]

Table 5.2: Pairwise LLM Win Rate for Debate vs. STRUM-LLM-like Baselines for Restaurants Datasets. 95% Confidence Interval provided next to each value.

Criterion	Debate vs. Baseline Contrastive Summarizer	Debate vs. Baseline Simple Summarizer
Contrast	**0.84** [0.77, 0.90]	**0.83** [0.77, 0.89]
Relevance	**0.59** [0.52, 0.67]	**0.58** [0.51, 0.66]
Diversity	**0.80** [0.72, 0.87]	**0.88** [0.82, 0.94]
Usefulness	**0.77** [0.70, 0.85]	**0.85** [0.79, 0.91]

Table 5.3: Pairwise LLM Win Rate for Debate vs. STRUM-LLM-like Baselines for Hotels Datasets. 95% Confidence Interval provided next to each value.

For the restaurant and hotel datasets, as we see in Tables 5.2 and 5.3, debate outperforms both baselines on all criteria, most by a wide margin. The Win Rate for debate is at or above 77% (with most well above 80%) for contrast, diversity, and usefulness and is only lower (but still above 50%) for relevancy. In the example in Figure 5.2, the LLM scores Comparison A (debate) as the winner for contrast, diversity, and usefulness, and a tie for relevancy, which is consistent with the typical results we see in Table 5.2. As well, when we qualitatively analyze these results, we see that all bullet points for both debate and the baseline are clearly relevant to the aspect of "City Views", which further validates the LLM's judgment. However, we see clear repetitiveness in the baseline, with 54th floor mentined twice, resulting in a debate win for diversity. The debate explanation also mentions pros and cons and additional relevant and useful information, which is consistent with the wins on contrast and usefulness.

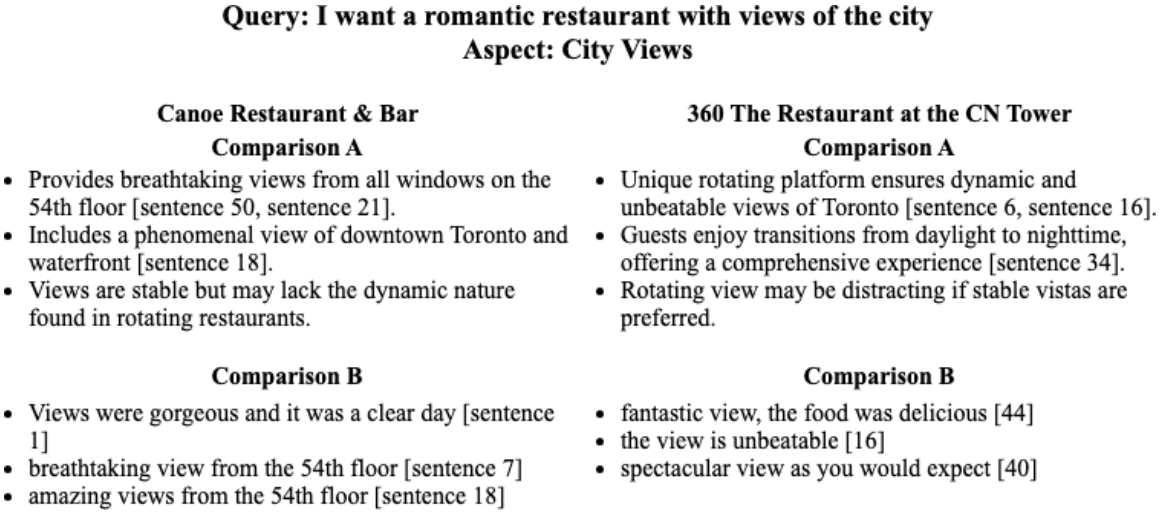

Figure 5.2: Restaurant Debate vs. STRUM-LLM-like Contrastive Summarizer Example. Comparison A is Debate and Comparison B is STRUM-LLM-like.

Criterion	Debate vs. Baseline Contrastive Summarizer	Debate vs. Baseline Simple Summarizer
Contrast	**0.64** [0.58, 0.70]	**0.79** [0.74, 0.84]
Relevance	**0.51** [0.46, 0.56]	**0.57** [0.52, 0.61]
Diversity	**0.55** [0.48, 0.61]	**0.70** [0.64, 0.75]
Usefulness	**0.61** [0.55, 0.67]	**0.73** [0.67, 0.78]

Table 5.4: Pairwise LLM Win Rate for Debate vs. STRUM-LLM-like Baselines for TravelDest. 95% Confidence Interval provided next to each value.

For the TravelDest dataset, the results in Table 5.4 show clear wins for debate over the Simple Summarizer baseline, with all confidence intervals being above 50% and the Win Rates for the contrast, diversity and usefulness criteria being above 70%. Debate also wins against the Contrastive Summarizer baseline, however, for relevance and diversity, the confidence interval overlaps below 50% and the wins are less pronounced. This warrants additional investigation as to why there was poorer performance on the more objective WikiVoyage data in the TravelDest dataset than the more subjective review data in the other TripAdvisor datasets. An example is provided in Figure 5.3 which the LLM has judged to be a tie on all criteria. We see in both comparisons that pros and cons are mentioned, indicating that both are contrastive. The diversity in chosen points is also observed to be mostly the same, with all 3 points being relevant and useful for both destinations and comparisons. All citations are the same for Sydney for both comparisons and two of the three citations for San Francisco are the same among both comparisons as well. This indicates a lack of diversity in the objective source content that may limit the ability of debate to extract additional diverse, useful, relevant, and contrastive content over the baseline. Therefore, these results indicate that debate performs better when there is a larger quantity of information and more opinionated data to debate about.

We see from Figures 5.4 and 5.5 that for the review-based datasets, most frequently, debate wins at three out of the four evaluation criteria for a given query, irrespective of comparison to the contrastive or simple baseline methodologies. From Figure 5.6, we see that for the more objectively-based TravelDest dataset, the most frequent result is that debate does not beat the contrastive baseline on any criteria for a given query. However, when compared to the simple baseline, debate often wins on three out of the four criteria as well.

Further analysis is done by aggregating the number of aspect wins per query. Since there are three aspects per query, the best outcome for a criteria is for all three aspects to win with high frequency. We see from Figures 5.7, 5.8, and 5.9 that debate never wins on the relevancy criteria for all 3 aspects and most frequently, for none of the aspects does relevancy win.

5.3 RQ2: Impact of aggressiveness in debate-style prompting

RQ2: Does the aggressiveness in debate-style prompting impact the quality of contrastive explanations?

In this section, we explore whether the level of aggressiveness in debate-style prompting influences the quality of contrastive explanations. We designed an experiment to assess the impact of varying levels of assertiveness during the debate stage and evaluated the outputs using the same criteria as

Query: Family friendly cities for vacations
Aspect: Family-Friendly Attractions

San Francisco

Comparison A

- Golden Gate Park is home to the California Academy of Sciences with a vast array of exhibits [19].
- Fisherman's Wharf is noted as 'a tourist trap' yet offers a lively atmosphere with street entertainers and museums [25].
- Some Civic Center attractions are less interactive for younger children [1].

Comparison B

- Golden Gate Park offers museums and historical sites for families to explore [19].
- The San Francisco Zoo is well-maintained and great for children interested in animals like penguins and lions [31].
- Fisherman's Wharf is a popular spot for tourists, with street entertainers and sea lions, though it's considered a tourist trap by locals [25].

Sydney

Comparison A

- Bondi and other beaches offer family-friendly sun and sand experiences [27].
- Wild Australian animals can be seen, appealing to families with children [34].
- Some attractions like the Sydney Opera House may not be as engaging for children seeking interactive experiences [23].

Comparison B

- Sydney Opera House and the Art Gallery of New South Wales are prominent landmarks for families to visit [23].
- Sydney offers the opportunity to see wild Australian animals in their natural habitat [34].
- Sydney beaches are great for families to enjoy both summer and winter outdoor activities [27].

Figure 5.3: TravelDest Debate vs. STRUM-LLM-like Contrastive Summarizer Example. Comparison A is Debate and Comparison B is STRUM-LLM-like.

RQ1.

5.3.1 Experimental Design

To investigate the impact of aggressiveness, we added a sentence to the debate prompt found in Listing 3.4 for different variations. For the 'nice' version, we added: 'Alice and Bob should both be nice and polite to each other.' For the 'aggressive' version, we added: 'Alice and Bob should both be aggressive and assertive with each other.' The rest of the prompt remained identical to the standard version. We used the same outputs from the Filter stage as input for each debate to ensure consistency, with the only variation being in the debate prompt itself. All experiments received the same inputs. We conducted the same pairwise Win Rate comparison as described in RQ1 (see Section 5.2.1), and ran evaluations with both pairwise orderings, as was done in that section, to prevent ordering bias.

5.3.2 Experimental Results

The experiments conducted on the Hotels dataset, comparing the 'standard', 'aggressive', and 'nice' versions are presented in Table 5.5. The results on this dataset particularly showed that the aggressive prompt outperformed the standard on all criteria, while the standard prompt performed better than the 'nice' prompt.

On both other datasets, Restaurants and TravelDest, shown in Tables 5.6 and 5.7, respectively, the standard debate beats out both aggressive and nice debate on all criteria, expect for relevance on the TravelDest dataset, where it ties. However, the 95% confidence intervals for all comparisons on all three datasets overlapped with 0.5, indicating that these results are not statistically significant.

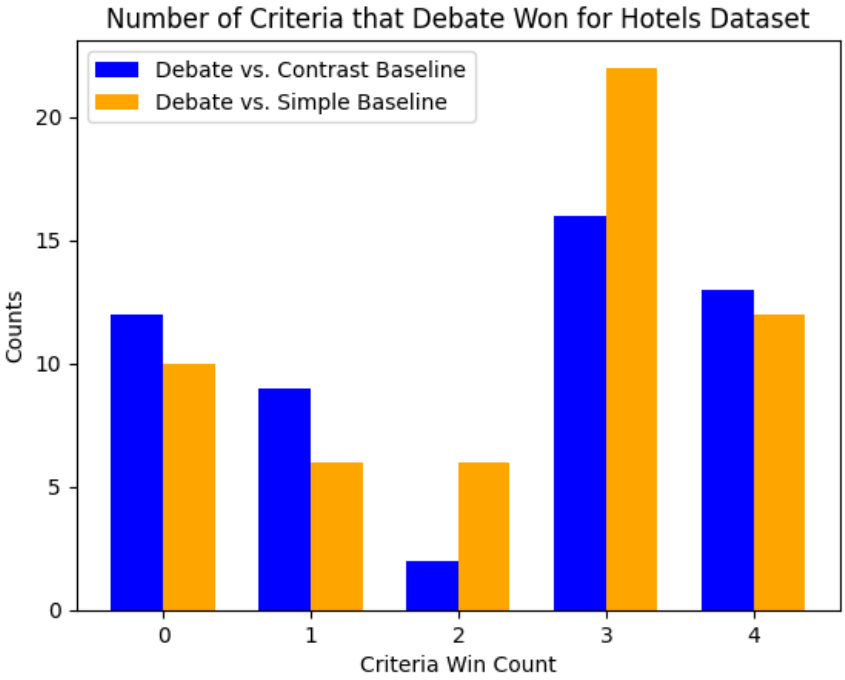

Figure 5.4: Criteria Wins for Debate on Hotels Dataset

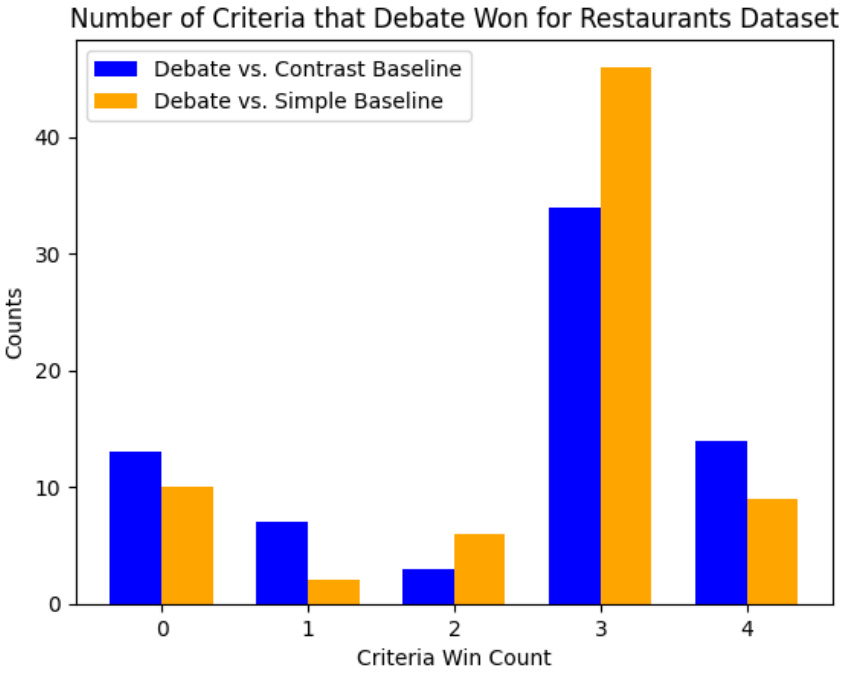

Figure 5.5: Criteria Wins for Debate on Restaurants Dataset

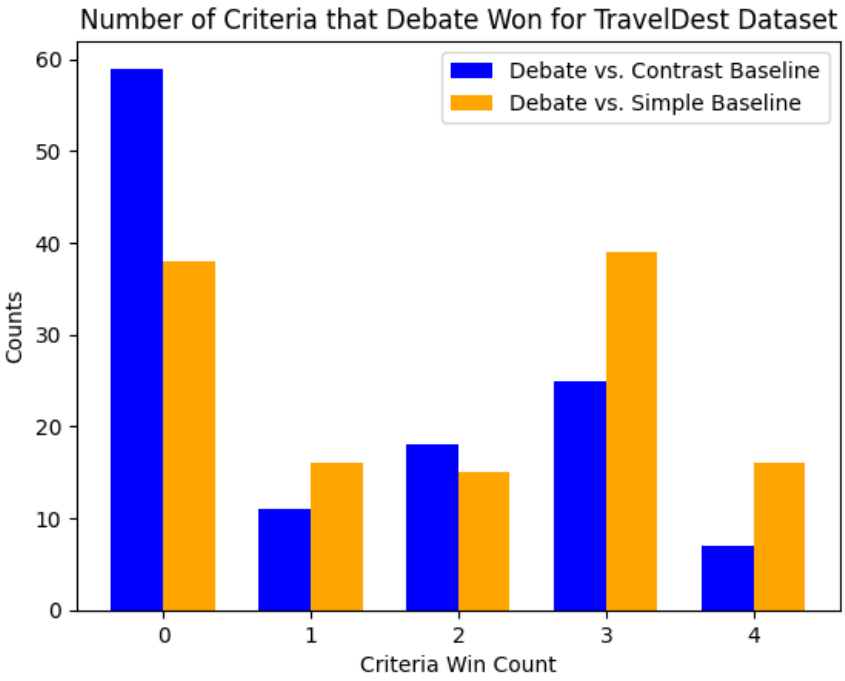

Figure 5.6: Criteria Wins for Debate on TravelDest Dataset

Criterion	Standard vs. Aggressive Debate	Standard vs. Nice Debate
Contrast	0.44 [0.32, 0.57]	**0.57** [0.49, 0.64]
Relevance	0.46 [0.36, 0.56]	**0.56** [0.49, 0.62]
Diversity	0.42 [0.28, 0.55]	**0.54** [0.46, 0.62]
Usefulness	0.43 [0.29, 0.56]	**0.55** [0.47, 0.63]

Table 5.5: Standard vs. Aggression Variations in Debate-style Prompting on the Hotels Dataset. 95% Confidence Interval provided next to each value.

Criterion	Standard vs. Aggressive Debate	Standard vs. Nice Debate
Contrast	**0.58** [0.50, 0.66]	**0.56** [0.47, 0.64]
Relevance	**0.55** [0.49, 0.60]	**0.56** [0.49, 0.62]
Diversity	**0.53** [0.44, 0.61]	**0.56** [0.48, 0.65]
Usefulness	**0.59** [0.50, 0.67]	**0.56** [0.48, 0.64]

Table 5.6: Standard vs. Aggression Variations in Debate-style Prompting on the Restaurants Dataset. 95% Confidence Interval provided next to each value.

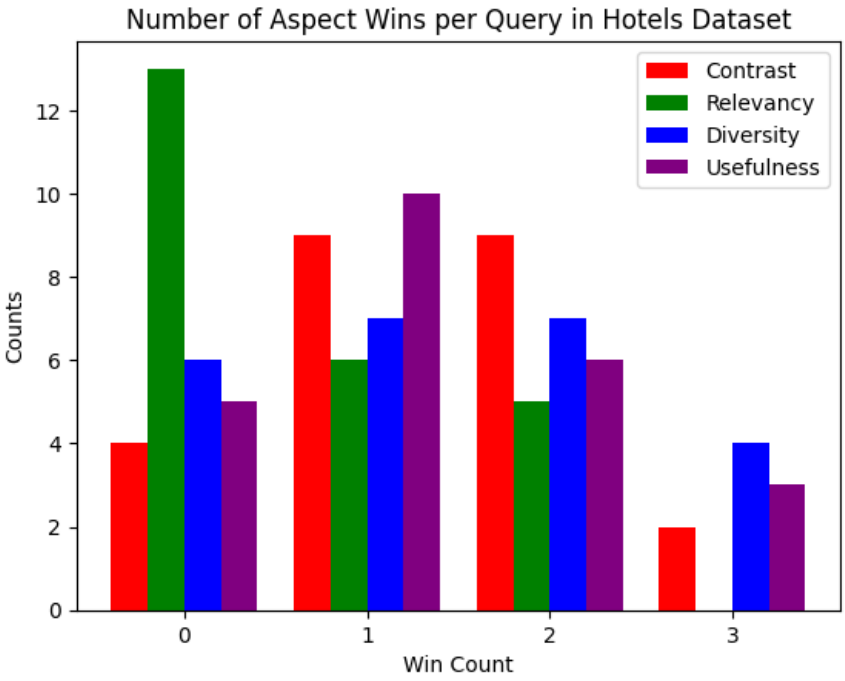

Figure 5.7: Number of Aspect Wins per Query for the Hotels Dataset for Debate vs. Contrast Baseline

Criterion	Standard vs. Aggressive Debate	Standard vs. Nice Debate
Contrast	**0.54** [0.50, 0.57]	**0.51** [0.47, 0.55]
Relevance	**0.51** [0.48, 0.55]	0.50 [0.46, 0.54]
Diversity	**0.52** [0.46, 0.58]	**0.51** [0.46, 0.57]
Usefulness	**0.54** [0.49, 0.60]	**0.51** [0.46, 0.56]

Table 5.7: Standard vs. Aggression Variations in Debate-style Prompting on the TravelDest Dataset. 95% Confidence Interval provided next to each value.

5.4 RQ3: Reliability of LLM-based Win Rate evaluation

RQ3: Can we trust pairwise LLM Win Rate evaluation of query-driven contrastive explanations?

To address RQ3, we designed a user study to validate the reliability of LLM-based pairwise Win Rate evaluations. The main objective was to determine whether the preferences expressed by human participants align with those made by the LLM in evaluating query-driven contrastive explanations. By comparing the results from human participants with the LLM's pairwise evaluations, we aim to validate the claims in the literature on how well LLM evaluations reflect human judgment in terms of the specific criteria used for evaluation in this study.

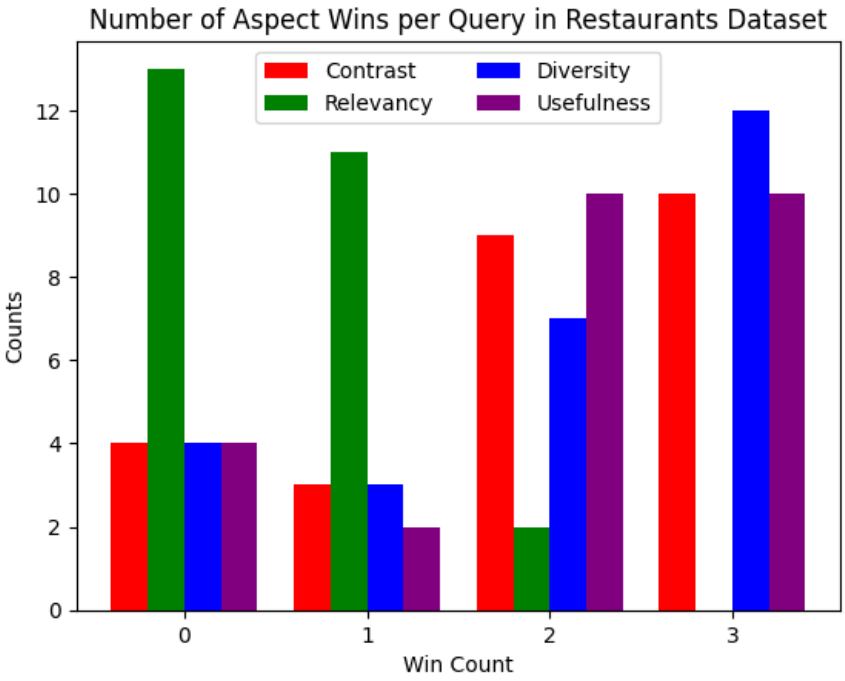

Figure 5.8: Number of Aspect Wins per Query for the Restaurants Dataset for Debate vs. Contrast Baseline

5.4.1 User Study Experimental Design

The user study involved collecting pairwise comparisons from human participants on the same explanations that were evaluated by the LLM. We selected 10 random aspects from the overall set of 78 aspects available from the Restaurants dataset (26 queries times 3 aspects). Since the STRUM-LLM-like Contrastive Summarizer is the strongest baseline, we showed users a comparison between that and our debate method. Participants were given the exact same criteria that the pairwise LLM evaluator used, which are defined in Section 5.2.

To ensure fairness, the ordering of whether debate or the baseline appears as comparison A or B was randomized evenly among the users. Additionally, the ordering of the 10 aspects was randomized evenly. The study was run among 10 participants. An example of one of the 10 examples shown to users can be seen in Figure 5.2.

To fairly compare the human evaluations with the LLM's evaluations and ensure no ordering bias, we considered an LLM win only if both directions (debate as A and baseline as B, and vice versa) agreed on the same winner. Otherwise, we counted the LLM's evaluation as a tie. Participants were presented with the query, aspect, and two explanations (Comparison A and Comparison B), similar to the setup used in the LLM evaluation process. Human evaluators were instructed to choose between the two explanations based on the same evaluation criteria used by the LLM: contrast, relevancy, diversity, and usefulness. To ensure the reliability of human judgments, participants were provided with identical definitions of each evaluation criterion as the LLM received.

To collect overall general feedback and sentiment from users, we also prompt them to provide commentary on the study, with the following question:

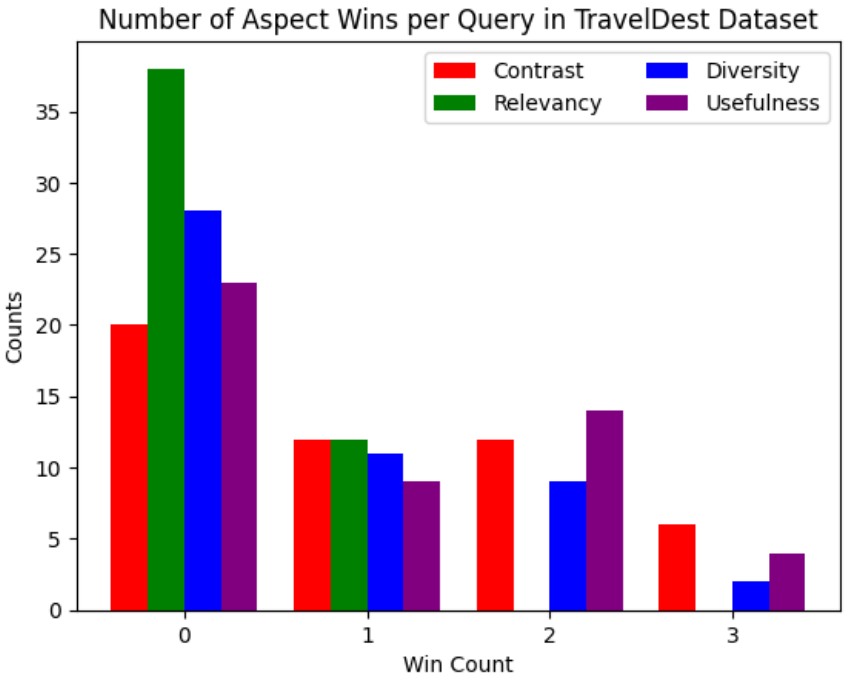

Figure 5.9: Number of Aspect Wins per Query for the TravelDest Dataset for Debate vs. Contrast Baseline

> Please provide feedback on the comparisons you saw in this survey. Did you notice anything notable or any patterns with any of the comparisons? Were there some characteristics that made some comparisons better than others on these criteria or in general? Feel free to provide any additional insight or information you want to share.

5.4.2 Experimental Results

To assess the agreement between human evaluators and LLM-based evaluations, we employed Cohen's Kappa as the primary metric for measuring inter-rater reliability, as it is the standard metric for this [42]. Cohen's Kappa was used in three distinct ways:

1. **Agreement Among Human Participants (Average)**: We measured the agreement level among the 10 human evaluators to determine the consistency in their judgments when comparing the explanations. This helped us understand how often human evaluators agreed on the same choice (either debate or the baseline). Formally, we define the agreement $\kappa_{i,j}$ between user i and user j, where $i \neq j$ and $i, j \in \{1, \ldots, n\}$, and $n = 10$ is the number of users in the study. The average agreement of user i among all other users is defined as:

$$\theta_i = \frac{\sum_{j \neq i} \kappa_{i,j}}{n-1}. \tag{5.1}$$

The overall average agreement among all users is then given by:

$$\kappa_{\text{ua}} = \frac{\sum_{i=1}^{n} \theta_i}{n}. \tag{5.2}$$

2. **LLM Agreement with Human Evaluators (Average)**: We also computed the average Cohen's Kappa value between the LLM's evaluations and each individual human evaluator's choices. This allowed us to gauge how closely the LLM's decisions aligned with those of the participants on average. A similar calculation to Equation 5.2 is computed here.

3. **Human Agreement with the Majority Vote (Average)**: We measured the average agreement between the human evaluators and their majority vote. The majority vote was defined as at least 6 out of 10 human evaluators choosing either debate or the baseline as the winner. If no majority was achieved, the evaluation was considered a tie. This measure helped assess the agreement on average of humans with a more consensus-based human decision. A similar calculation to Equation 5.2 is computed here.

4. **LLM Agreement with the Majority Vote**: Lastly, we measured the agreement between the LLM's evaluations and the majority vote of human evaluators. This measure helped assess the reliability of the LLM's evaluations in comparison to a more consensus-based human decision.

Table 5.8 provides a summary of these four metrics. We see that the highest Cohen's Kappa values seen for each criterion are usually for LLM agreement with human evaluators when compared to the average agreement among humans (noted in bold in the second column). For contrast, average agreement among human participants is slightly higher but both values are close. Furthermore, the LLM agreement with majority vote is higher than the average agreement among human participants for all criteria. However, when comparing the LLM agreement with majority vote to the average human evaluator agreement with majority vote, we see that for three of the criteria, the human evaluators have higher agreement with the majority vote than the LLM. However, the values are close and the LLM agreement significantly exceeds that of human evaluators with the majority vote for the diversity criteria. We also calculate Fleiss' Kappa among all human evaluators in Table 5.9 and find that the human agreement with each other is quite low. These results are close to or below the average agreement of the LLM with human evaluators and is far below the LLM agreement with the majority human vote. We also show the individual Cohen's Kappa scores for each criteria for each user with the LLM in Table 5.10.

Therefore, this is a good indication that, as the existing literature suggested [40], LLM pairwise evaluation is well-aligned with (and sometimes even better than) human to human alignment on pairwise evaluation. Thus, to answer the RQ, we can trust pairwise LLM evaluation of query-driven contrastive explanations.

Criterion	Agreement Among Human Evaluators (Average)	LLM Agreement with Human Evaluators (Average)	Agreement of Human Evaluators with Majority Vote (Average)	LLM Agreement with the Majority Vote
Contrast	**0.087584**	0.06483	**0.24545**	0.210526
Relevance	0.143305	**0.217614**	**0.399715**	0.310345
Diversity	0.112613	**0.215235**	0.23009	**0.411765**
Usefulness	0.132112	**0.206982**	**0.281494**	0.268293

Table 5.8: Summary of Cohen's Kappa Statistics for LLM & Human Agreement

Criterion	Fleiss' Kappa for Human Evaluators
Contrast	0.067246
Relevance	0.125931
Diversity	0.104792
Usefulness	0.094024

Table 5.9: Fleiss' Kappa for Human Participants

Criterion	User 1	User 2	User 3	User 4	User 5	User 6	User 7	User 8	User 9	User 10
Contrast	0.02439	0.137931	0.189189	0.189189	-0.176471	0.107143	0.02439	0	0.152542	0
Relevancy	0.365079	0.130435	0.354839	0.090909	0.122807	0.242424	0.344262	0.245283	0.189189	0.090909
Diversity	0.090909	0.02439	0.756098	-0.111111	0.411765	0.090909	0.512195	0.210526	0.166667	0
Usefulness	0.038462	0.074074	0.268293	0.375	0	0.206349	0.444444	0.444444	0.21875	0

Table 5.10: Pairwise Cohen's Kappa Between Each User and the LLM

5.4.2.1 Additional Insights

Based on the final open question we asked the human participants to complete, we gathered the following insights:

- Users preferred longer points with more descriptors, adjectives, and mention of specific details, such as explicitly mentioning specific servers or dishes at a restaurant. Users found that these points provided more context and quotes, which they preferred.
- However, users noted that a downside of longer points is that they tended to be more repetitive, which users did not prefer.
- Users liked diversity and the emphasis on differences between the entities, noting that they especially liked to see cons being mentioned.

These findings are consistent with the literature on evaluating contrastive explanations. Higher diversity and contrast are favored and a focus on relevant and useful information is preferred.

Chapter 6

Conclusion

The increasing reliance on NLP for decision support has highlighted the need for effective contrastive explanations, particularly in contexts that involve nuanced, query-driven decisions. While LLMs have revolutionized the field, challenges persist in generating high-quality, query-driven contrastive explanations that balance contrast, relevance, diversity, and usefulness. To address these gaps, we introduced a novel debate-style prompting methodology aimed at leveraging LLMs to enhance contrastive explanations. The core research questions (RQs) addressed were:

- **RQ1**: Does debate-style prompting improve query-driven contrastive explanations?
- **RQ2**: Does the aggressiveness in debate-style prompting impact the quality of contrastive explanations?
- **RQ3**: Can we trust pairwise LLM Win Rate evaluation of query-driven contrastive explanations?

Our approach incorporated structured prompting stages, culminating in a debate component designed to elicit a deeper exploration of contrasting perspectives. The methodology was evaluated against existing baselines using diverse datasets, including both objective (WikiVoyage) and subjective (TripAdvisor reviews) sources.

6.1 Summary of Contributions

In this work, we provide the following contributions to address the research challenges:

6.1.1 RQ1: Debate-style prompting methodology

We demonstrated that debate-style prompting significantly enhances the quality of query-driven contrastive explanations compared to STRUM-LLM-inspired baselines. Across datasets focusing on travel destinations, restaurants, and hotels:

- Debate-style prompting outperformed baselines on contrast, diversity, relevancy, and usefulness criteria for subjective review datasets (restaurants and hotels), achieving Win Rates exceeding 80% in most cases.

- For objective data (TravelDest), debate prompting still showed improved performance over baselines, particularly for contrast and usefulness, but the results were less pronounced due to the limited diversity in source content.
- The debate-style methodology consistently captured richer pros and cons, with higher specificity and contrastiveness in its outputs.

6.1.2 RQ2: Impact of aggressiveness in debate-style prompting

The level of aggressiveness in debate-style prompting did not significantly influence the quality of contrastive explanations:

- For the hotels dataset, aggressive debate prompting outperformed both standard and nice prompting, achieving higher win rates across all evaluation criteria.
- For the other datasets (restaurants and TravelDest), the standard prompting version consistently outperformed both aggressive and nice variations, suggesting domain and context-dependent effects.
- Confidence intervals indicated that the differences, while present, were not statistically significant across all datasets, warranting further investigation.

6.1.3 RQ3: Reliability of LLM-based Win Rate evaluation

We validated the reliability of pairwise LLM Win Rate evaluation for assessing query-driven contrastive explanations through a user study:

- LLM-based Win Rate evaluations were closely aligned with human judgments, often exceeding the inter-rater agreement observed among human evaluators. Cohen's Kappa values for the LLM agreement with majority human votes were higher than the agreement among human evaluators themselves.
- The study revealed that humans preferred explanations with higher diversity, rich descriptors, and explicit pros and cons—traits that the debate-style methodology excels at.
- These findings further indicate that LLMs are a trustworthy tool for evaluating contrastive explanations, supporting their broader adoption in evaluation frameworks.

6.2 Future Directions

The future directions of this research seek to broaden the capabilities and applicability of our debate model, enhancing its versatility and impact across different domains and methodologies. Below, we outline key areas for future exploration:

- **Top-k Entity Comparisons**: In our current work, we focused on comparing just two entities at a time, consistent with the existing literature and the debate methodology introduced in Chapter 3. A potential extension would involve expanding this comparison to the top-k entities. This could allow for a more nuanced understanding of the relative merits among multiple

options, enhancing the practical utility of the model in real-world decision-making scenarios, where users often compare more than two entities.

- **Multi-Modal Inputs**: Incorporating multi-modal inputs such as images, videos, or other forms of media could enrich the decision-making process. For instance, in Chapter 4, we focused on datasets involving textual reviews and objective descriptions, like those from restaurants, hotels, and travel destinations. Including visual or auditory data could improve the depth and richness of the model's understanding and enhance its ability to provide contrastive explanations, particularly in domains like travel or product reviews.

- **Application to Other Domains**: While our experimentation in Chapter 4 primarily focused on restaurant and hotel reviews, as well as travel destinations using datasets such as TravelDest, future work could expand into other domains. Comparisons of products, hiking trails, or even abstract concepts could provide a broader understanding of the capabilities and limitations of our approach, potentially identifying unique challenges and solutions in these diverse areas.

- **Exploring Subjectivity**: In Chapters 4 & 5, we utilized both objective data from the TravelDest dataset and subjective review data from TripAdvisor to evaluate the performance of our debate methodology. Interestingly, subjective data proved to be more effective in generating compelling and user-aligned debates. However, an important avenue for further research involves analyzing how subjective information is interpreted by the debate model. Future work could investigate how differing or conflicting opinions within subjective data are managed, reconciled, or contrasted during debates. This exploration would help understand how subjective nuances are portrayed to users, particularly when reviews or opinions diverge.

- **Testing with Other LLMs**: The experiments in Chapter 5 utilized GPT-4o for evaluating the effectiveness of our debate model, along with a user study to verify our evaluation. It would be beneficial to explore how other large language models, such as Claude or LLaMA models, perform with the same debate and contrastive explanation methods. This exploration could highlight whether model-specific behaviors impact the quality of debates or if the methodology generalizes well across different LLMs.

- **Multi-Agent Debates**: The methodology explored in Chapter 3 involved a self-debate framework, where a single agent debates both sides of a comparison. Future work could explore the use of multi-agent debate, where multiple agents provide different perspectives. This approach is an existing area of research but has not been broadly applied to contrastive explanations [16, 36]. Applying multi-agent systems to enhance these tasks could result in more dynamic and robust argumentative outputs, providing diverse viewpoints that a single agent might not capture.

- **Investigate Ordering Bias**: When performing LLM pairwise evaluation as discussed in Chapter 5,there is a potential bias introduced by the ordering in which different methodologies are presented. An insightful line of inquiry would be to determine whether the LLM consistently selects the same "winner" irrespective of the presentation order. If the choice varies with ordering, it would be essential to explore ways to mitigate this bias and understand the underlying reasons behind it.

- **Investigate User Study**: The user study outlined in Chapter 5 reported notably low Kappa scores, both among human evaluators themselves and between humans and LLMs. Future investigations could focus on the difficulty of the evaluation task and potential strategies for improvement. For example, it may be worth examining whether the low inter-rater agreement is partly driven by statistical noise and could be improved by recruiting a larger or more specialized pool of participants. Moreover, the impact of subjectivity in evaluating explanation quality should be carefully studied, including whether alternative or more structured evaluation frameworks might yield higher agreement among evaluators.

In this thesis, we have introduced a novel debate framework for generating contrastive explanations, focusing specifically on two-entity comparisons across various domains. Our methodology leverages prompting large language models to perform debate, demonstrating significant potential for enhancing user decision-making. The work presented here lays a strong foundation for future research, offering a structured approach that can be expanded and adapted to diverse settings to further enhance its utility and impact.

Bibliography

[1] Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. ASPECTNEWS: Aspect-oriented summarization of news documents. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[2] Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. Aspect-controllable opinion summarization. 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 6578–6593, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[3] Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293, 2021.

[4] Alejandro Ariza-Casabona, Ludovico Boratto, and Maria Salamó. A comparative analysis of text-based explainable recommender systems. In *Proceedings of the 18th ACM Conference on Recommender Systems*, RecSys '24, page 105–115, New York, NY, USA, 2024. Association for Computing Machinery.

[5] Lochan Basyal and Mihir Sanghvi. Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models. *arXiv preprint arXiv:2310.10449*, 2023.

[6] 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.

[7] Alessandro Castelnovo, Riccardo Crupi, Nicolò Mombelli, Gabriele Nanino, and Daniele Regoli. Evaluative item-contrastive explanations in rankings, 2023.

[8] Hanxiong Chen, Xu Chen, Shaoyun Shi, and Yongfeng Zhang. Generate natural language explanations for recommendation. In *SIGIR 2019 Workshop on ExplainAble Recommendation and Search*, 2021.

[9] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *IJCNLP (Findings)*, pages 361–374, 2023.

[10] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[11] Raffel Colin. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140–1, 2020.

[12] 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. Chain-of-verification reduces hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[14] J.P. Dillard and L. Shen. *The SAGE Handbook of Persuasion*. Sage Handbooks. SAGE Publications, 2013.

[15] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. A survey on in-context learning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[16] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023.

[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, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, 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, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, 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, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss,

Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024.

[18] Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*,

2024.

[19] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacafarm: A simulation framework for methods that learn from human feedback, 2023.

[20] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should i explain? a comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4):367–382, 2014.

[21] Lukas Gienapp, Harrisen Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. Evaluating generative ad hoc information retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 1916–1929. ACM, July 2024.

[22] H. Paul Grice. Logic and conversation. In Maite Ezcurdia and Robert J. Stainton, editors, *The Semantics-Pragmatics Boundary in Philosophy*, page 47. Broadview Press, 2013.

[23] Beliz Gunel, Sandeep Tata, and Marc Najork. Strum: Extractive aspect-based contrastive summarization. In *Companion Proceedings of the ACM Web Conference 2023*, WWW '23 Companion, page 28–31, New York, NY, USA, 2023. Association for Computing Machinery.

[24] Beliz Gunel, James B. Wendt, Jing Xie, Yichao Zhou, Nguyen Vo, Zachary Fisher, and Sandeep Tata. Strum-llm: Attributed and structured contrastive summarization, 2024.

[25] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. Effects of interactivity and presentation on review-based explanations for recommendations. In Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen, editors, *Human-Computer Interaction – INTERACT 2021*, pages 597–618, Cham, 2021. Springer International Publishing.

[26] Diana C Hernandez-Bocanegra and Jürgen Ziegler. Explaining recommendations through conversations: Dialog model and the effects of interface type and degree of interactivity. *ACM Transactions on Interactive Intelligent Systems*, 13(2):1–47, 2023.

[27] Diana C. Hernandez-Bocanegra and Jürgen Ziegler. Explaining recommendations through conversations: Dialog model and the effects of interface type and degree of interactivity. *ACM Trans. Interact. Intell. Syst.*, 13(2), April 2023.

[28] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[29] Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 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*, pages 113–122, 2021.

[30] Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. Chain-of-symbol prompting elicits planning in large langauge models, 2024.

[31] Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. Comparative opinion summarization via collaborative decoding. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[32] 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.

[33] Kevin Lerman and Ryan McDonald. Contrastive summarization: an experiment with consumer reviews. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics, companion volume: Short papers*, pages 113–116, 2009.

[34] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[35] Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *International Conference on Learning Representations*, 2024.

[36] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[37] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[38] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), January 2023.

[39] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, December 2023. Association for Computational Linguistics.

[40] Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. Aligning with human judgement: The role of pairwise preference in large language model evaluators, 2024.

[41] Sebastian Lubos, Thi Ngoc Trang Tran, Alexander Felfernig, Seda Polat Erdeniz, and Viet-Man Le. Llm-generated explanations for recommender systems. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*, pages 276–285, 2024.

[42] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[43] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 06 2017.

[44] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez,

Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[46] Richard E. Petty and John T. Cacioppo. *The Elaboration Likelihood Model of Persuasion*, pages 1–24. Springer New York, New York, NY, 1986.

[47] Pearl Pu and Li Chen. Trust building with explanation interfaces. In *Proceedings of the 11th international conference on Intelligent user interfaces*, pages 93–100, 2006.

[48] 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.

[49] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3:333–389, 01 2009.

[50] Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2024.

[51] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., USA, 1986.

[52] Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. Explaining legal concepts with augmented large language models (gpt-4), 2023.

[53] Matthias Schildwächter, Alexander Bondarenko, Julian Zenker, Matthias Hagen, Chris Biemann, and Alexander Panchenko. Answering comparative questions: Better than ten-blue-links? In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, page 361–365, New York, NY, USA, 2019. Association for Computing Machinery.

[54] Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. ProxyQA: An alternative framework for evaluating long-form text generation with large language models. 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 6806–6827, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[55] Nava Tintarev and Judith Masthoff. Effective explanations of recommendations: user-centered design. In *Proceedings of the 2007 ACM conference on Recommender systems*, pages 153–156, 2007.

[56] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[57] Thi Ngoc Trang Tran, Viet Man Le, Muesluem Atas, Alexander Felfernig, Martin Stettinger, and Andrei Popescu. Do users appreciate explanations of recommendations? an analysis in the movie domain. In *Proceedings of the 15th ACM Conference on Recommender Systems*, RecSys '21, page 645–650, New York, NY, USA, 2021. Association for Computing Machinery.

[58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[59] Hongning Wang, Yue Lu, and Chengxiang Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '10, page 783–792, New York, NY, USA, 2010. Association for Computing Machinery.

[60] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is ChatGPT a good NLG evaluator? a preliminary study. In Yue Dong, Wen Xiao, Lu Wang, Fei Liu, and Giuseppe Carenini, editors, *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 1–11, Singapore, December 2023. Association for Computational Linguistics.

[61] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. 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 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[62] Zilong Wang, Hao Zhang, Chun-Liang Li, Julian Martin Eisenschlos, Vincent Perot, Zifeng Wang, Lesly Miculicich, Yasuhisa Fujii, Jingbo Shang, Chen-Yu Lee, and Tomas Pfister. Chain-of-table: Evolving tables in the reasoning chain for table understanding, 2024.

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[64] Qianfeng Wen, Yifan Liu, Joshua Zhang, George Saad, Anton Korikov, Yury Sambale, and Scott Sanner. Elaborative subtopic query reformulation for broad and indirect queries in travel destination recommendation. In *The 1st Workshop on Risks, Opportunities, and Evaluation of Generative Models in Recommender Systems (ROEGEN@RecSys 2024)*, 2024.

[65] Ziyou Yan. Evaluating the effectiveness of llm-evaluators (aka llm-as-judge). *eugeneyan.com*, Aug 2024.

[66] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: deliberate problem solving with large language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[67] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[68] Yao Yao, Zuchao Li, and Hai Zhao. GoT: Effective graph-of-thought reasoning in language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2901–2921, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[69] Hang Yu and Jiawei Han. Survey of query-based text summarization. *arXiv preprint arXiv:2211.11548*, 2022.

[70] Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[71] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[72] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2023.

[73] Xufeng Zhao, Mengdi Li, Wenhao Lu, Cornelius Weber, Jae Hee Lee, Kun Chu, and Stefan Wermter. Enhancing zero-shot chain-of-thought reasoning in large language models through logic. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6144–6166, Torino, Italia, May 2024. ELRA and ICCL.

[74] Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. Towards a unified multi-dimensional evaluator for text generation. In Yoav

Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.