Dialogue2Data (D2D): Transforming Interviews into Structured Data Final Project Report

Sienko Ikhabi, Dominic Lam, Wangkai Zhu, Yun Zhou

Table of contents

1 Executive Summary 2

2 Project Introduction 3
2.1 Problem Statement . . . 3
2.2 Data Formats . . . 3
2.3 Project Objectives . . . 3
2.4 Solution Overview . . . 4

3 Data Science Techniques 5
3.1 Description of Techniques Used . . . 5
3.2 Challenges and Shortcomings . . . 5
3.3 Evaluation . . . 5

4 Data Product and Results 6
4.1 Description of Data Product . . . 6
4.2 Key Results . . . 6
4.3 Future Improvements . . . 10

5 Conclusion and Recommendations 12

6 References 13

1 Executive Summary

Open-ended text data holds immense potential for discovering user insights, but extracting those insights efficiently remains a major challenge—particularly in unstructured formats. Our capstone partner, Fathom, specializes in surfacing insights from raw survey data and interview transcripts to support client decision-making. While structured survey responses follow a question-and-answer format, interview transcripts are conversational and free-flowing, making them more difficult and time-consuming to analyze reliably.

To address this challenge, we designed and implemented a data product that automates the extraction of relevant responses from unstructured interview transcripts. Our solution leverages a Retrieval-Augmented Generation (RAG) framework, which consists of:

- A **retriever** that selects the most relevant passages from the transcript based on a guideline question,
- A **generator** that formulates concise answers using the retrieved text and prompt,
- And an **evaluator** that scores each generated response on five key quality metrics and flags uncertain cases for human review.

This system significantly reduces the need for manual transcript annotation while improving accuracy and consistency. In testing, it increased correct response identification by 87% compared to the previous baseline, and streamlined the review process by automatically surfacing low-confidence matches.

The final pipeline is fully functional and ready for integration into Fathom's existing analytics workflow. By enabling structured analysis of unstructured interviews, our tool enhances Fathom's ability to deliver richer insights to clients and scale their operations to handle more complex, conversational data. Ultimately, this project extends the capabilities of conversational survey analytics, allowing organizations to extract meaningful information with greater speed, precision, and depth.

2 Project Introduction

2.1 Problem Statement

Organizations often rely on surveys and interviews to understand their users, assess needs, and guide decision-making. While traditional surveys provide structured data that is relatively easy to analyze, open-ended formats—especially in interviews—offer deeper, more nuanced insights. However, the richness of these responses comes at a cost: they are difficult to process at scale due to their unstructured and free-flowing nature.

Our capstone partner, Fathom, specializes in analyzing open-text data from surveys and interviews. Their platform already supports structured survey responses, but scaling up analysis of interview transcripts has proven to be a key challenge. Interview responses are less uniform and harder to map directly to the original guideline questions. As a result, Fathom's team must rely heavily on large language models and manual transcript review to extract relevant content—an approach that is time-consuming, error-prone, and difficult to scale across large volumes of data.

2.2 Data Formats

The dataset used for this project consists of over 200 unstructured, conversational interview transcripts, alongside more than 20 sets of guideline questions to which the responses must be mapped. These transcripts vary widely in length, tone, and structure, posing a realistic challenge for scalable NLP systems. The output of the pipeline is a structured `.csv` file in which each row corresponds to an interviewee and each column contains the extracted response to a specific guideline question, enabling seamless downstream analysis and integration into existing workflow.

2.3 Project Objectives

To address this challenge, we refined the problem into the following tangible data science objectives:

1. **Automate the mapping of interview transcript content to original guiding questions**, reducing human effort and improving consistency. Currently, Fathom relies heavily on large language models (LLMs) for response extraction, followed by manual review to ensure quality. This approach is time-intensive and difficult to scale. Automating the mapping process allows for faster, more consistent analysis and reduces dependence on manual labor.

2. **Ensure high response quality** by integrating an evaluation module that scores generated outputs and flags uncertain responses for manual review. Manual validation of LLM outputs is costly and error-prone. A built-in evaluator provides a systematic way to assess the quality of generated responses across key metrics, surfacing only ambiguous or low-confidence cases for human review—saving time while preserving accuracy.

3. **Deliver an end-to-end pipeline** that can integrate into Fathom's existing workflow, enabling scalable and repeatable analysis of new interview data. For this solution to be truly impactful, it must work within Fathom's current analytics infrastructure. An integrated, end-to-end pipeline ensures seamless adoption and allows the team to expand from survey data to conversational interview data with minimal friction.

2.4 Solution Overview

Our solution is built using a **Retrieval-Augmented Generation (RAG)** architecture, leveraging NLP techniques including **embedding-based retrieval** and **large language models** for answer generation. The pipeline also includes an evaluator that assesses output across five quality dimensions: **correctness**, **faithfulness**, **precision**, **recall**, and **relevance**.

By framing the problem around semantic matching and response generation, we developed a scalable and effective data science pipeline that meets Fathom's analytical needs. This system empowers them to analyze open-ended interview transcripts more efficiently and deliver richer, faster, and more actionable insights to their clients.

3 Data Science Techniques

3.1 Description of Techniques Used

3.2 Challenges and Shortcomings

3.3 Evaluation

To assess the quality of D2D's output, we have designed the evaluator to ensure that the answers are not only accurate in meaning, but also generated through a reliable and precise process.

Inspired by RAGAS (Retrieval-Augmented Generation Assessment)[1] and partner's internal documentation Daedalus v4[2], the evaluation framework provides five core metrics:

- **Correctness**: Measures how well the answer is consistent with the reference (ground truth).
- **Faithfulness**: Evaluates whether the answer is fully supported by the retrieved context and avoids hallucinations.
- **Precision**: Assesses the proportion of the answer that is actually supported by the retrieved chunks.
- **Recall**: Captures how many relevant facts from the context are included in the answer.
- **Relevance**: Reflects how closely the answer relates to the original guideline question. This metric is used only to assess questionnaire quality, not processor performance.

These metrics are computed using LLM-based prompting, with carefully designed templates and decision logic for edge cases such as ambiguous or evasive responses. Compared to the standard RAGAS pipeline, our customized evaluator introduces three key enhancements:

First, each metric uses a tailored LLM prompt designed to calculate metric score and handle edge cases such as vague or non-informative answers. Nonspecific or uninformative answers are detected through keyword matching and scored conservatively to ensure evaluation accuracy.

Second, built-in feedback mechanism: Each score includes LLM-generated feedback, improving interpretability and helping users understand and validate scoring results.

Third, flexible scoring and low-score highlighting: Users can customize metric weights and set thresholds to flag low-scoring responses, helping streamline validation and adapt to varied evaluation needs.

Finally, to validate metric correctness, we have built a small but diverse golden sample set across ten topics, such as climate change, food, NBA, and workplace culture. By varying the number of interviews across topics, the sample better reflects real-world diversity.

4 Data Product and Results

4.1 Description of Data Product

Dialogue2Data (D2D) is a Python package that extracts answers from the unstructured interview transcripts based on the guideline questions for further statistical analysis. It includes two modules: the Processor, which applies embeddings and LLMs to match and generate responses to predefined guideline questions; and the Evaluator, which scores the outputs across five metrics such as correctness and faithfulness. The system supports csv/json outputs, logs, and flexible configuration via .env files.

D2D supports synchronous processing for high efficiency, incorporates detailed error handling for user-friendly operation, and integrates answer reference and evaluation feedback to aid manual verification. We quantified repeatability through multiple runs to ensure the system produces consistent and reliable results.

4.2 Key Results

To understand the performance of D2D processor, we conducted experiments using golden samples. We compared developed models against the client's baseline (vanilla LLM method), analyzed the average scores and distributions of different models across all metrics, and tuned retrieval parameters (top_k and top_p) to recommend optimal settings based on golden samples.

According to Figure 1, all of our developed models achieved much higher correctness scores compared to the client's baseline method. The baseline model scored only 2.20, while all developed models scored above 3.90, and the best-performing model top-k with gpt4-1, reached 4.22. This clear improvement demonstrates the strength of our RAG approach.

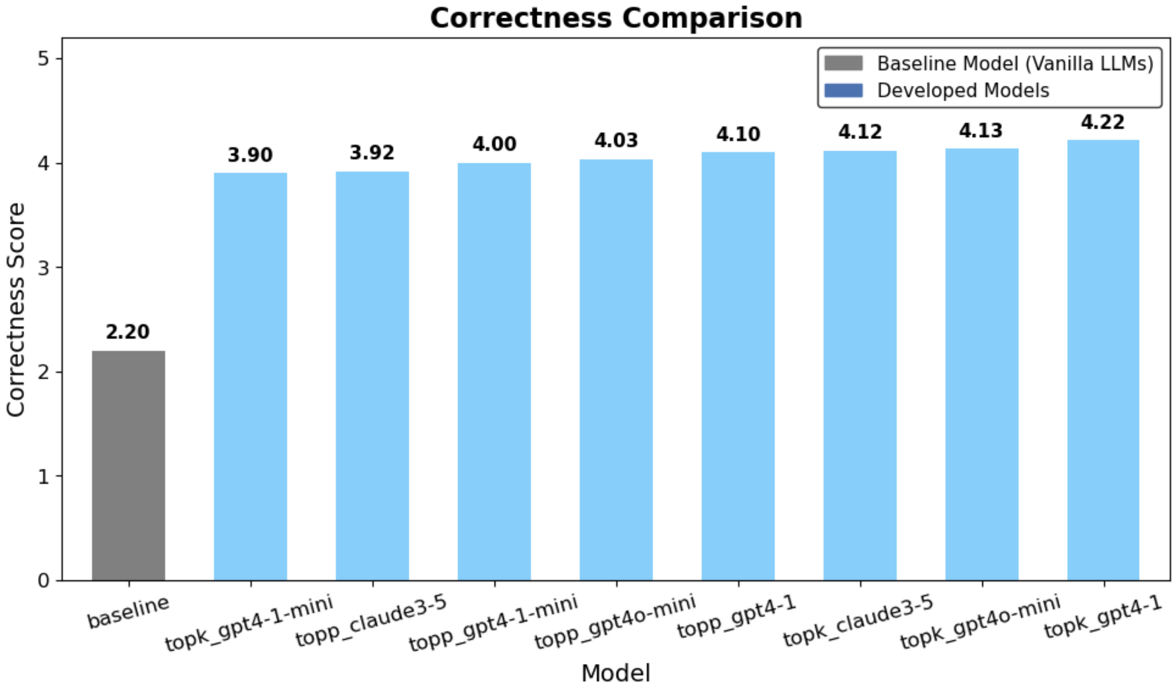

Figure 1: Correctness Comparison to Baseline

Then, we compared multiple model variants across all six evaluation metrics under the setting of k = 5 and p = 0.5, as shown in Figure 2. Overall, top-k configurations performed slightly better than their top-p counterparts. Among models using the same retriever, GPT-4.1 and GPT-4.1-mini showed slightly better performance than Claude 3.5 and GPT-4o-mini. We also observed that the variance across 10 runs was minimal, indicating stable and repeatable results. Moreover, the differences between models were relatively small across all metrics. Taking both performance and cost into account, we recommend top-k = 5 with GPT-4o-mini as the default setting.

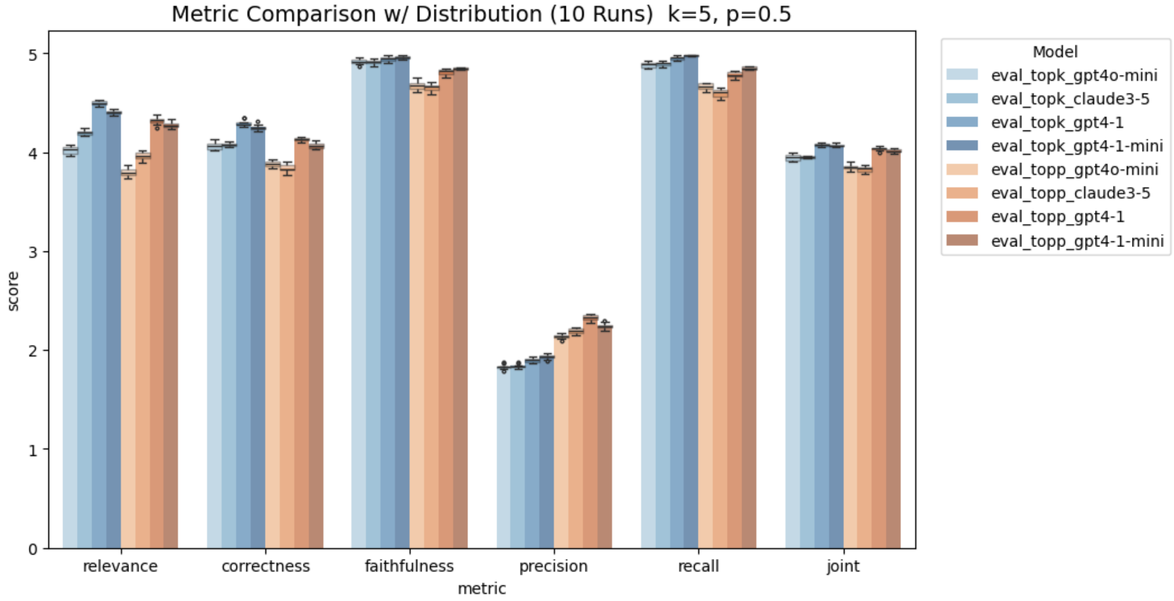

Figure 2: Metric Comparison with Distribution

Finally, we tuned the retriever parameters using our golden sample set. As shown in Figure 3, for top-k retrieval, k = 5 yielded the best correctness and joint scores, making it the recommended setting. Similarly, Figure 4 shows the performance of models using top-p retrievers across different p values. The optimal performance was observed in the 0.52–0.56 range, suggesting that future p tuning efforts should focus within this interval.

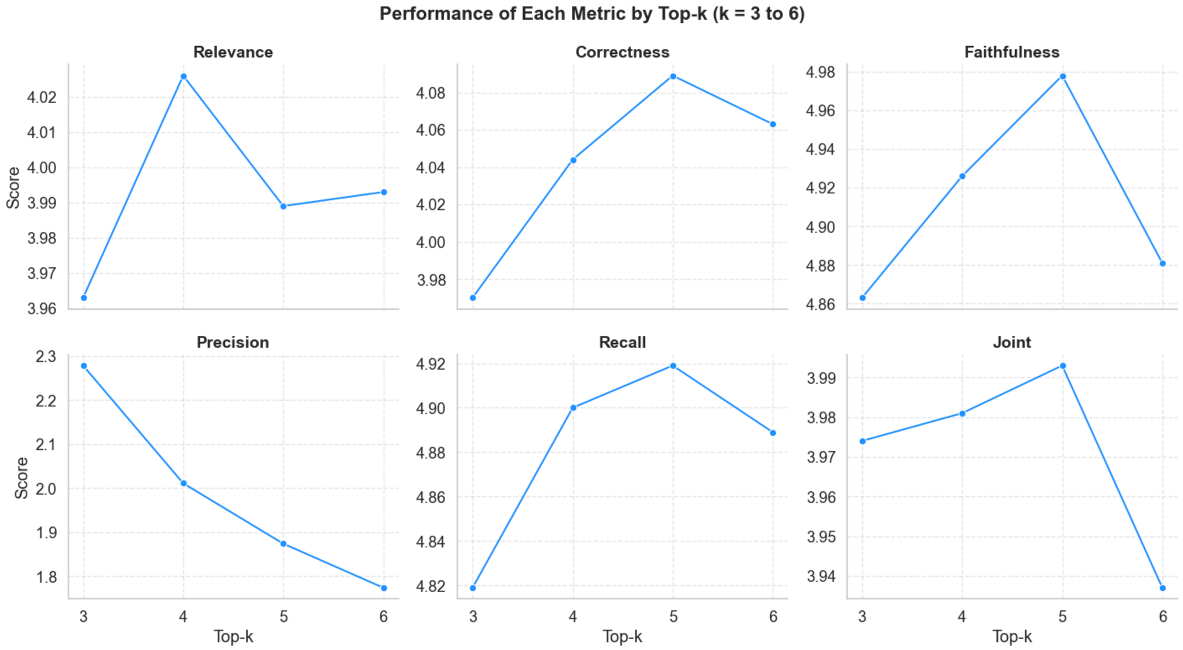

Figure 3: Model Performance Across Top-k Values (k = 3 - 6)

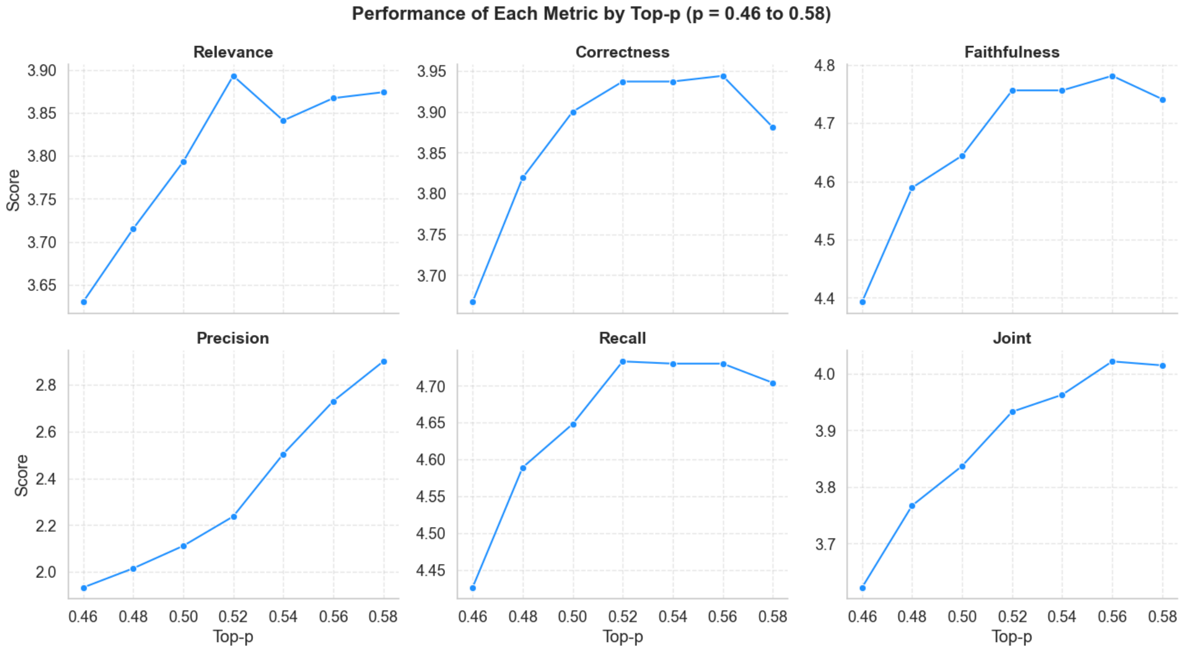

Figure 4: Model Performance Across Top-p Values (p = 0.46 - 0.58)

4.3 Future Improvements

The current release delivers practical business use. But as any project, new versions will be released with improved features and more flexibility.

Below are the improvements that can be developed in the next few releases by the partner: -

- **Integration Testing**: so far our evaluation has been restricted to using the evaluator and our personal judgement. The first and most obvious next step would be for our capstone partner to test the data product in their standard pipeline with the sample data that they had provided
- If you visualize the heatmap of the how QA Pairs match up to interview guideline questions based on cosine similarity of the embeddings, you will see that it is perfectly linear (along the diagonal from top-left to bottom right) in the synthetic dataset in Figure 5. Generating realistic dialogue was not easy, and the "perfectly linear" nature of the similarity matching shows this. Nonetheless, the heatmap generated from one of the partner's actual datasets show that our pipeline is robust enough and still finds the required matches when the interview length and format varies (Figure 6). Despite this, with the real-world data, we still see the diagonal pattern as much as it is less pronounced. Another area of testing for robustness would be to process some **non-sequential interviews** whose flow does not follow the same general flow as guideline questions
- Still referring to Figure 6, we notice that in most interviews, there are some dialogue sections that have very low similarity matches to any of the guideline questions. A more precise snapshot is highlighted in Figure 7 for one interview, with such sections marked with dotted lines. To improve the current process, a **post-processing step** for the matches can be added. The first objective of this step would be to fill in the gaps to any previously un-answered guideline question. For example, in Figure 7 Guideline Question 2 (second column) had "[No relevant response found]" output due to the low similarity matches. However, in the transcript, this question was answered by the interview dialogue pairs 2 and 3 (with the respective row numbers), given that Guideline Question 1 is answered in the first exchange (row 1) and Guideline Question 3 is answered in row 4. After this post-processing, any remaining un-used dialogue portions might be "out of topic" exchanges (for example "exit pleasantries" like *"Thank you for making time to speak with me."* at the end [rows 16 & 17]) or they could be additional information that was not covered by the guideline questions but could be useful to extract. Our partner had set this as a stretch goal for the capstone project. Despite having limited time we had to implement it, our work has set a good stage for this objective to be achieved

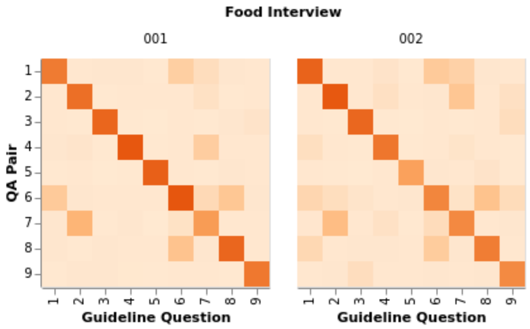

Figure 5: Heatmap with QA Pairs similarity to Guidelines (Synthetic Data)

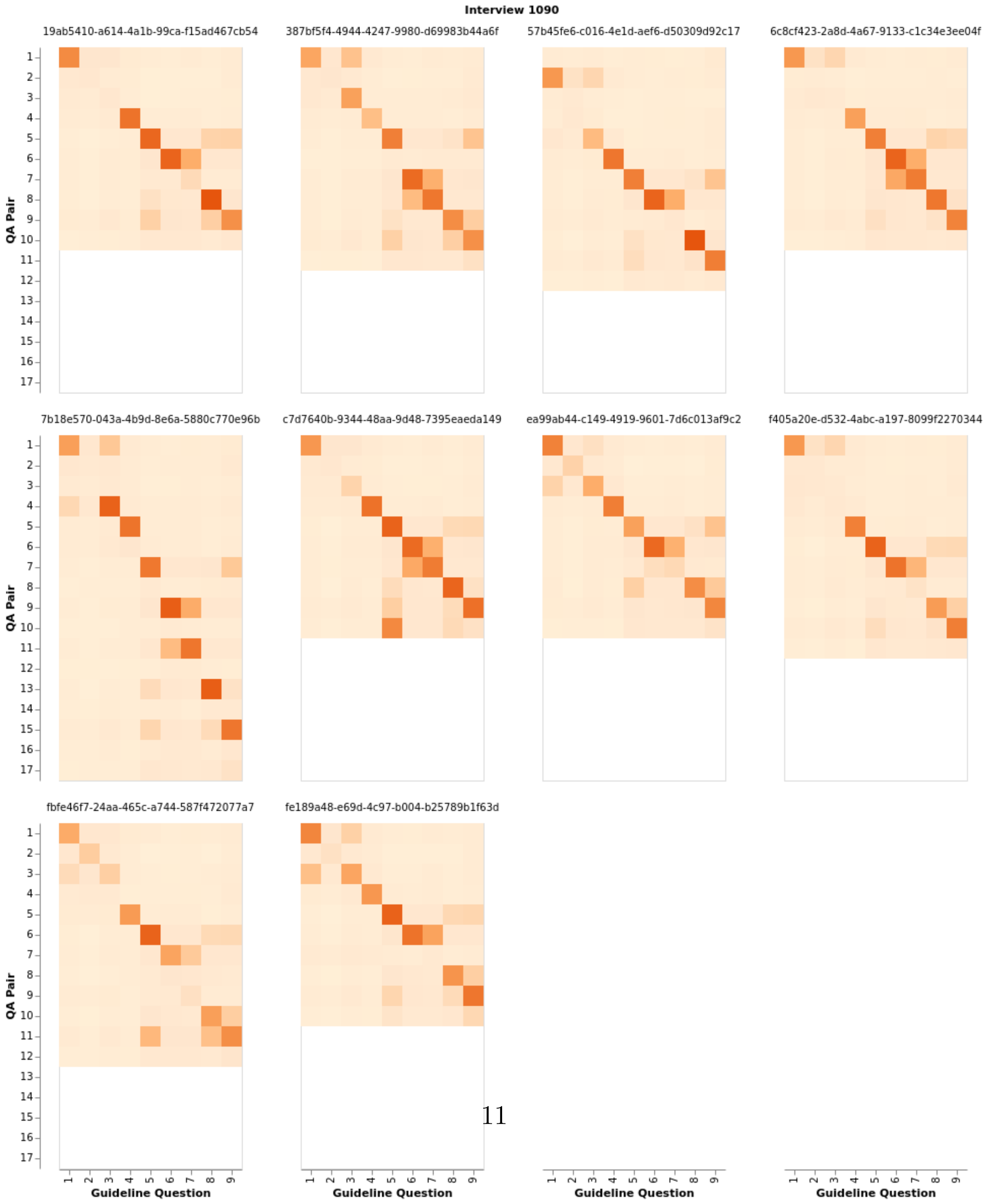

Figure 6: Heatmap with QA Pairs similarity to Guidelines (Real Data)

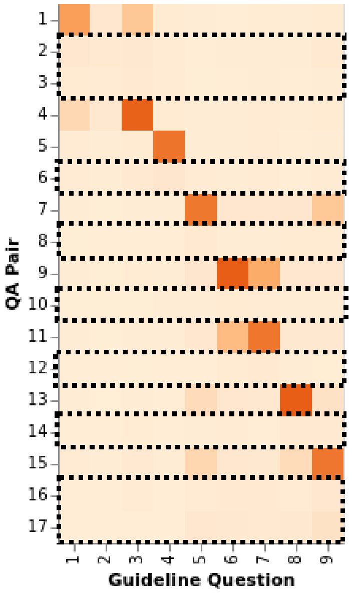

Figure 7: Heatmap showing Unused QA Pairs

5 Conclusion and Recommendations

In our capstone project, we designed and developed a functional pipeline that extracts concise summaries from interview transcripts and outputs structured data suitable for integration into our partner's downstream analysis workflows. Alongside this `processor` pipeline, we built an `evaluator` framework that enables our partner to assess the quality of extracted outputs. Both components are flexible and include hyperparameters that can be adjusted and customized.

The pipeline has been tested across a variety of scenarios and interview topics drawn from the synthetic data set in the public repository and the partner's private dataset (not included in the public repository), demonstrating its robustness and adaptability. We believe the methods and results presented here offer a strong foundation for future development by our partner, while already showing practical value for real-world use.

6 References

- [1] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert, “RAGAS: Automated Evaluation of Retrieval Augmented Generation,” arXiv preprint arXiv:2309.15217, Sep. 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.15217

- [2] DS-Daedalus Team, “DS-Daedalus v4: Increments and Updates,” Internal Presentation Report, Mar. 2025