A generalized two step finetuning Insight Generator for survey data

Jiahao Zhang*[1,2]

[1]Grapedata
[2]UC Berkeley

March 30, 2024

1 Enhanced Fine-tuning Workflow

The workflow 1 integrates specific analysis and prompt design steps for both open and closed choice questions, aiming to enhance the Language Model (LLM)'s capability to generate insights directly from survey questions.

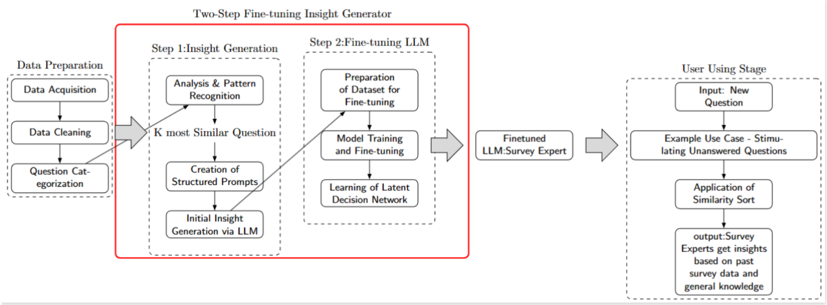

Two-Step Fine-tuning Insight Generator
Data Preparation
Data Acquisition
Data Cleaning
Question Cat-egorization
Step 1:Insight Generation
Analysis & Pattern Recognition
K most Similar Question
Creation of Structured Prompts
Initial Insight Generation via LLM
Step 2:Fine-tuning LLM
Preparation of Dataset for Fine-tuning
Model Training and Fine-tuning
Learning of Latent Decision Network
Finetuned LLM:Survey Expert
User Using Stage
Input: New Question
Example Use Case - Stimu-lating Unanswered Questions
Application of Similarity Sort
output:Survey Experts get insights based on past survey data and general knowledge

Figure 1: Insight Generator Overview flow

2 Data Preparation Stage

After cleaning the original data from Grapedata's Knowledge Base,the dataset can be recogonized as question-responses pairs,we distinguish the questions into 2 parts:open-ended questions like "What is your favorite software" the answer space is infinite and hard to deal with,Chioce-Based Questions which has finite

---

*email:jiahao.zhang.public@gmail.com or edu:jiahao.zhang_public@berkeley.edu

questions for survey respondents to choose.We designed different prompt template for these questions using the prompt technology - Chain of Thoughts(CoT) and few shot learning(or few shot examples)

Templates for Analysis

Choice-Based Questions:

- *Available Choices*: List all possible answer choices.
- *Response Distribution*: Show the distribution of responses across the choices.
- *Descriptive Statistics*: Include total responses, most/least common choices, and other relevant statistics.
- *Insights*: Generate insights highlighting patterns, preferences, and outliers.

Open-Ended Questions:

- *Response Summary*: Summarize the key themes and sentiments.
- *Topic Modeling*: Identify and describe the main topics from the responses.
- *Sentiment Analysis*: Analyze the overall sentiment of the responses.
- *Insights*: Highlight notable patterns, trends, and observations for further exploration.

Incorporating the Similarity Sort Method

A key component of our methodology is the *Similarity Sort Method*, employed during both the Insight Generation and User Interaction stages. This method involves the following steps:

1. *Textual Similarity Calculation*: Utilizing techniques such as TF-IDF and cosine similarity to calculate the textual similarities between survey questions.
2. *Similarity Sorting*: Ranking the questions based on their similarities to a target question.
3. *Top-M Similar Question Selection*: Selecting the top M similar questions to the target question for further analysis or for constructing the training dataset.

This method enables the model to focus on the most relevant data during the fine-tuning process and allows for more accurate and contextually relevant insights during user interaction.

Similarity Sort Method

During Insight Generation: The Similarity Sort Method is used to identify and select the most relevant questions similar to the target question. This ensures that the insights generated are based on a focused set of data, enhancing the quality and relevance of the generated insights.
During User Interaction: When users query the model with new survey questions, the Similarity Sort Method is applied to find similar questions within the dataset. This allows the model to leverage previous insights and analyses, providing users with informed and contextually relevant responses.

3 Two-Step Fine-tuning Stage

Step 1: Insight Generation

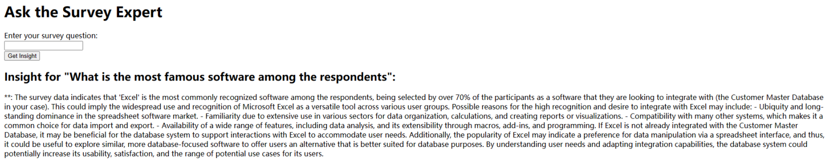
is critical for preparing the model to understand the complex patterns within the survey data. This step involves:

1. *Statistical Analysis and Pattern Recognition*: Analyzing survey responses to identify underlying patterns, preferences, and statistical correlations among the responses.
2. *Structured Prompt Creation*: Designing prompts that encapsulate the essence of the survey question along with derived statistical insights, readying them for the LLM's consumption.
3. *Initial Insight Generation with LLM*: Leveraging the LLM's capability to synthesize the provided information into coherent, insightful analyses reflective of respondent behaviors and preferences.

Step 2: Fine-tuning

builds on the insights generated in Step 1 to further refine the model's ability to autonomously generate insights from new survey questions without explicit statistical context. This involves:

1. *Dataset Preparation*: Using the insights generated in Step 1 alongside the original survey questions to compile a targeted dataset for model training.
2. *Model Training*: Fine-tuning the LLM with the prepared dataset, where the survey questions serve as inputs and the corresponding insights as the target outputs.
3. *Latent Decision Network Expression*: Through this training, the LLM develops an internal representation, or ”latent decision network,” that encapsulates the decision-making patterns, preferences, and the simulated ”choice selection space” of the survey respondents.



Ask the Survey Expert

Insight for "What is the most famous software among the respondents":

**: The survey data indicates that 'Excel' is the most commonly recognized software among the respondents, being selected by over 70% of the participants as a software that they are looking to integrate with (the Customer Master Database in your case). This could imply the widespread use and recognition of Microsoft Excel as a versatile tool across various user groups. Possible reasons for the high recognition and desire to integrate with Excel may include: - Ubiquity and long-standing dominance in the spreadsheet software market. - Familiarity due to extensive use in various sectors for data organization, calculations, and creating reports or visualizations. - Compatibility with many other systems, which makes it a common choice for data import and export. - Availability of a wide range of features, including data analysis, and its extensibility through macros, add-ins, and programming. If Excel is not already integrated with the Customer Master Database, it may be beneficial for the database system to support interactions with Excel to accommodate user needs. Additionally, the popularity of Excel may indicate a preference for data manipulation via a spreadsheet interface, and thus, it could be useful to explore similar, more database-focused software to offer users an alternative that is better suited for database purposes. By understanding user needs and adapting integration capabilities, the database system could potentially increase its usability, satisfaction, and the range of potential use cases for its users.

Figure 2: Enter Caption

This sequential approach ensures that the LLM not only learns to replicate the surface-level task of generating insights but also understands and mimics the deeper cognitive and decision-making processes of the survey participants.

4 User Interaction Stage

1. **Input: New Question:** Users can query the model with new survey questions.
2. **Application of Similarity Sort:** Identifying similar questions in the dataset to leverage past insights.(Latent for users)
3. **Output: Survey Insights:** The model generates insights based on past survey data and general knowledge, guiding users in their exploration of unanswered questions.

usage example

Here is a straightforward example illustrated in Figure 2, where we input the query "What is the most famous software among the respondents" to our fine-tuned survey expert. Even in the absence of direct survey data pertaining to this specific question, our expert adeptly narrows its search scope by identifying and sorting the top K most similar questions from our dataset. It then proceeds to comprehend the latent decision network characteristic of the respondent group. Essentially, the survey expert simulates the decision-making framework of this population, deriving insights not only from its comprehensive knowledge base but also leveraging the extensive repository of Grapedata. This process exemplifies how the expert harnesses its general understanding and specific data insights to formulate informed responses.