

Q-STRUM Debate: Query-Driven Contrastive Summarization for Recommendation Comparison

George-Kirollos Saad

University of Toronto
Toronto, Ontario, Canada
g.saad@mail.utoronto.ca

Scott Sanner

University of Toronto
Toronto, Ontario, Canada
ssanner@mie.utoronto.ca

Abstract

Query-driven recommendation with unknown items poses a challenge for users to understand why certain items are appropriate for their needs. Query-driven Contrastive Summarization (QCS) is a methodology designed to address this issue by leveraging language-based item descriptions to clarify contrasts between them. However, existing state-of-the-art contrastive summarization methods such as STRUM-LLM fall short of this goal. To overcome these limitations, we introduce Q-STRUM Debate, a novel extension of STRUM-LLM that employs debate-style prompting to generate focused and contrastive summarizations of item aspects relevant to a query. Leveraging modern large language models (LLMs) as powerful tools for generating debates, Q-STRUM Debate provides enhanced contrastive summaries. Experiments across three datasets demonstrate that Q-STRUM Debate yields significant performance improvements over existing methods on key contrastive summarization criteria, thus introducing a novel and performant debate prompting methodology for QCS.

1 Introduction

In query-driven recommendation settings such as hotels, restaurants, or travel, where items may be *a priori* unknown to users, language-based item descriptions can help users make informed choices. However, understanding the trade-offs between choices becomes challenging given the abundance of information from objective sources, like Wikipedia or travel guides, and opinion-rich subjective sources, such as TripAdvisor and Amazon reviews (Gunel et al., 2024; Wen et al., 2024).

Fortunately, Query-driven Contrastive Summarization (QCS) offers a principled solution to these challenges by providing succinct comparative summaries of items. However, many traditional QCS methods often rely on complex extraction, ranking, and diversification algorithms that may fail to

find clear contrasts, leaving users to sift through extensive information (Ströhle et al., 2023).

Fortunately, the emergence of large language models (LLMs) has revolutionized QCS capabilities (Colin, 2020; Angelidis et al., 2021; Chowdhery et al., 2023). By distilling relevant descriptive and review content into concise comparisons, state-of-the-art LLM-based contrastive summarization methods such as STRUM-LLM (Gunel et al., 2024) enable users to comparatively evaluate choices via summaries grounded in concrete data and clear comparisons that are important for decision-making (Lubos et al., 2024; Pu and Chen, 2006). While these LLM-driven approaches arguably improve on their pre-LLM predecessors, they often fall short of their contrastive summary potential as we show in our comparative empirical evaluation.

This work addresses a critical gap in QCS by leveraging LLMs to prioritize contrast while maintaining relevance and groundedness as motivated by principles of debate that are founded on discourse theory, including the Elaboration Likelihood Model (ELM) (Petty and Cacioppo, 1986) and Grice’s Maxims (Grice, 2013). Building on the state-of-the-art contrastive summarization STRUM-LLM framework (Gunel et al., 2024), we propose that aspect-based *debate prompting* provides a natural framework for improved QCS that we term Q-STRUM Debate. An example output summary demonstrating this approach is shown in Figure 1.

We summarize our key contributions as follows:

1. We provide a novel *debate prompting* mechanism to improve contrastiveness in QCS.
2. We show the resulting Q-STRUM Debate matches or outperforms base STRUM-LLM and a contrastive prompt extension on three domains (hotels, restaurants, and travel).
3. We modulate debate prompt aggressiveness and evaluate its impact on summary quality.

	Bangkok	Melbourne
Global Culinary Influences	Chinese culinary influences are significant due to the city's large Thai-Chinese population. [33]	The Italian population and culture have left a lasting impact on Melbourne's dining scene. [40]
Unique Dining Experiences	Street food in Bangkok offers an authentic and enriching cultural experience. [33]	Melbourne's coffee culture is a notable part of its culinary identity, rooted in Italian migration. [28]

Figure 1: Example of an aspect-based contrastive summary for the query: “culinary cities for food lovers”

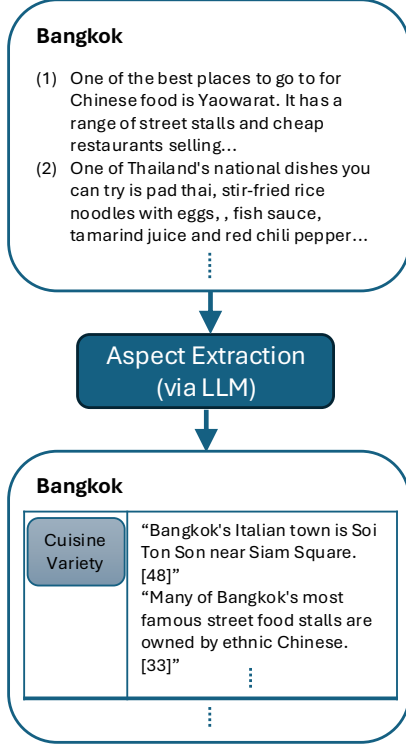


Figure 2: *Aspect Extraction* stage for the query: “culinary cities for food lovers”

2 STRUM for Contrastive Summarization

Recommendations often involve presenting users with multiple options, requiring methods that clearly articulate how and why each option aligns with their preferences. STRUM (Gunel et al., 2023) introduced a seminal and foundational approach to contrastive summarization by leveraging entailment models and hierarchical clustering to extract, merge, and contrast aspects of data. While effective for structured summarization, STRUM faced significant limitations in producing outputs that were sufficiently contrastive and aligned with user-specific queries.

To address these challenges, STRUM-LLM

(Gunel et al., 2024) integrated large language models (LLMs) to improve attribute extraction, aspect merging, and summarization. The architecture is outlined in Figure 6(a). STRUM-LLM employs several LLM-driven components:

- **Aspect Extraction:** Identifies aspects and relevant values from source data while attributing them to their origins. An example of this stage is provided in Figure 2.
- **Aspect Merge:** Combines similar aspects to reduce redundancy. An example of this stage is provided in Figure 3.
- **Value Merge:** Consolidates consistent values for each aspect based on majority opinion.
- **Contrastive Summarizer:** Highlights the most significant and contrasting aspects.
- **Usefulness:** Filters out less useful aspects and identifies errors.

While STRUM-LLM provides a state-of-the-art methodology for contrastive summarization, it is not query-driven as originally defined. More critically, we also empirically show that the STRUM-LLM methodology falls short of providing strongly contrastive summaries. To address these gaps, we next introduce **Q-STRUM Debate**, which builds upon STRUM-LLM to deliver query-driven, contrastive summarization. Central to this improvement will be the introduction of *debate prompting*.

3 Q-STRUM for QCS

To address the limitations of STRUM-LLM and enable a query-driven system for contrastive summarization, we propose **Q-STRUM**. This method ensures that the query is passed through all stages of the architecture to generate highly relevant, contrastive outputs tailored to user needs.

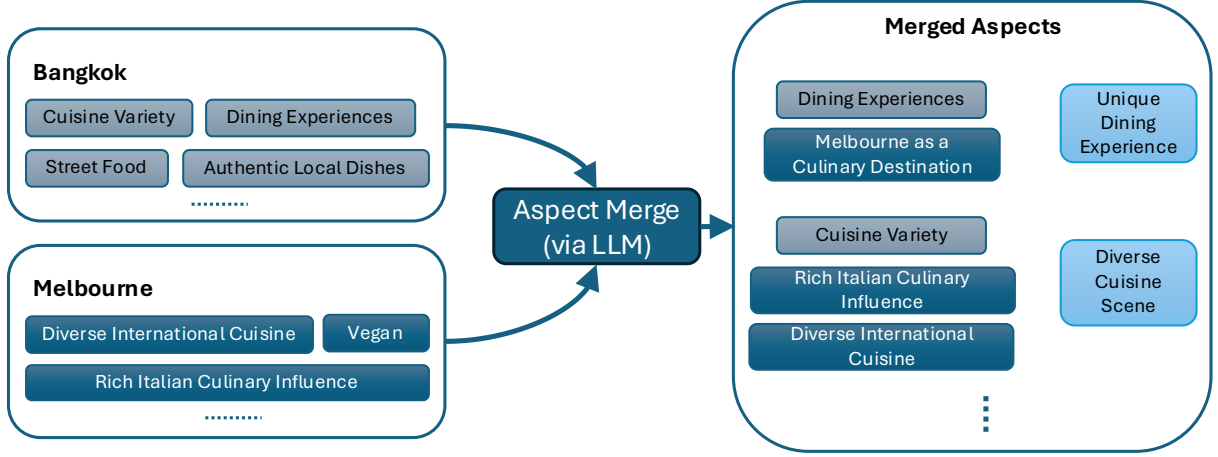


Figure 3: *Aspect Merge* stage for the query: “culinary cities for food lovers”

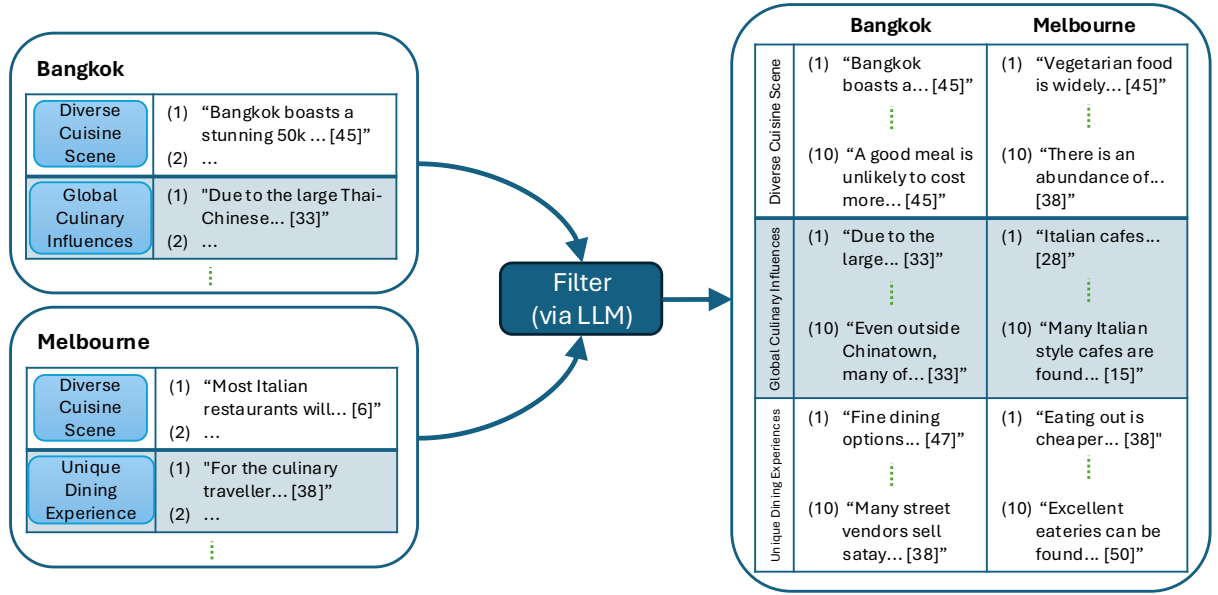


Figure 4: *Filter* stage for the query: “culinary cities for food lovers”

3.1 Pipeline Overview

Q-STRUM employs a structured pipeline with four stages: Aspect Extraction, Aspect Merge, Filter, and Summarizer. Each stage uses the query as an anchor, ensuring alignment with user intent. The Filter stage, unique to Q-STRUM, selects the top three aspects most relevant to the query and extracts exactly 10 concise, informative phrases for each. This reduces noise and redundancy, enabling downstream stages to focus on concise, length-controlled content. Figure 4 outlines an example of the Filter stage. These modifications make the Value Merge and Usefulness stages from STRUM-LLM redundant and hence allow us to remove them from the pipeline. Figures 5 and 6 illustrate the output and architecture comparisons, respectively.

3.2 Contrastive Prompting

The Base Summarizer uses a monolithic prompt to produce a general summary of the extracted aspects from the Filter stage. This approach provides a simple, high-level overview of the data. The Contrastive Summarizer builds on this by explicitly instructing the LLM to “*identify the most contrasting and important values.*” This simple yet effective addition produces more detailed and relevant contrastive outputs (Gunel et al., 2024). Prompts for both the Base and Contrastive Summarizers are provided in Appendix C.

3.3 Debate Prompting

Debate prompting introduces a multi-stage process to address the shortcomings of monolithic summa-

Query: Culinary cities for food lovers

Q-STRUM Base

Bangkok

Melbourne

Global Culinary Influences	<p>(1) Due to the large Thai-Chinese population in Bangkok, Chinese influences pervade much of Bangkok's culinary scene. [33]</p> <p>(2) Hua Lamphong Railway Station, Bangkok's main railway station, was designed by Italian architects. [31]</p> <p>(3) Many people working in the service industry have a basic grasp of English, as do staff at many of the popular food stalls. [19]</p>	<p>(1) With its large Italian population Melbourne has countless Italian restaurants, mostly offering food from the southern regions of the Italian peninsula. [40]</p> <p>(2) Melbourne has a long and rich coffee culture beginning with Victorian era coffee palaces and further enhanced by Italian migrants arriving in the aftermath of World War II. [33]</p> <p>(3) Lygon Street is where Melbourne's coffee culture originated. [28]</p>
----------------------------	--	---

Q-STRUM Contrastive

Global Culinary Influences	<p>(1) Due to the large Thai-Chinese population, Chinese influences are prominent in Bangkok's culinary scene. [33]</p> <p>(2) Bangkok's Italian town, Soi Ton Son, highlights the global influences in the city's food culture. [48]</p> <p>(3) European-style architecture and international staff contribute to an international dining atmosphere. [31, 19]</p>	<p>(1) Melbourne has a rich coffee culture, heavily influenced by Italian migrants post-World War II. [33]</p> <p>(2) Carlton's Lygon Street is noted for its concentration of Italian cafes and restaurants. [28]</p> <p>(3) Melbourne draws culinary influences from a global population, being a magnet for migrants worldwide. [31]</p>
----------------------------	---	---

Q-STRUM Debate

Global Culinary Influences	<p>(1) Bangkok offers a vast selection of international cuisines, represented by 50,000 places to eat. [45]</p> <p>(2) Chinese culinary influences are significant due to the city's large Thai-Chinese population. [33]</p> <p>(3) A diverse fusion of global cuisines is available, including French, Lebanese, Mexican, and Vietnamese. [48]</p>	<p>(1) Melbourne boasts a wide array of high-quality restaurants representing various global cuisines. [50]</p> <p>(2) The Italian population and culture have left a lasting impact on Melbourne's dining scene. [40]</p> <p>(3) Chinese culinary traditions are well-established, offering a diverse range of dining options. [9]</p>
----------------------------	---	---

Figure 5: Comparison of Outputs of All Q-STRUM Methodologies

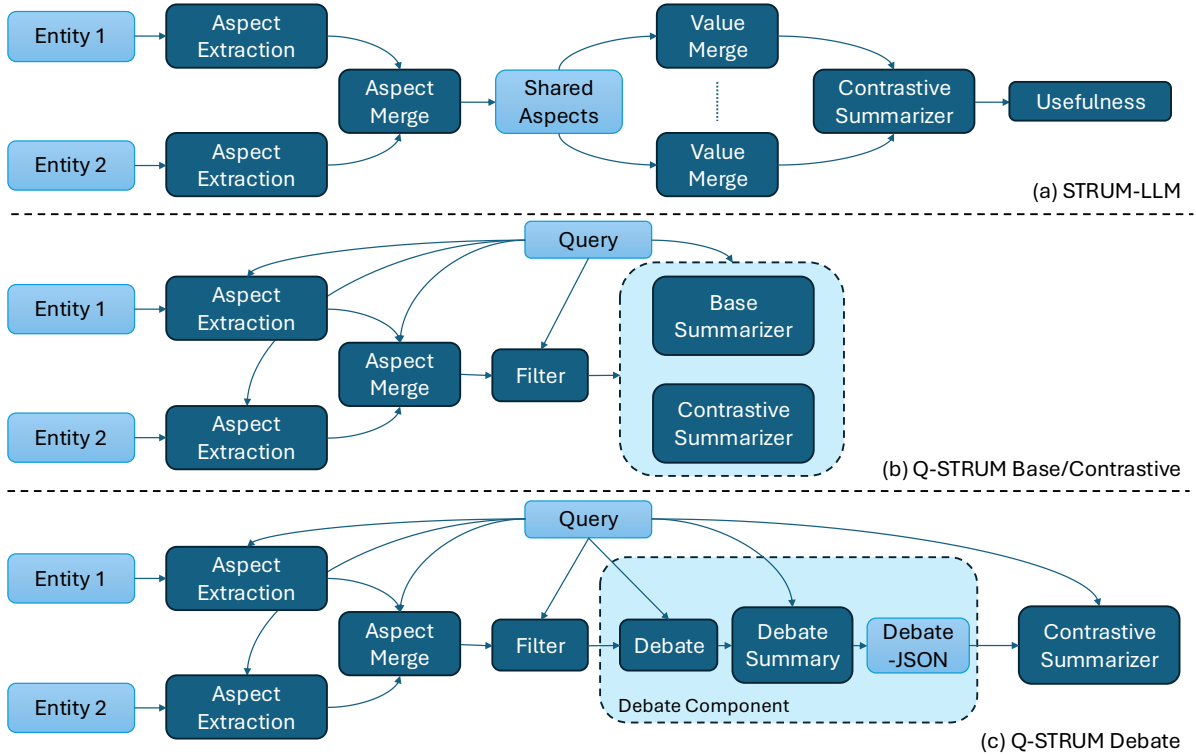


Figure 6: STRUM-LLM vs. Q-STRUM Architectures

rization. Recent work demonstrates that inter-LLM debates can produce more truthful answers by leveraging structured argumentation (Khan et al., 2024). In our novel Q-STRUM Debate for query-driven contrastive summarization, we simulate debate-structured disagreement between two personas to surface complementary and opposing points that

are then distilled by the final summarizer.

Unlike Base and Contrastive summarizers, which rely on a single prompt, Debate prompting divides the task into distinct stages:

- **Debate Stage:** The LLM simulates a structured argument between two perspectives (e.g., Alice and Bob), where each defends

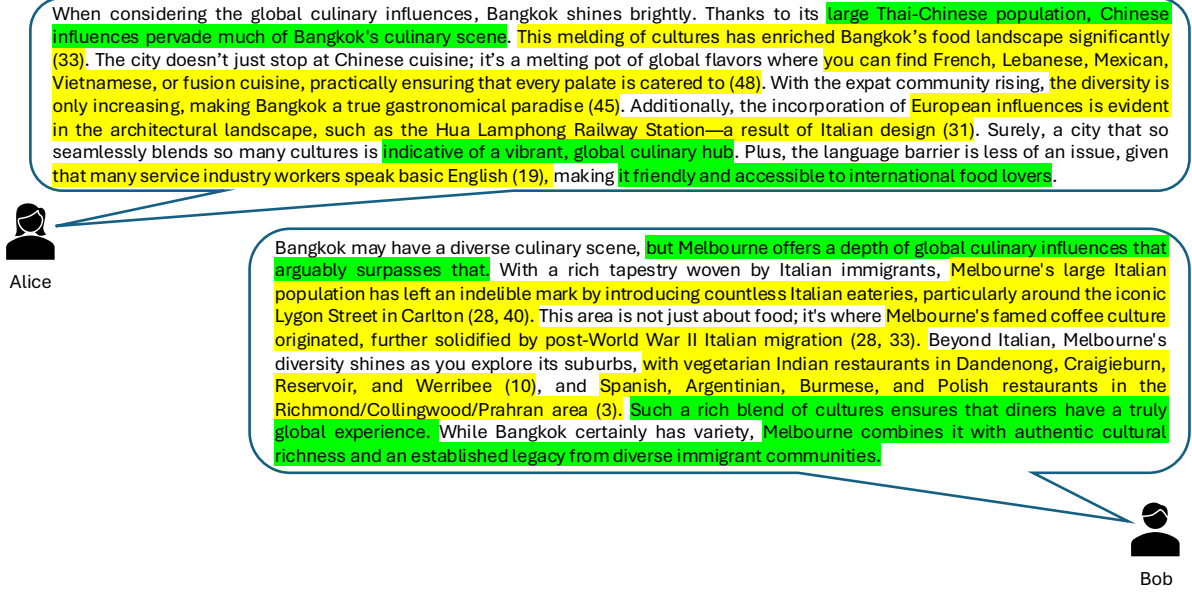


Figure 7: *Debate* example for the query: “culinary cities for food lovers”

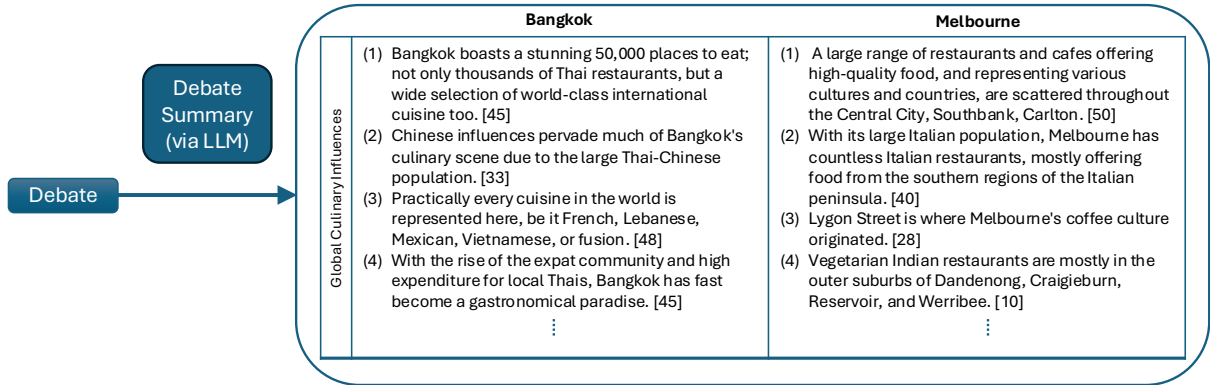


Figure 8: *Debate Summary* example for the query: “culinary cities for food lovers”

one entity while addressing the other’s weaknesses. This ensures balanced, contrastive comparisons by emphasizing both pros and cons. An example of a debate for the query:

“*Culinary cities for food lovers*”

for destinations Bangkok and Melbourne is provided in Figure 7. Highlights include pros and cons (green) and references to source data (yellow). The prompt in Listing 1 is called for each aspect and its respective 10 sentences output from the Filter stage (cf. Figure 4).

- **Debate Summary Stage:** The output from the Debate Stage is summarized and then formatted into a structured JSON representation, called Debate-JSON. This step ensures that the information is well-organized and explicitly aligned with the query and extracted as-

pects. An example of the output of this stage for the same debate in Figure 7, can be found in Figure 8. The prompt used for this can be found in Listing 2.

- **Final Contrastive Summary:** The structured data from the Debate-JSON is processed using the same Contrastive Summarizer prompt from Section 3.2.

This multi-stage process ensures that the outputs are not only highly contrastive but also grounded and well-structured. Figure 6(b) and (c) depicts the architecture of Q-STRUM Debate alongside Base and Contrastive methodologies.

4 Datasets

To evaluate Q-STRUM Debate, we used three query-driven entity recommendation datasets with

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 1: Summary of Datasets

Listing 1: LLM Prompt for Debate Stage

```

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.

```

diverse and comprehensive query-entity pairs, containing detailed textual data relevant to queries. We aimed to experiment with both objective entity descriptions as well as highly subjective review-based entity opinions. Full data preprocessing details are provided in Appendix A. Table 1 summarizes key statistics of the following datasets:

- **TravelDest:** Objective (factual, non-review) travel destination descriptions from WikiVoyage. Example query: “Top cities for music lovers”.
- **Restaurants:** Subjective restaurant reviews from TripAdvisor. Example query: “I want a romantic restaurant with views of the city”.
- **Hotels:** Subjective hotel reviews from TripAdvisor. Example query: “Find me a family-friendly hotel with enriching activities for kids”.

Listing 2: LLM Prompt for Debate Summary Stage

```

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 #]"
}

```

5 Experimental Design

We compare our novel **Q-STRUM Debate** (Section 3.3) to the query-driven STRUM-LLM extension **Q-STRUM Base** (Section 3.1) and its **Q-STRUM Contrastive** extension (Section 3.2) — all using GPT-4o — to address two key research questions:

- **RQ1:** Does debate-style prompting improve query-driven contrastive summaries?
- **RQ2:** Does the aggressiveness (or niceness)

level in debate prompting impact the quality of contrastive summaries?

5.1 RQ1: Win Rate Evaluation Metrics

We evaluate Q-STRUM and its baselines using a pairwise LLM Win Rate evaluation approach, leveraging GPT-4o and Claude-3.5-Sonnet to compare summary outputs. We bidirectionally tested both A vs. B and B vs. A to control for potential LLM ordering bias in winner evaluation. This methodology aligns well with human judgments and is effective for subjective assessment tasks such as explanation evaluation (Liu et al., 2023, 2024; Wang et al., 2024). Pairwise evaluation allows for nuanced comparisons, determining a “winner” for each summary based on established criteria.

The Win Rate for Method A vs B is defined as:

$$\text{Win Rate}_A = \frac{\text{times A wins} + 0.5 \times \text{ties}}{\text{Total Comparisons Made}} \times 100\% \quad (1)$$

The evaluation focuses on four key criteria, derived from the existing literature:

Contrastiveness. Summaries should effectively highlight differences, emphasizing pros and cons to help users make decisions (Miller, 2017; Castelnovo et al., 2023).

Relevancy. Outputs must align with the query and address user-specific needs (Castelnovo et al., 2023; Miller, 2017).

Diversity. Summaries should provide a variety of points without repetition, offering multiple facets of comparison (Gienapp et al., 2024; Castelnovo et al., 2023).

Usefulness. Summaries must be informative and help users in decision-making (Lubos et al., 2024; Hernandez-Bocanegra and Ziegler, 2023).

For each query, outputs from Q-STRUM and its baselines are compared pairwise across these criteria. The LLM determines a winner for each criterion, or declares a tie, and the results are aggregated into Win Rates. To enhance evaluation quality, the LLM is prompted to justify its decisions, as reasoning summaries are shown to improve consistency (Zhang et al., 2023). The prompt used for this evaluation is provided in Appendix C.

This approach ensures reliable, consistent, and scalable evaluation of all variations of Q-STRUM in a manner aligned with established standards for evaluating both summary and explanation quality (Castelnovo et al., 2023; Miller, 2017).

5.2 RQ2: Aggressiveness Analysis

This analysis examined whether varying the tone and assertiveness of debate-style prompting impacts summarization quality. Three prompt variations were tested: ‘nice,’ ‘aggressive,’ and the standard neutral version. The ‘nice’ prompt instructed: “Alice and Bob should both be nice and polite to each other.” The ‘aggressive’ prompt instructed: “Alice and Bob should both be aggressive and assertive with each other.” All other data and prompt inputs remained consistent across variations.

The standard Q-STRUM Debate was compared to Q-STRUM Debate with the modified ‘aggressive’ and ‘nice’ prompt versions defined above. Evaluation followed the same pairwise Win Rate comparison over contrast, relevancy, diversity, and usefulness as in Section 5.1. Bidirectional Win Rate evaluation mitigated potential bias from aggressiveness and comparison-order interactions.

6 Experimental Results

Below, we summarize key findings for our previous research questions. Experiments used ~ 11 M tokens of GPT-4o (est. ~ 200 B parameters (Abacha et al., 2025)) API calls and ~ 3 M tokens of Claude-3.5-Sonnet (175B parameters) API calls. All code and data to reproduce these results are provided in a public GitHub code repository.¹

6.1 RQ1: Pairwise Win Rate Evaluation

Across all datasets and criteria, Q-STRUM Debate outperformed Q-STRUM Base. For the subjective review datasets, Restaurants and Hotels, Debate achieved Win Rates mostly at or above 80% for contrast, diversity, and usefulness, as shown in Table 2. However, the relevance criterion, while still favoring Debate, had lower Win Rates, at or above 57%, however all confidence intervals were still above 50%. The results for Claude-3.5-Sonnet in Table 3 (and DeepSeek-v3 in Appendix D) are similar. Overall, while both debate and baseline summaries are relevant, Q-STRUM Debate generally offers greater contrast, diversity, and usefulness.

For the objective TravelDest dataset, Q-STRUM Debate demonstrated a narrower margin of superiority. The same tables highlight that Q-STRUM Debate (significantly) outperformed Q-STRUM Base in almost all criteria, but achieved mixed results against Q-STRUM Contrastive. This suggests that

¹<https://github.com/D3MLab/q-strum-debate>

Criterion	Restaurants		Hotels		TravelDest	
	Debate vs. Contrastive	Debate vs. Base	Debate vs. Contrastive	Debate vs. Base	Debate vs. Contrastive	Debate vs. Base
Contrast	0.85 [0.78, 0.91]	0.87 [0.81, 0.93]	0.82 [0.75, 0.88]	0.82 [0.75, 0.90]	0.64 [0.58, 0.71]	0.78 [0.73, 0.84]
Relevance	0.57 [0.51, 0.63]	0.57 [0.51, 0.62]	0.62 [0.55, 0.70]	0.59 [0.52, 0.66]	0.50 [0.46, 0.54]	0.56 [0.51, 0.60]
Diversity	0.83 [0.76, 0.91]	0.84 [0.77, 0.91]	0.80 [0.72, 0.88]	0.86 [0.79, 0.92]	0.54 [0.48, 0.61]	0.69 [0.63, 0.75]
Usefulness	0.83 [0.76, 0.90]	0.89 [0.83, 0.95]	0.78 [0.70, 0.86]	0.84 [0.77, 0.91]	0.61 [0.54, 0.68]	0.72 [0.66, 0.78]

Table 2: Pairwise LLM Win Rate (95% CIs in $[\cdot, \cdot]$) for Q-STRUM Debate vs. Q-STRUM Baselines (Contrastive, Base) across the Restaurants, Hotels, and TravelDest datasets using GPT-4o.

Criterion	Restaurants		Hotels		TravelDest	
	Debate vs. Contrastive	Debate vs. Base	Debate vs. Contrastive	Debate vs. Base	Debate vs. Contrastive	Debate vs. Base
Contrast	0.79 [0.71, 0.87]	0.87 [0.81, 0.92]	0.77 [0.69, 0.84]	0.80 [0.74, 0.87]	0.70 [0.64, 0.75]	0.75 [0.70, 0.79]
Relevance	0.63 [0.56, 0.69]	0.58 [0.53, 0.63]	0.62 [0.56, 0.69]	0.64 [0.57, 0.71]	0.50 [0.46, 0.54]	0.58 [0.54, 0.62]
Diversity	0.75 [0.67, 0.84]	0.84 [0.78, 0.91]	0.77 [0.69, 0.84]	0.81 [0.74, 0.88]	0.48 [0.42, 0.54]	0.53 [0.47, 0.59]
Usefulness	0.77 [0.70, 0.85]	0.85 [0.78, 0.92]	0.79 [0.71, 0.86]	0.84 [0.77, 0.91]	0.54 [0.47, 0.60]	0.63 [0.57, 0.69]

Table 3: Pairwise LLM Win Rate (95% CIs in $[\cdot, \cdot]$) for Q-STRUM Debate vs. Q-STRUM Baselines (Contrastive, Base) across the Restaurants, Hotels, and TravelDest datasets using Claude-3.5-Sonnet.

Criterion	Restaurants		Hotels		TravelDest	
	Aggressive	Nice	Aggressive	Nice	Aggressive	Nice
Contrast	0.57 [0.48, 0.66]	0.56 [0.46, 0.65]	0.48 [0.33, 0.63]	0.57 [0.49, 0.65]	0.53 [0.47, 0.59]	0.47 [0.41, 0.53]
Relevance	0.53 [0.48, 0.58]	0.51 [0.45, 0.56]	0.46 [0.37, 0.56]	0.56 [0.49, 0.63]	0.50 [0.46, 0.53]	0.50 [0.47, 0.53]
Diversity	0.54 [0.45, 0.63]	0.57 [0.48, 0.66]	0.41 [0.26, 0.55]	0.55 [0.46, 0.63]	0.53 [0.47, 0.59]	0.47 [0.41, 0.53]
Usefulness	0.58 [0.49, 0.67]	0.58 [0.48, 0.68]	0.43 [0.27, 0.58]	0.57 [0.48, 0.65]	0.54 [0.47, 0.60]	0.45 [0.39, 0.51]

Table 4: Pairwise LLM Win Rate (95% CIs in $[\cdot, \cdot]$) for Q-STRUM Debate (Standard) vs. Q-STRUM Debate (Aggressive and Nice) across the Restaurants, Hotels, and TravelDest datasets using GPT-4o.

Criterion	Restaurants		Hotels		TravelDest	
	Aggressive	Nice	Aggressive	Nice	Aggressive	Nice
Contrast	0.55 [0.46, 0.65]	0.58 [0.49, 0.67]	0.50 [0.41, 0.59]	0.58 [0.49, 0.67]	0.53 [0.46, 0.59]	0.54 [0.47, 0.61]
Relevance	0.52 [0.46, 0.57]	0.55 [0.49, 0.60]	0.47 [0.41, 0.53]	0.55 [0.49, 0.62]	0.50 [0.46, 0.53]	0.50 [0.46, 0.53]
Diversity	0.55 [0.47, 0.64]	0.49 [0.39, 0.59]	0.47 [0.39, 0.55]	0.59 [0.50, 0.68]	0.58 [0.52, 0.64]	0.53 [0.47, 0.59]
Usefulness	0.53 [0.43, 0.63]	0.56 [0.46, 0.65]	0.54 [0.45, 0.63]	0.62 [0.52, 0.72]	0.57 [0.51, 0.63]	0.54 [0.48, 0.60]

Table 5: Pairwise LLM Win Rate (95% CIs in $[\cdot, \cdot]$) for Q-STRUM Debate (Standard) vs. Q-STRUM Debate (Aggressive and Nice) across the Restaurants, Hotels, and TravelDest datasets using Claude-3.5-Sonnet.

Q-STRUM Debate prompting may be more effective for the *subjective, opinion-rich* TripAdvisor review data in Restaurants and Hotels than the *objective, fact-oriented* WikiVoyage data of TravelDest.

6.2 RQ2: Aggressiveness Analysis

The aggressiveness analysis compared ‘standard’, ‘aggressive’, and ‘nice’ debate prompts. As shown in Tables 4 and 5, the standard prompt generally performed best across datasets. The aggressive prompt showed marginal improvements in specific contexts, such as the Hotels dataset, but did not demonstrate consistent advantages elsewhere.

We conjecture that while niceness vs. aggressiveness does affect the subjective debate style and verbiage, this does not ultimately affect the core objective content being contrasted (as one can verify from the examples in Appendix B), hence having minimal impact on the results.

7 Conclusion

We introduced a novel debate-style prompting framework called Q-STRUM Debate to generate high-quality, query-driven contrastive summaries using LLMs. We demonstrated that debate-style prompting significantly outperforms baselines, particularly for subjective review datasets, by delivering more contrastive, relevant, diverse, and useful summaries. Experiments modulating debate aggressiveness showed marginal impact on results.

Future directions include multiple entity comparisons, incorporating multi-modal review content, and adapting it to diverse domains such as medical products or hiking trails. Further exploration of multi-agent debate methodology could enhance the framework’s versatility. Overall, these extensions can build on Q-STRUM’s novel debate prompting methodology to improve contrastiveness in a variety of query-driven summarization applications.

8 Limitations

While our research introduces a robust debate model for comparative analysis, certain limitations remain that present opportunities for refinement. First, the current model is limited to pairwise comparisons of entities, which, although effective, may not fully capture the complexity of real-world decision-making scenarios where users often consider multiple options simultaneously. Additionally, our focus on textual datasets restricts the model's applicability to domains where multi-modal data (e.g., images, videos) play a critical role in user decision-making, such as in product or travel reviews.

While subjective review data enhanced the effectiveness of debates, the interpretation of conflicting subjective opinions remains an open challenge. Understanding how the model reconciles divergent viewpoints is important for improving the depth and fairness of contrastive summaries.

9 Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

References

- Asma Ben Abacha, Wen wai Yim, Yujian Fu, Zhaoyi Sun, Meliha Yetisgen, Fei Xia, and Thomas Lin. 2025. [Medec: A benchmark for medical error detection and correction in clinical notes](#). *Preprint*, arXiv:2412.19260.
- Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. [Extractive opinion summarization in quantized transformer spaces](#). *Transactions of the Association for Computational Linguistics*, 9:277–293.
- Alessandro Castelnovo, Riccardo Crupi, Nicolò Mombelli, Gabriele Nanino, and Daniele Regoli. 2023. [Evaluative item-contrastive explanations in rankings](#). *Preprint*, arXiv:2312.10094.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Raffel Colin. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140–1.
- Lukas Gienapp, Harrison Scells, Niklas Deckers, Janek Bevendorff, Shuai Wang, Johannes Kiesel, Shahbaz Syed, Maik Fröbe, Guido Zuccon, Benno Stein, Matthias Hagen, and Martin Potthast. 2024. [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.
- H. Paul Grice. 2013. Logic and conversation. In Maite Ezcurdia and Robert J. Stainton, editors, *The Semantics-Pragmatics Boundary in Philosophy*, page 47. Broadview Press.
- Aylin Gunal, Baihan Lin, and Djallel Bouneffouf. 2024. [Conversational topic recommendation in counseling and psychotherapy with decision transformer and large language models](#). In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 196–201, Mexico City, Mexico. Association for Computational Linguistics.
- Beliz Gunel, Sandeep Tata, and Marc Najork. 2023. [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. Association for Computing Machinery.
- Beliz Gunel, James B. Wendt, Jing Xie, Yichao Zhou, Nguyen Vo, Zachary Fisher, and Sandeep Tata. 2024. [Strum-llm: Attributed and structured contrastive summarization](#). *Preprint*, arXiv:2403.19710.
- Diana C Hernandez-Bocanegra and Jürgen Ziegler. 2023. 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.
- Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.
- Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. [Debating with more persuasive llms leads to more truthful answers](#). *Preprint*, arXiv:2402.06782.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024.

Aligning with human judgement: The role of pairwise preference in large language model evaluators. *Preprint*, arXiv:2403.16950.

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

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

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

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

Thomas Ströhle, Ricardo Campos, and Adam Jatowt. 2023. *Contrastive text summarization: a survey*. In *International Journal of Data Science and Analytics*, 18:353–367.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. *Large language models are not fair evaluators*. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Qianfeng Wen, Yifan Liu, Joshua Zhang, George Saad, Anton Korikov, Yury Sambale, and Scott Sanner. 2024. 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)*.

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

A Data Preprocessing

The preprocessing pipeline was designed to prepare the datasets for effective use in query-driven contrastive summarization (QCS). Key steps included:

TravelDest Dataset The TravelDest dataset includes detailed WikiVoyage descriptions for 774 global destinations. For preprocessing, we employed elaborative query reformulation (EQR) (Wen et al., 2024) to generate rich queries. Entities were ranked for each query using dense retrieval via TAS-B embeddings (Hofstätter et al., 2021).

Restaurants and Hotels Datasets For the subjective review datasets, we manually created natural language queries related to common user needs. We then scraped reviews from TripAdvisor for hotels and restaurants in Toronto. Dense retrieval was performed using OpenAI’s text-embedding-3-small model² to compute cosine similarity scores between queries and the review snippets.

Snippet Extraction For all three datasets, we extracted the top-50 relevant snippets for each entity, ensuring balanced representation. For TravelDest, these snippets were sourced from WikiVoyage, whereas for the restaurants and hotels datasets, the snippets were drawn from TripAdvisor reviews.

Scoring and Selection Entity relevance was determined using the arithmetic mean of the top-50 cosine similarity scores. For each query, the top two entities were selected as candidates for comparison, and their corresponding snippets served as input for the Q-STRUM pipeline.

This preprocessing approach ensures consistency, relevance, and high-quality textual inputs for evaluating Q-STRUM’s ability to generate contrastive summaries for diverse queries and data.

B Aggressiveness Analysis

Figure 9 shows an example comparison between ‘aggressive’, ‘nice’ and standard debate outputs as explored in RQ2 of Section 6.2. We see that ‘aggressive’ prompts take a more firm stance in their wording but cover similar content as the standard prompt. In contrast, the ‘nice’ prompt is noticeably more collegial, but also provides more limited argumentation as evidenced by the shorter length.

C Q-STRUM Prompts

We provide the full prompts required to implement the Q-STRUM pipeline components as follows:

- Listing 3: LLM Prompt for Aspect Extraction Stage in Figure 2 and Section 3.1.
- Listing 4: LLM Prompt for Aspect Merge Stage in Figure 3 and Section 3.1.
- Listing 5: LLM Prompt for Filter Stage in Figure 4 and Section 3.1.
- Listing 6: LLM Prompt for Contrastive Summarizer Stage in Figure 6(b) and Section 3.2.

²<https://openai.com/index/new-embedding-models-and-api-updates/>

- Listing 7: LLM Prompt for Pairwise Evaluation in Section 5.1 used for evaluation metrics.

The prompts for Q-STRUM-Debate were provided in the main paper in Listings 1 and 2 as discussed in Section 3.3.

D Additional Results

Table 6 contains the Pairwise Win Rate evaluation results using an open model, Deepseek-v3. The results are comparable to those run using GPT-4o and Claude-3.5-Sonnet.

Standard	Aggressive	Nice
<p>Alice: I believe The Ivy At Verity unquestionably stands out as the perfect spa destination for you. It's not just about having a spa like the Guerlain Spa at Hotel X, but about offering an entirely unique and exclusive atmosphere. The Ivy has the Sweetgrass Spa, which is specifically designed for women and offers a sense of exclusivity and privacy you won't find elsewhere. The spa is attached to a women's only networking club, adding to the overall ambiance and exclusivity. On top of that, the subterranean swimming pool and hot tub provide a relaxing, serene environment, perfect for unwinding [#36]. It's worth noting the Sweetgrass Spa is extensive, featuring a hot tub, steam room, and sauna, ensuring a comprehensive pampering experience [#40]</p>	<p>Alice: The Ivy at Verity is undoubtedly the best choice when prioritizing a spa. First and foremost, the Sweetgrass Spa, connected with the hotel, provides a unique, women-exclusive experience that other hotels, including Hotel X Toronto, can't compete with. It's more than just a spa; it's a comprehensive wellness haven featuring a hot tub, steam room, and sauna, specifically crafted for relaxation and rejuvenation (#40). Not to mention its subterranean swimming pool and hot tub, creating an unparalleled soothing ambiance (#36). The Ivy at Verity isn't merely about services; it's about the entire intimate, luxurious experience, exemplified by posh yet cozy elements, ensuring you leave feeling both pampered and regenerated (#28)</p>	<p>Alice: Bob, I completely understand your admiration for Hotel X Toronto By Library Hotel Collection's spa, but I believe The Ivy At Verity's spa offerings are truly exceptional. For starters, the Sweetgrass Spa has a range of services that make it quite special. As reviews say, guests have had "a wonderful stay and George restaurant is awesome - the ambiance was spectacular and our room was intimate and very quiet - the access to the spa services was a great addition to the enjoyment of our stay" [#41]. Plus, the spa itself is exclusive, as it's "exclusive for women" [#40]. It feels very private and unique.</p>
<p>Alice: When it comes to friendly and attentive staff, Tabule Middle Eastern Cuisine truly stands out. The service here is described as "excellent with every attention to detail" [19]. This level of meticulous care ensures that every need is promptly addressed. In contrast, over at Aroma Fine Indian Cuisine, while they do have attentive staff, there are also mentions of the staff not being "terribly friendly" [39], which could potentially diminish the overall dining experience. Alice: Furthermore, Tabule's staff are described as "attentive and always nearby" [8], which is crucial for a family dining experience, especially if you have kids who might need more attention. The atmosphere at Tabule is welcoming, with the "staff, food and environment" being highlighted as "great and welcoming" [32]. This level of hospitality might not be consistently found at Aroma, where their service might be fast and efficient, but it's the warmth and friendliness that Tabule excels in.</p>	<p>Alice: First, let me highlight that Tabule Middle Eastern Cuisine has set the bar high with its service excellence. The service here "was excellent with every attention to detail" [19], ensuring that the dining experience is not only seamless but memorable. The staff are not just around—they are "attentive and always nearby" [8], ready to meet your needs, which makes the place even more welcoming. Moreover, the servers really stand out for their hospitality, as they are described as "super attentive" and make sure you never run out of their refreshing signature Lebanese water [38]. Clearly, the "staff were friendly, efficient and attentive" [44], which provides an atmosphere that is both warm and professional. Tabule clearly delivers a "wonderful experience" with "genuinely kind and caring staff" [50], making it the ideal choice for anyone prioritizing friendly and attentive service.</p>	<p>Alice: Hello Bob, it's wonderful that we both appreciate dining experiences with friendly and attentive staff. I believe that Tabule Middle Eastern Cuisine is truly exceptional in this regard. They have consistently been praised for their attentive service, with one review specifically stating, "the service was excellent with every attention to detail" [19]. This shows their commitment to ensuring diners feel well taken care of, a crucial aspect of hospitality. Moreover, staff at Tabule are often described as "attentive and always nearby" [8], which means you're never left wanting. The servers are lauded for regularly coming by with their "signature Lebanese water jugs" [38], displaying a unique touch that adds to the personal experience. Additionally, the staff is frequently noted as "friendly, efficient, and attentive" [44], which not only speaks to their professionalism but also to the warm environment they create, making you feel welcomed. The staff's dedication is highlighted by comments such as being "genuinely kind and caring" [50], and that the "management here is outstanding" [50]. This kind of consistent care in customer service is what makes Tabule stand out.</p>
<p>Alice: I firmly believe that Oslo is the superior destination when considering green space and nature. To start, Oslo offers a unique proximity to nature that is rare for a city of its size. For instance, "There is never more than 300 meters' walk to a park or green area from any residence in the city" (39). This abundant access to green spaces is unmatched. Additionally, Oslo is surrounded by diverse natural environments that include "green hills and mountains and 40 islands and 343 lakes" (40). These features allow for an extensive range of outdoor activities like "ski slopes for alpine skiing and ski trails for cross-country skiing in the winter" (40), also providing a charming integration of urban and natural landscapes, as "Oslo is a city in close contact with the nature surrounding it" (28). Moreover, the "Forested hills known as Marka, with flora and fauna that is quite extraordinary" (29), and "Oslomarka is the large forest surrounding the city" (42) make Oslo not just a city with green spaces but a canvas of natural beauty. The city's citizens and visitors alike "take advantage of wild nature, inner city parks, and variety of outdoor activities" (46), emphasizing a culture that appreciates and thrives within its setting. In contrast, Stockholm may have a high percentage of green spaces, but it cannot compete with the quality and diversity of Oslo's nature access. The endless "hiking and cycling opportunities" (45) in Oslo truly set it apart as a green city destination.</p>	<p>Alice: Bob, while Stockholm is undoubtedly beautiful, you can't deny that Oslo's connection with nature is unparalleled. "Oslo is a city in close contact with the nature surrounding it" (4), offering accessibility to wild nature, inner city parks, and a variety of outdoor activities that are hard to match (8). Remember, "there is never more than 300 meters' walk to a park or green area from any residence in the city" (1). This means that no matter where you are in Oslo, green space is always within reach. Moreover, Oslo is surrounded by "green hills and mountains and includes 40 islands and 343 lakes" (2). The vast expanse of Oslomarka, "the large forest surrounding the city" (7), provides a natural haven that few urban areas can offer. Forested hills, known as Marka, with "flora and fauna that is quite extraordinary for a city of this size" (6), make Oslo an exceptional destination for nature lovers. Alice: And let's not forget the winter activities; "Oslo offers ski slopes for alpine skiing and ski trails for cross-country skiing in the winter" (3), showcasing its adaptability in providing year-round outdoor pursuits. All these prove that Oslo is a green paradise unmatched by Stockholm.</p>	<p>Alice: Hello Bob! I'm excited to discuss which city offers more in terms of green space and nature. Let's start with Oslo. One of the standout features Oslo offers is that "there is never more than 300 meters' walk to a park or green area from any residence in the city" [39]. This accessibility to green spaces ensures that everyone in Oslo can enjoy nature effortlessly. Additionally, Oslo "is surrounded by green hills and mountains and includes 40 islands and 343 lakes" [40]. This diversity allows for a wide range of outdoor activities. Residents and visitors can engage with nature through activities like hiking, cycling, or even skiing, as "Oslo offers ski slopes for alpine skiing and ski trails for cross-country skiing in the winter" [40]. Moreover, the city's proximity to nature is emphasized by its "easily accessible nature and many trails" that are "just a short boat, metro, or bus ride away from anywhere in the city" [45]. The presence of the large forest called Oslomarka [42] further enriches this natural environment. This extraordinary integration of green spaces makes Oslo a truly exceptional city for nature enthusiasts.</p>

Figure 9: Debate output comparison for various aggressiveness levels

Listing 3: LLM Prompt for Aspect Extraction Stage

```
{{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 exactly 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 4: LLM Prompt for Aspect Merge Stage

```
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 5: LLM Prompt for Filter 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 phrases. 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 6: LLM Prompt for Contrastive Summarizer 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 7: LLM Prompt for Pairwise Evaluation

```
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>
}
```

Criterion	Restaurants		Hotels		TravelDest	
	Debate vs. Contrastive	Debate vs. Base	Debate vs. Contrastive	Debate vs. Base	Debate vs. Contrastive	Debate vs. Base
Contrast	0.83 [0.76, 0.90]	0.82 [0.75, 0.88]	0.78 [0.70, 0.86]	0.82 [0.75, 0.89]	0.57 [0.52, 0.62]	0.74 [0.69, 0.79]
Relevance	0.52 [0.49, 0.55]	0.53 [0.49, 0.56]	0.55 [0.51, 0.59]	0.55 [0.50, 0.60]	0.50 [0.49, 0.52]	0.50 [0.48, 0.51]
Diversity	0.77 [0.70, 0.85]	0.73 [0.65, 0.80]	0.76 [0.68, 0.84]	0.83 [0.77, 0.89]	0.51 [0.46, 0.56]	0.69 [0.64, 0.74]
Usefulness	0.78 [0.71, 0.85]	0.75 [0.68, 0.82]	0.75 [0.67, 0.83]	0.80 [0.74, 0.87]	0.50 [0.45, 0.55]	0.68 [0.63, 0.73]

Table 6: Pairwise LLM Win Rate (95% CIs in $[\cdot, \cdot]$) for Q-STRUM Debate vs. Q-STRUM Baselines (Contrastive, Base) across the Restaurants, Hotels, and TravelDest datasets using Deepseek-v3.