

# Lab 2: Improving Agent Quality through AI Instructions and Example Queries

## Pre-requisites

1. You have completed Lab 1 – Getting Started

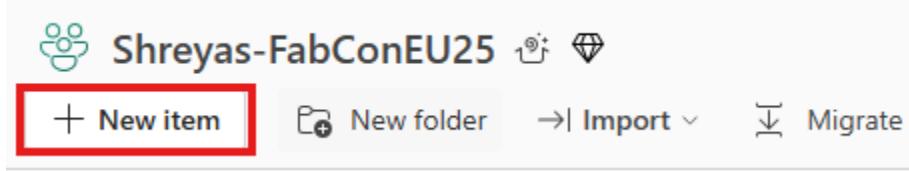
## Learning Objectives

1. Connect a data source to a data agent.
2. Learn how to use Few Shot Queries, AI Instructions, and Data Source Instructions to improve response quality.

## Lab Steps

### Step 1: Create a Data Agent

1. Select “+ New Item”



2. Search for “Data Agent” and Select “Data agent”

New item

Favorites All items

Data Agent

Analyze and train data

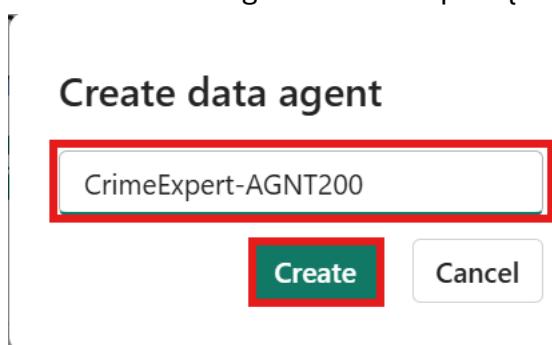
Propose hypotheses, train models, and explore your data to make decisions and predictions.

Data agent (preview)

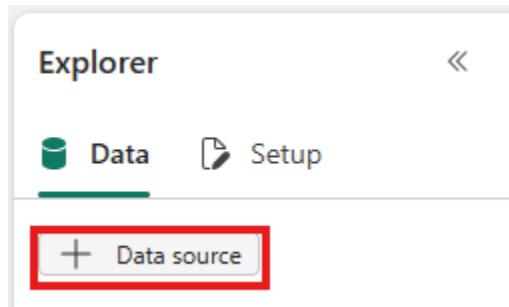
Build a generative AI agent that understands your data and can answer complex questions in a variety of conversational interfaces.

A screenshot of a search results page. At the top, there's a search bar with the text "Data Agent" (highlighted with a red box). Below the search bar, there are two buttons: "Favorites" and "All items". The main area shows a card titled "Data agent (preview)" which contains the text: "Build a generative AI agent that understands your data and can answer complex questions in a variety of conversational interfaces." There's also a small green icon next to the title.

3. Name the Data Agent “CrimeExpert-{YOUR\_WORKSPACE\_NAME}” -> “Create”



4. Click on “+ Data source” in the Explorer



5. Search for your “Kaggle” Data source using its name, select it, and hit “Add”

A screenshot of the "Add a data source" search results page. The search bar at the top right contains the text "Kaggle" with a red border. Below the search bar is a table with columns: Name, Type, Owner, Refreshed, Endorsement, and Sensitivity. One row in the table is highlighted with a red border, corresponding to the "Kaggle" entry in the search results.

Name	Type	Owner	Refreshed	Endorsement	Sensitivity
Kaggle	Lakehouse	Markus Cozowicz	—	—	—

6. You should see the lakehouse with all the tables we’ve loaded added to the data agent:

Explorer <<

No tables selected yet. The data agent can't answer questions until at least one table for a data source is selected.

+ Data source

- ▼ Kaggle
  - ▼ GreaterManchesterCrime
    - >  greatermanchestercrime
  - ▼ Pesticide
    - >  resultsdata15
    - >  sampledata15
  - > StudentMathScore
  - > TheHistoryofBaseball
  - > WorldSoccerDataBase

## Step 2: Data exploration through Data Agent

For this lab, we will focus on the “GreaterManchesterCrime” Dataset.

1. Start by selecting all the tables under the “GreaterManchesterCrime” dataset:

Explorer <<

+ Data source

- ▼ Kaggle
  - ▼ GreaterManchesterCrime
    - >  greatermanchestercrime

2. Let's start by understanding the response structure for the Data Agent.

Ask the Data Agent to "give me the top 10 rows in the data"

give me the top 10 rows in the data



Created with AI. Mistakes are possible. Review terms

### 3. Let's Analyze the Response:

Here are the top 10 rows from the Greater Manchester crime data:

1. CrimeID: 8F:02:C8:40:54:44 | Date: 8/12/2018 | Location: Buxton Lane | Type: Violence and sexual offences | Outcome: Unable to prosecute suspect
2. CrimeID: 90:2E:9C:3D:06:64 | Date: 8/1/2018 | Location: Ravendale Close | Type: Violence and sexual offences | Outcome: Under investigation
3. CrimeID: CC:86:69:39:F1:5F | Date: 8/25/2018 | Location: Shirley Road | Type: Vehicle crime | Outcome: No suspect identified
4. CrimeID: 77:0E:CF:ED:BF:34 | Date: 8/8/2018 | Location: Parking Area | Type: Shoplifting | Outcome: Under investigation
5. CrimeID: EC:30:31:16:BF:2C | Date: 8/5/2018 | Location: Tithe Barn Street | Type: Criminal damage and arson | Outcome: No suspect identified
6. CrimeID: 45:3B:8E:41:42:52 | Date: 8/4/2018 | Location: Shannon Road | Type: Criminal damage and arson | Outcome: No suspect identified
7. CrimeID: 1F:49:BE:8D:DE:F4 | Date: 8/14/2018 | Location: Barncroft Gardens | Type: Vehicle crime | Outcome: No suspect identified
8. CrimeID: A4:69:9B:85:8D:F7 | Date: 8/22/2018 | Location: Thomas Street | Type: Burglary | Outcome: Under investigation
9. CrimeID: BC:B3:9F:10:6F:51 | Date: 8/12/2018 | Location: St Brendan'S Road North | Type: Criminal damage and arson | Outcome: No suspect identified
10. CrimeID: 41:D0:4D:3B:E9:8D | Date: 8/24/2018 | Location: Moorcroft Street | Type: Vehicle crime | Outcome: No suspect identified

Let me know if you need more details or analysis!

1 step completed ▾

⌚ Response time: 13 sec

Data agent will always respond in Natural Language. In this case, it takes the values from the table and outputs 10 rows from the dataset.

We can also take this further and view the steps Data Agent took to get this result by clicking on "1 step completed"

Let me know if you need more details or

1 step completed ▾

The expanding on the "Run Step":

⌚ Analyzed ShreyasFabconEU25Data Lakehouse for: "Show the top 10 rows from the greatermanchestercrime table." ▾

</> Query code

Sql

```
SELECT TOP 10 *
FROM [GreaterManchesterCrime].[greatermanchestercrime];
```

Query output

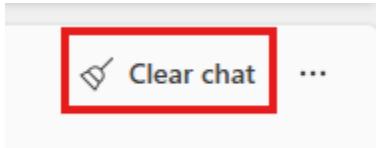
CrimID	CrimeTS	Location	LSOA	Type	Outcome	theGeom	lon
8F:02:C8:40:54:44	8/12/2018 6:51:23 AM	On or near Buxton Lane	Stockport 029C	Violence and sexual offences	Unable to prosecute suspect	POINT(-2.072135 53.390292)	-2.072135 53.390292

This should give you an SQL query that was used to generate the response. These can be used by the user to understand what steps Data Agent took to generate a response.

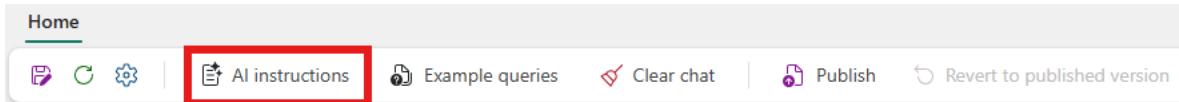
- Now you try! Ask some additional data exploration questions. Take this time to understand the results of the query and how data agent came up with the result. You may ask the following questions to start:
  - “What are all the unique values in the “Type” column”
  - “What are all the values in the Location column?”
  - “What was the most common location for a crime?”

### Step 3: Improving Results Using Agent Instructions

- Let’s start by clicking “Clear Chat” on the top right corner



- Select the “AI Instructions” In the Ribbon



- Add the following instruction to the Agent:

“When displaying results, I want you to always use table output for tabular results in

Markdown format.”

The screenshot shows the Data Agent interface. At the top, there's a section titled "Agent instructions" with a green icon. Below it is a text area containing a system prompt: "Guide the data agent to the best data sources for different types of questions by writing a detailed system prompt." A message input field contains the text: "When displaying results, I want you to always use table output for tabular results in Markdown format." At the bottom of the input field, it says "102/15000 characters used".

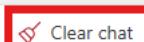
4. Try asking Data Agent some exploratory questions again:

- “give me the top 10 rows in the data”
- “What are all the unique values in the “Type” column”
- “What are all the values in the Location column?”
- “What was the top 5 most common location for a crime?”

Do you notice a change in how the results are displaying in the chat canvas?

## Step 4: Improving Results Using Few Shots Example Queries

1. Let’s start by clicking “Clear Chat”



Data Agent uses the chat history as part of the context, so let’s begin from a fresh slate.

2. Ask the Question “Which Streets in Manchester have the highest crime rates?”  
*note: I expect Data Agent to ONLY give me streets from the Dataset where Crimes have occurred. You may receive a result below:*

Which Streets in Manchester have the highest crime rates?

The streets in Manchester with the highest crime rates, based on recorded crime counts, are:

1. On or near Parking Area – 150 crimes
2. On or near Shopping Area – 133 crimes
3. On or near Supermarket – 114 crimes
4. On or near Petrol Station – 88 crimes
5. On or near Nightclub – 63 crimes
6. On or near Sports/Recreation Area – 39 crimes
7. On or near Pedestrian Subway – 28 crimes
8. On or near Piccadilly – 23 crimes
9. On or near Prison – 20 crimes
10. On or near Park/Open Space – 18 crimes

None of these are street names.

3. Let's add a Few Shot example query to help the Agent answer this question. Select "Example queries" in the ribbon, and hit the edit button:



Example queries X

For each data source, add common questions and how to interpret them.

Kaggle

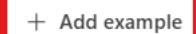
Example SQL queries Edit 

0 examples added

4. Click on "Add example"

### Example SQL queries

Add examples of how a natural language question would become a question. [Learn more](#)

 Add example

Import from json

Download

Showing 0 items (0 selected)

5. Input the below question and query pair:

**Question:** Which Streets in Manchester have the highest crime rates?

**Query:**

```
SELECT Location AS StreetName, COUNT(*) AS NumberOfCrimes
FROM [GreaterManchesterCrime].[greatermanchestercrime]
-- note that we ALWAYS want to perform case insensitive searches
-- when the question asks about streets, ALWAYS look for 'street' in the location column
WHERE LOWER(Location) LIKE '%street%'
```

```
GROUP BY Location  
ORDER BY NumberOfCrimes DESC
```

□ Question      SQL query

Which Streets in Manchester have the highest crime rates?

```
SELECT Location AS StreetName, COUNT(*) AS NumberOfCrimes  
FROM [GreaterManchesterCrime].[greatermanchestercrime]  
-- note that we ALWAYS want to perform case insensitive searches
```

57/500 characters used

Close the window.

#### 6. Clear the chat



#### 7. Ask the question once again:

“Which Streets in Manchester have the highest crime rates?”

Sample questions

Which Streets in Manchester have the highest crime rates?



Do you notice a change in the result?

As part of the run steps, you may also notice flavors of the example query in your result:

Analyzed Kaggle Lakehouse for: "List the streets in Manchester with the highest number of reported crimes.  
Provide the top streets by number of crimes reported."

</> Query code

Sql

```
SELECT Location AS StreetName, COUNT(*) AS NumberOfCrimes  
FROM [GreaterManchesterCrime].[greatermanchestercrime]  
-- note that we ALWAYS want to perform case insensitive searches  
-- when the question asks about streets, ALWAYS look for 'street' in the location column  
WHERE LOWER(Location) LIKE '%street%'  
GROUP BY Location  
ORDER BY NumberOfCrimes DESC
```

## Step 5: Improving Results Using Datasource Instructions

#### 1. Let's start by clicking “Clear Chat”



#### 2. Ask the question: “Where is the place with the largest number of violent crime events?”

It is likely that Data Agent will know the values in the table. Hence, data agent is unable to determine how to parse “violent crime” in the SQL query:

Analyzing ShreyasFabconEU25Data LakehouseTables database ...

Details ^

Analyzing ShreyasFabconEU25Data Lakehouse for: "Find the location with the largest number of violent crime events and provide the number of events."

Query code

Sql

```
SELECT TOP 1 Location, COUNT(*) AS NumberOfViolentCrimes
FROM [GreaterManchesterCrime].[greatermanchestercrime]
WHERE LOWER(Type) LIKE '%violent%'
GROUP BY Location
ORDER BY NumberOfViolentCrimes DESC
```

3. In order to help the Data Source with this question, let's add some Data Source Instructions:

Select the icon to go back to the AI Instructions Configurations -> Expand the Kaggle Instructions:

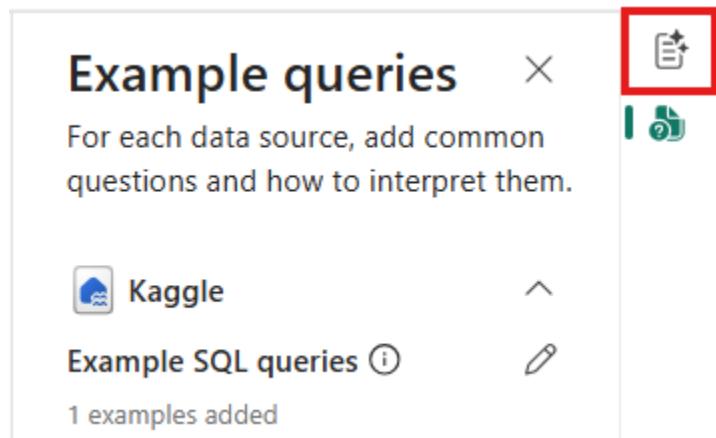
Example queries X

For each data source, add common questions and how to interpret them.

Kaggle

Example SQL queries ⓘ

1 examples added



## Data source instructions

For each connected data source, help the data agent understand its data and how to use it most effectively.



Kaggle instructions



4. Add the following Instructions:

```
## Table descriptions
```

Below are the possible values in the type column:

- Violence and sexual offences
- Public order
- Criminal damage and arson
- Vehicle crime
- Burglary
- Other theft
- Shoplifting
- Other crime
- Robbery
- Theft from the person
- Possession of weapons
- Bicycle theft
- Drugs



Kaggle instructions



```
## Table descriptions
```

Below are the possible values in the type column:

- Violence and sexual offences
- Public order
- Criminal damage and arson
- Vehicle crime



298/15000 characters used

5. Let's clear the chat, and try asking the question again:  
"Where is the place with the largest number of violent crime events?"  
Was Data Agent able to provide a response this time? 😊

## Step 6: Challenge – Use what you've learned to answer additional questions

1. From what you've learned above, can you configure the Data Agent to answer the following questions:
  - a. "Which Type of crime happens the most in Salford?"  
For this question, I want the Data agent to use the LSOA column instead of location!
  - b. "What are all the crimes that are undergoing or have undergone a trial?"  
Hint: Take a look at the Outcome column

**Congrats, you have officially finished the lab!**

## Lab Feedback

Once you have completed the lab, please fill out the following form to provide feedback for the session so far: <https://forms.microsoft.com/r/FMRdFcuzTz>