SQL Agent and how to run evals:

Phase 1:
Run through the golden set and check answers via code grading
Phase 2:
Run through rubric with LLM as Judge to find linguistic similarities (BLEU scores)
Phase 3:
Run through moderation API to check Tone, Harm

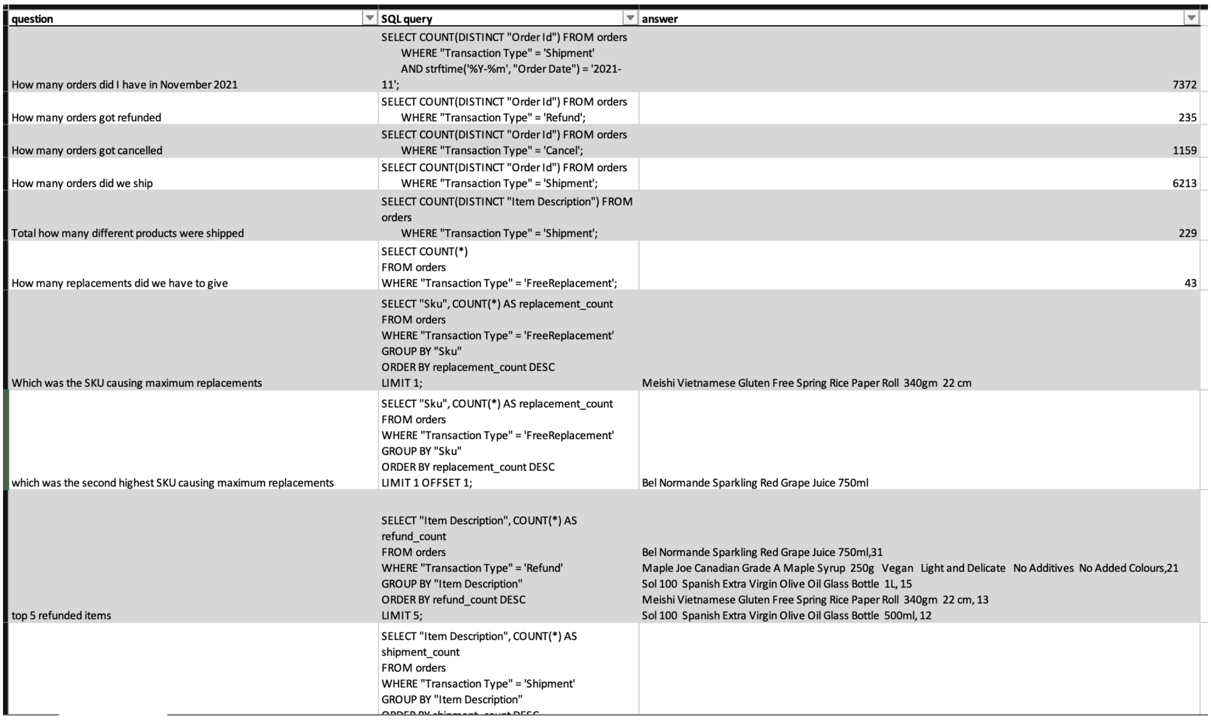
Run	Model	Prompt	Tool
First	Execution	Basic	No
Second	Reasoning	Basic	No
Third	Execution	Basic	Yes
Fourth	Reasoning	Basic	Yes
~~Fifth~~	~~Execution~~	~~Add Schema~~	~~No~~
~~Sixth~~	~~Reasoning~~	~~Add Schema~~	~~No~~
Seventh	Execution	Add Schema	Yes
Eighth	Reasoning	Add Schema	Yes

Per Run metrics:

Metric	Type	Grading per question	Grading Method	Accumulated Score
Coherence/Hallucination avoidance	Making sense from a language perspective	Yes/No	Rubric (LLM as Judge)	Percentage of total
Relevance	Could be sounding very intelligent but besides the point	1-5	Rubric (LLM as Judge)	?
Harm	Aggression,	Yes/No	Rubric (LLM as	Percentage of

	Abuse		Judge)	total
Tone	Formal, Empathetic	Yes/No	Rubric (LLM as Judge)	Percentage of total
Accuracy		Yes/No	Code eval, BLEU Score	Percentage of total
Don’t Know		Yes/No	Code eval	Percentage of total
Tool Use - True Positive	% of total times it was supposed to be called and it did	Ideally 100%	Code eval	% of total times it was supposed to be called and it did
Tool Use - False Positive	Number of times it was called when not needed	Ideally 0 times	Code eval	% of total times it was not supposed to be called and it did

Golden Question Set:



question	SQL query	answer

How many orders did I have in November 2021	SELECT COUNT(DISTINCT "Order Id") FROM orders
	WHERE "Transaction Type" = 'Shipment'
	AND strftime('%Y-%m', "Order Date") = '2021-11';
		7372

How many orders got refunded	SELECT COUNT(DISTINCT "Order Id") FROM orders
	WHERE "Transaction Type" = 'Refund';
		235

How many orders got cancelled	SELECT COUNT(DISTINCT "Order Id") FROM orders
	WHERE "Transaction Type" = 'Cancel';
		1159

How many orders did we ship	SELECT COUNT(DISTINCT "Order Id") FROM orders
	WHERE "Transaction Type" = 'Shipment';
		6213

Total how many different products were shipped	SELECT COUNT(DISTINCT "Item Description") FROM orders
	WHERE "Transaction Type" = 'Shipment';
		229

How many replacements did we have to give	SELECT COUNT(*)
	FROM orders
	WHERE "Transaction Type" = 'FreeReplacement';
		43

Which was the SKU causing maximum replacements	SELECT "Sku", COUNT(*) AS replacement_count
	FROM orders
	WHERE "Transaction Type" = 'FreeReplacement'
	GROUP BY "Sku"
	ORDER BY replacement_count DESC
	LIMIT 1;
		Meishi Vietnamese Gluten Free Spring Rice Paper Roll 340gm 22 cm

which was the second highest SKU causing maximum replacements	SELECT "Sku", COUNT(*) AS replacement_count
	FROM orders
	WHERE "Transaction Type" = 'FreeReplacement'
	GROUP BY "Sku"
	ORDER BY replacement_count DESC
	LIMIT 1 OFFSET 1;
		Bel Normande Sparkling Red Grape Juice 750ml

top 5 refunded items	SELECT "Item Description", COUNT(*) AS refund_count
	FROM orders
	WHERE "Transaction Type" = 'Refund'
	GROUP BY "Item Description"
	ORDER BY refund_count DESC
	LIMIT 5;
		Bel Normande Sparkling Red Grape Juice 750ml,31
		Maple Joe Canadian Grade A Maple Syrup 250g Vegan Light and Delicate No Additives No Added Colours,21
		Sol 100 Spanish Extra Virgin Olive Oil Glass Bottle 1L, 15
		Meishi Vietnamese Gluten Free Spring Rice Paper Roll 340gm 22 cm, 13
		Sol 100 Spanish Extra Virgin Olive Oil Glass Bottle 500ml, 12

	SELECT "Item Description", COUNT(*) AS shipment_count
	FROM orders
	WHERE "Transaction Type" = 'Shipment'
	GROUP BY "Item Description"	

Generating Data:
Manual runs

Experiment	Model Type	Tool Use	Prompt	Run result

Run 1, Question 1	Execution	No	You are an expert business analyst that generates SQL queries and executes them for an amazon orders database.	'run_id': np.int64(1), 'question': 'How many orders did I have in November 2021', 'expected': '7372', 'used_tool': True, 'response': 'SQL ERROR: Execution failed on sql \'SELECT COUNT(*) AS total_orders \nFROM orders \nWHERE EXTRACT(MONTH FROM order_date) = 11 AND EXTRACT(YEAR FROM order_date) = 2021;\': near "FROM": syntax error'}
Run 1, Question 1	Execution	No	Basic+ - You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to sqlite.	{'run_id': np.int64(1), 'question': 'How many orders did I have in November 2021', 'expected': '7372', 'used_tool': False, 'response': "To find out how many orders you had in November 2021, you can use the following SQL query:\n\n```sql\nSELECT COUNT(*) AS total_orders\nFROM orders\nWHERE strftime('%Y-%m', order_date) = '2021-11';\n```\n\nThis query will count the total number of orders that occurred in November 2021 from your `orders` table. You can execute this query in your SQLite database to get the total number of orders in November 2021."}
Run 2, Question 1	Reasoning	No	Basic+ - You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to sqlite.	{'run_id': np.int64(2), 'question': 'How many orders did I have in November 2021', 'expected': '7372', 'used_tool': False, 'response': "To find out how many orders you had in November 2021, you can use the following SQL query:\n\n```sql\nSELECT COUNT(*) AS total_orders\nFROM orders\nWHERE strftime('%Y-%m', order_date) = '2021-11';\n```\n\nIn this query:\n- We are selecting the count of all entries from the

				`orders` table where the `order_date` column falls within November 2021.\n- We are using the `strftime` function to extract the year and month from the `order_date` column and compare it to '2021-11' for November 2021.\n\nYou can execute this query against your Amazon orders database to obtain the total number of orders you had in November 2021."}
Run 3, Question 1	Execution	Yes	Basic+ - You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to sqlite.	{'run_id': np.int64(3), 'question': 'How many orders did I have in November 2021', 'expected': '7372', 'used_tool': True, 'response': "SQL ERROR: Execution failed on sql 'SELECT COUNT(*) FROM orders WHERE strftime('%Y-%m', order_date) = '2021-11'": no such column: order_date"}
Run 3, Question 1	Execution	Yes	Basic++ - You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to sqlite.When referring to	{'run_id': np.int64(3), 'question': 'How many orders did I have in November 2021', 'query': 'SELECT COUNT(*) AS num_orders FROM orders WHERE strftime(\'%Y\', "Order Date") = \'2021\' AND strftime(\'%m\', "Order Date") = \'11\'', 'expected': '7372',

			column names, always use double quotes (e.g., \"Order Date\") if the column name contains spaces or special characters.	'used_tool': True, 'response': '7539'} **Accuracy: 97.73%**
Run 4, Question 1	Reasoning	Yes	Basic++ - You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to sqlite.When referring to column names, always use double quotes (e.g., \"Order Date\") if the column name contains spaces or special characters.	{'run_id': np.int64(4), 'question': 'How many orders did I have in November 2021', 'query': 'SELECT COUNT(*) AS order_count FROM orders WHERE strftime(\'%Y-%m\', "Order Date") = \'2021-11\'', 'expected': '7372', 'used_tool': True, 'response': '7539'} **Accuracy:97.73%**
Run 4, Question 2	Reasoning	Yes	Basic++ - You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to	{'run_id': np.int64(4), 'question': 'How many orders got refunded', 'query': 'SELECT COUNT(*) AS refund_count FROM orders WHERE "Order Status" = \'Refunded\'', 'expected': '235', 'used_tool': True, 'response': '0'}

			sqlite.When referring to column names, always use double quotes (e.g., \"Order Date\") if the column name contains spaces or special characters.	**Accuracy:0%**
Run 3/4, Question 2	Execution/Reasoning	Yes	Basic++ WITH BUSINESS LOGIC- You are an expert business analyst that generates SQL queries and executes them for an amazon orders database. This database is in sqlite format so ensure your SQL queries cater to sqlite. When referring to column names, always use double quotes (e.g., "Order Date") if the column name contains spaces or special characters. Use the business logic For sales or orders, the Transaction Type is 'Shipment'. For refunds, the Transaction Type is 'Refund'. For Cancellations, the Transaction Type is 'Cancel'. For Replacements, the Transaction Type is 'FreeReplacemen	{'run_id': np.int64(4), 'question': 'How many orders got refunded', 'query': 'SELECT COUNT(*) as num_refunds FROM orders WHERE "Transaction Type" = \'Refund\'', 'expected': '235', 'used_tool': True, 'response': '235'} **Accuracy:100%**

			t'	
Run 3/4, Question 2	Execution/Reasoning	Yes	“”	{'run_id': np.int64(3), 'question': 'How many orders got cancelled', 'query': 'SELECT COUNT(*) AS \'Cancelled Orders\' FROM Orders WHERE "Transaction Type" = \'Cancel\';', 'expected': '1159', 'used_tool': True, 'response': '1159'}