Gen AI at Swiggy

Jairaj Sathyanarayana
VP Data Science & Analytics [https://linkedin.com/in/jairajs]

Abhinav Nigam
Director of Engineering [https://linkedin.com/in/nigamabhinav/]

PROSUS AI MARKETPLACE 2023
LLMS IN PRODUCTION

A systematic approach to institutionalizing gen AI

- Established a gen AI task-force (in Q1 of 2023)

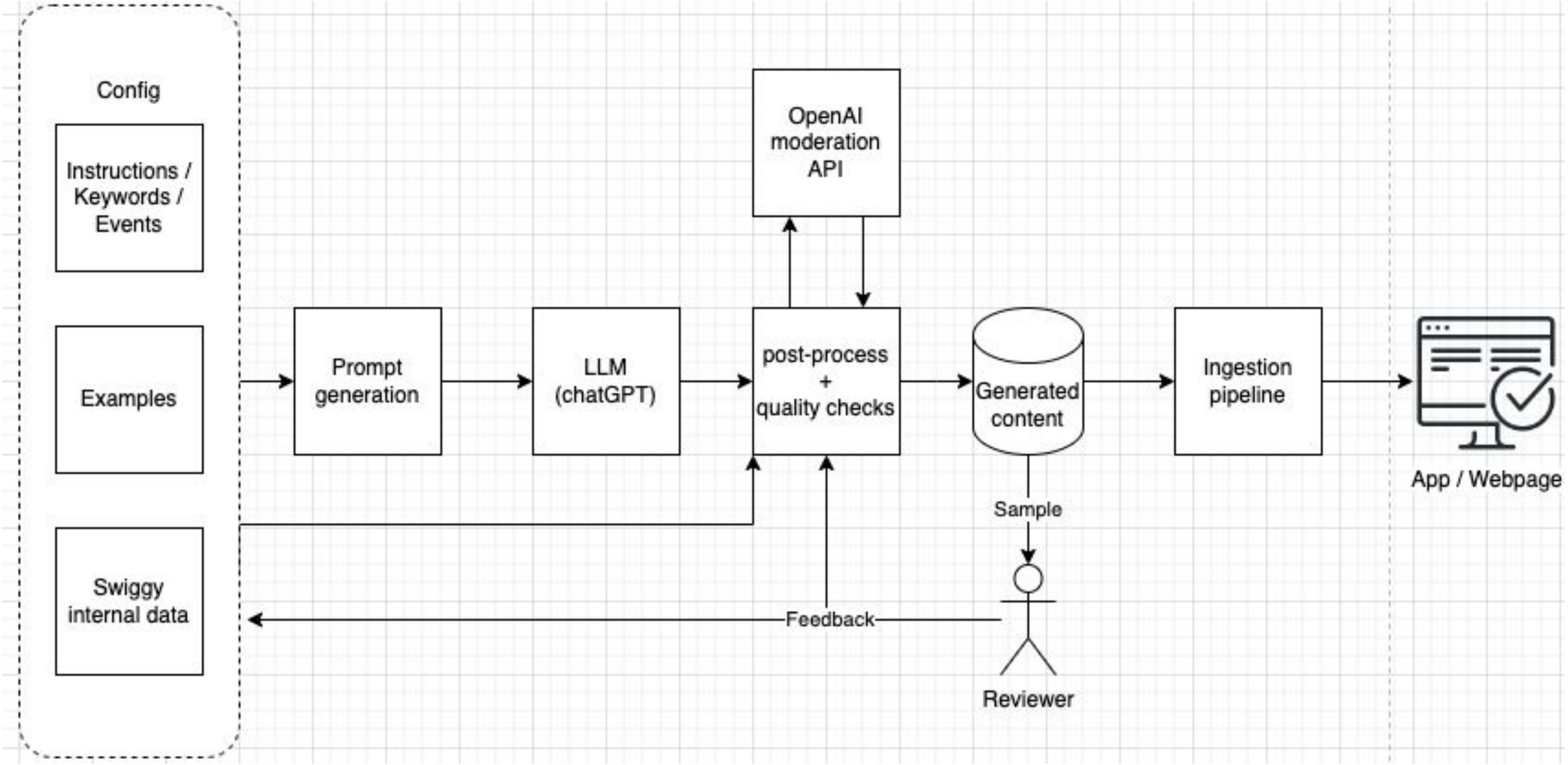
- Based on learnings, set up gen AI as a company-wide thrust with a single-threaded leader (starting Q2 of 2023)

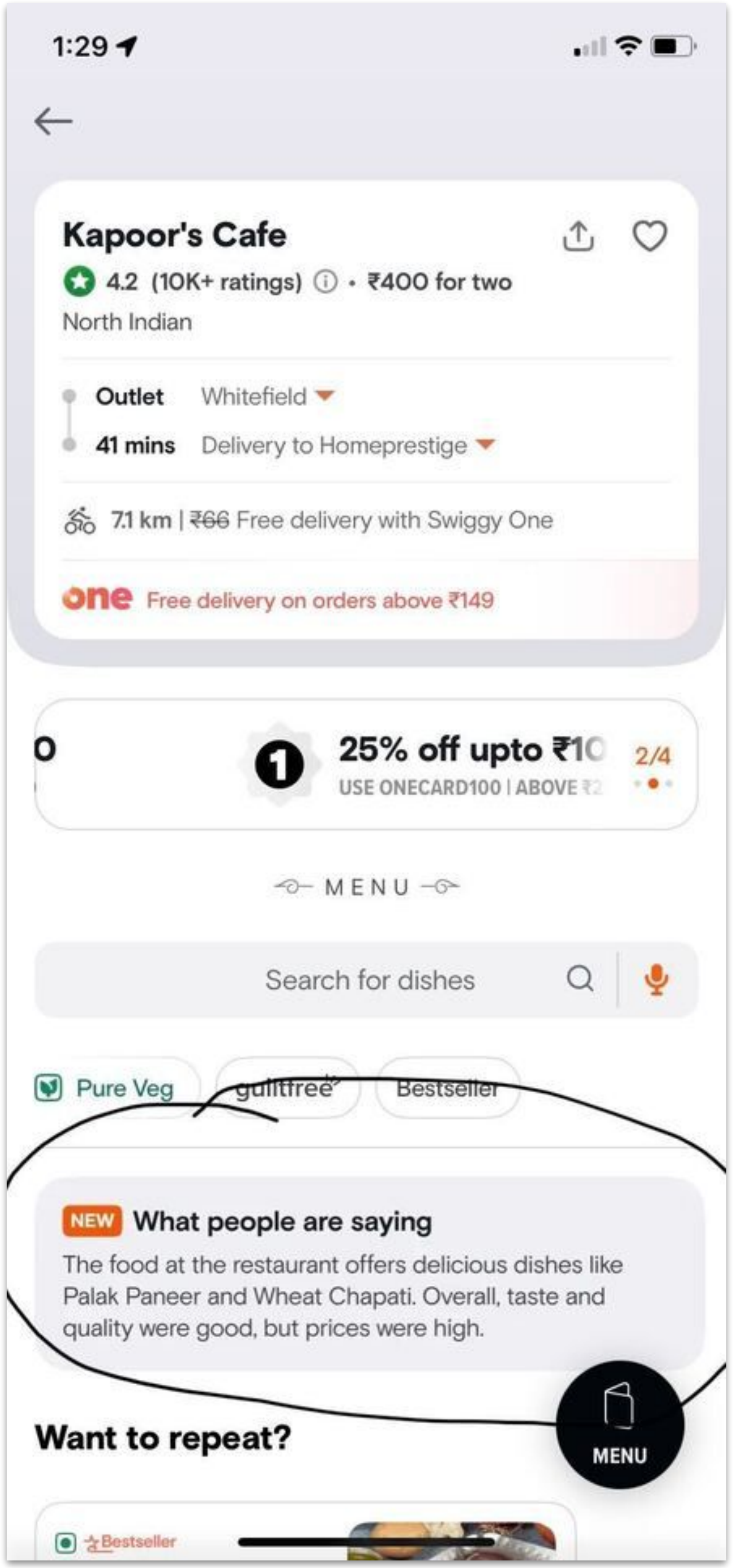
- Run the thrust like any other Annual Operating Plan (AOP) program… but with some flexibility for course corrections and fuzzy bets
  - Funding-- headcount, program management
  - Evangelization & enablement-- hackathons, tools
  - Monthly reviews with the CXO team

- 3 broad classes of gen AI initiatives
  - Discovery & Search
  - Productivity (excluding no-brainer use-cases like adopting Github Copilot)
  - Automation

Discovery & Search track: Catalog Enrichment

Currently in production: Text generation (GPT3.5 Turbo)





- Prompt engineering + pre/post processing
- Summary/ text needs to be truthful but not overly negative

500K+ descriptions (99%+ acceptance rate); 1K+ SEO pages

A/B with a few thousand restaurants under way in MUM, HYD

Discovery & Search track: Catalog Enrichment

Image generation using ControlNet-guided base SD

How to generate images in the style of a given restaurant, at scale? Prompt engineering led solutions are not scalable (hard to control, not realistic, Indian context, etc.)

Item image (real/generated)

Background (real/generated)

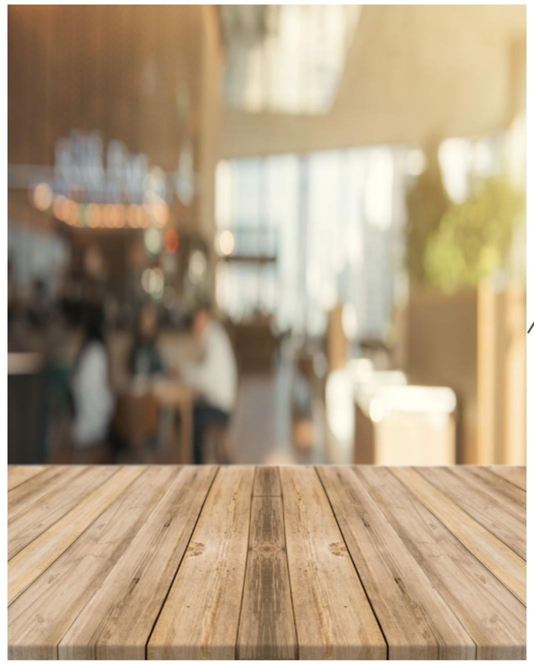
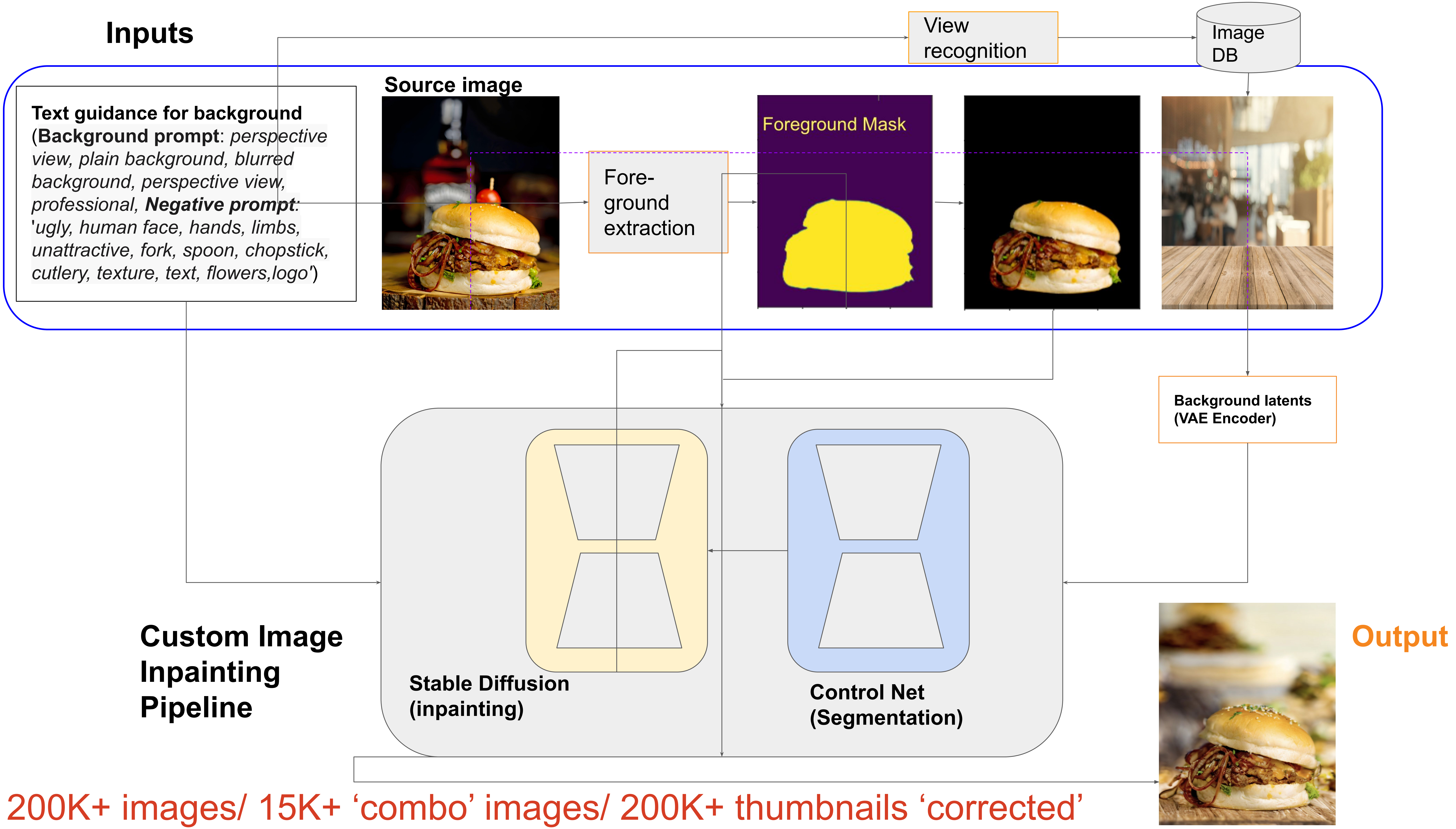


Image blending

Fixed text guidance for background (**Background prompt**: *perspective view, plain background, blurred background, perspective view, professional*, ***Negative prompt***: *'ugly, human face, hands, limbs, unattractive, fork, spoon, chopstick, cutlery, texture, text, flowers,logo'*)

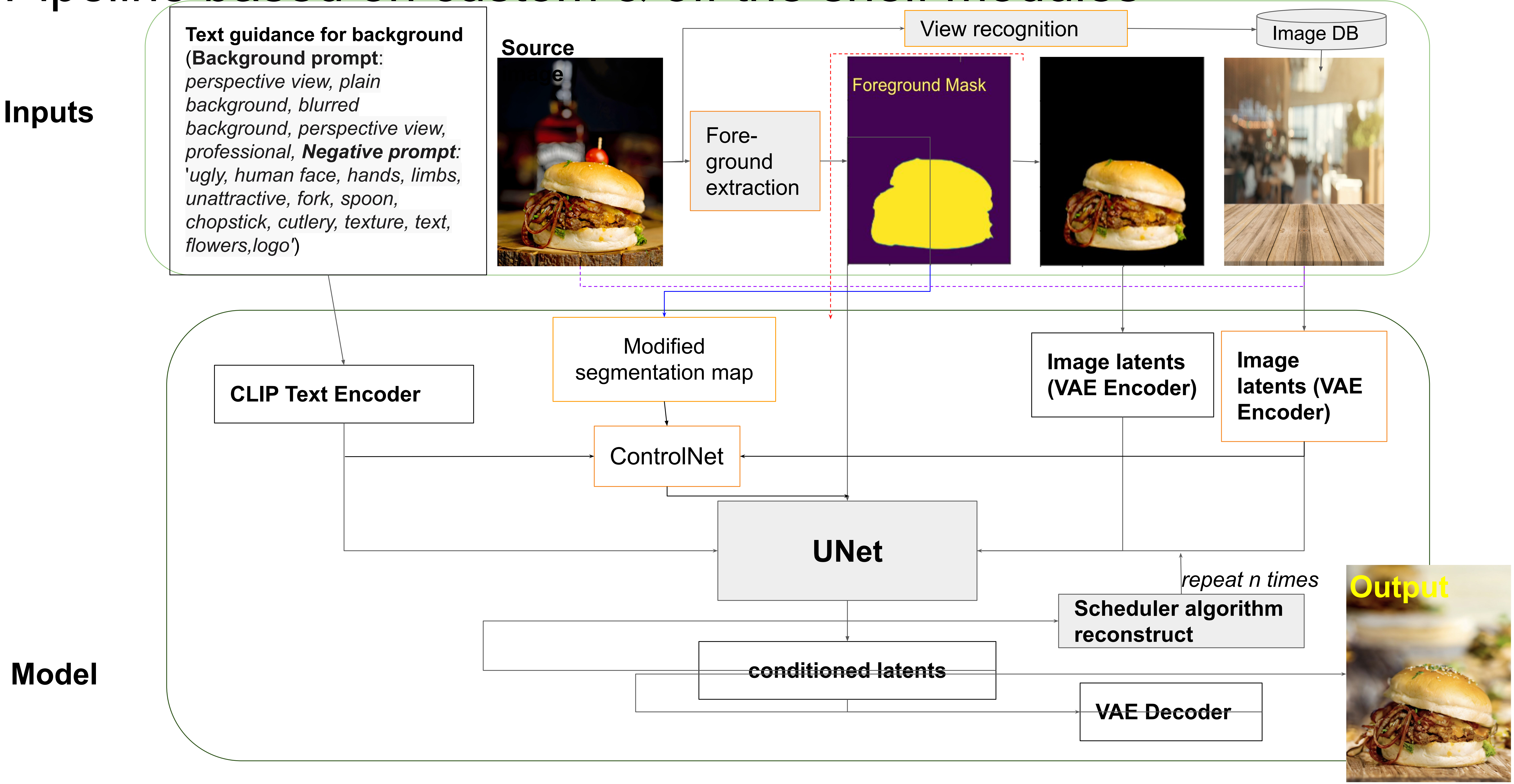
Item blended in the style of the given background (i.e., in the style of a given restaurant)



Inputs
View recognition
Image DB
Source image
Text guidance for background (Background prompt: perspective view, plain background, blurred background, perspective view, professional, Negative prompt: 'ugly, human face, hands, limbs, unattractive, fork, spoon, chopstick, cutlery, texture, text, flowers,logo')
Fore-ground extraction
Foreground Mask
Background latents (VAE Encoder)
Custom Image Inpainting Pipeline
Stable Diffusion (inpainting)
Control Net (Segmentation)
Output
200K+ images/ 15K+ 'combo' images/ 200K+ thumbnails 'corrected'

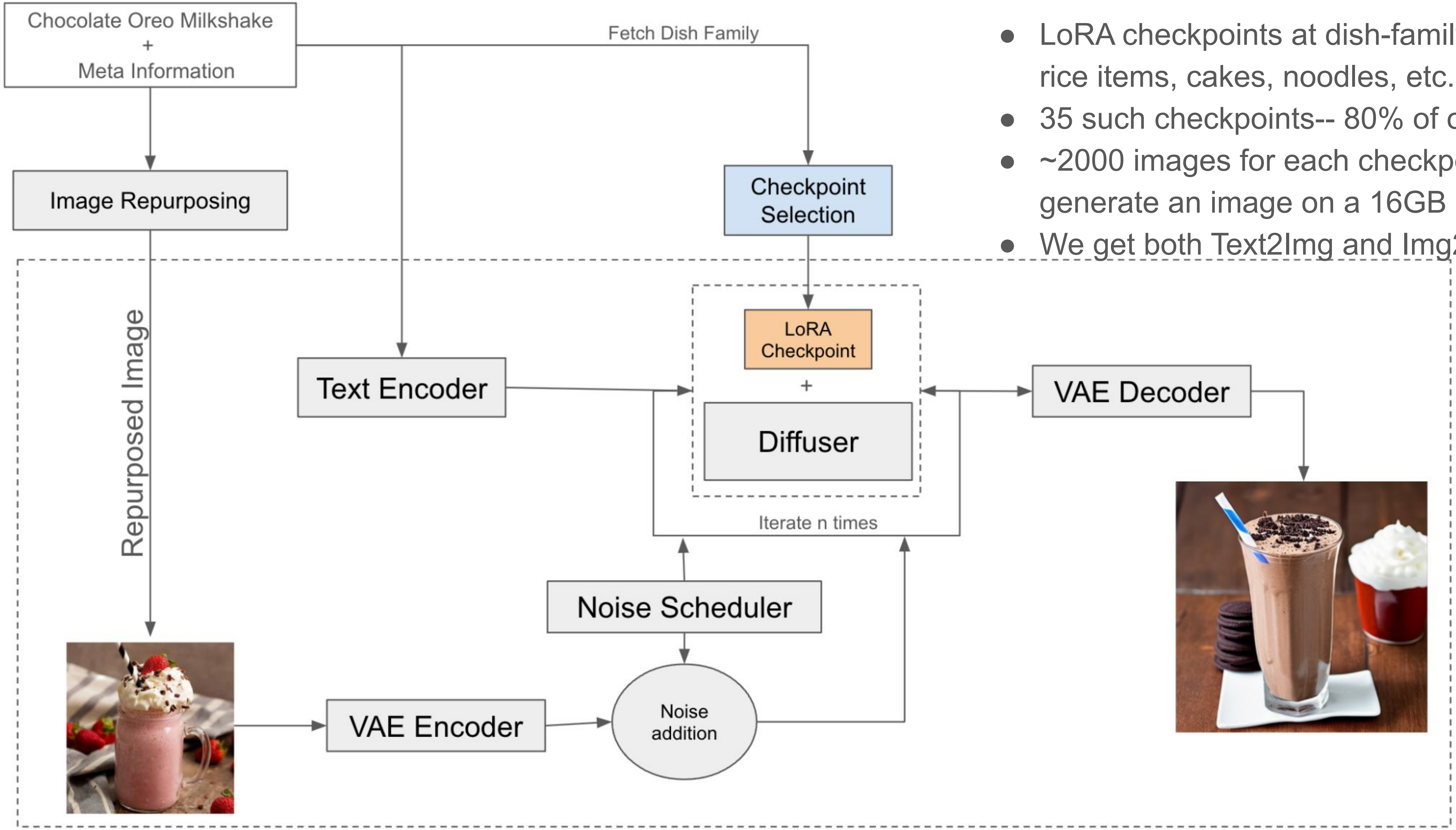
Discovery & Search track: Catalog Enrichment

Pipeline based on custom & off the shelf modules



Discovery & Search track: Catalog Enrichment

Image gen: We can do better by LoRA-tuning SD on our data



- LoRA checkpoints at dish-family level (for ex, pizzas, rice items, cakes, noodles, etc.)
- 35 such checkpoints-- 80% of order volume
- ~2000 images for each checkpoint training; 8 sec to generate an image on a 16GB GPU
- We get both Text2Img and Img2Img capabilities

Discovery & Search track: Catalog Enrichment

Image gen: We can do better by LoRA-tuning SD on our data

Txt2Img

Grilled Chicken Burger

Img2Img

Tandoori Naan

Discovery & Search track: Catalog Enrichment

Image gen: We can do better by LoRA-tuning SD on our data

Name: **Dry Chilli Chicken**	Image Blending	Img2Img
Reference Image 		

Discovery & Search track: Catalog Enrichment

Image gen: We can do better by LoRA-tuning SD on our data

Pipeline	Approval rate	Best-Image tag
Image Blending	51%	29%
Txt2Img	62%	26%
Img2Img	69%	45%

Discovery & Search track: Enhancing our Search

Traditional search

Customer has exact requirements
1. **Aloo paratha**
2. **Paneer butter masala**
3. **Mushroom Biryani**

Search query

Dish related | Restaurant related

1. Keyword based matching
2. Key phrase based matching
3. Word embeddings of dish items

Neural search

Customer is trying to explore
1. **Healthy juice after workout**
2. **Diabetic friendly ice creams**
3. **Karnataka special sweets**

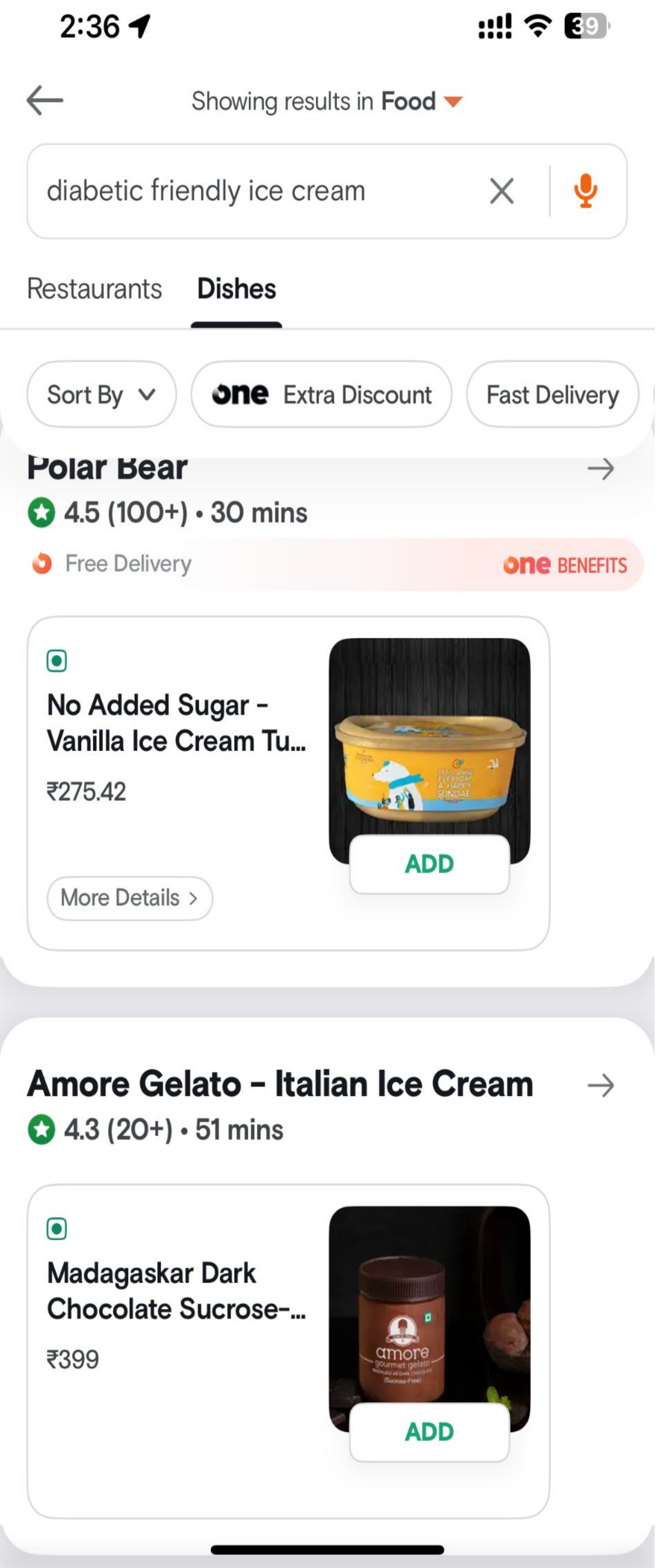
Natural language queries

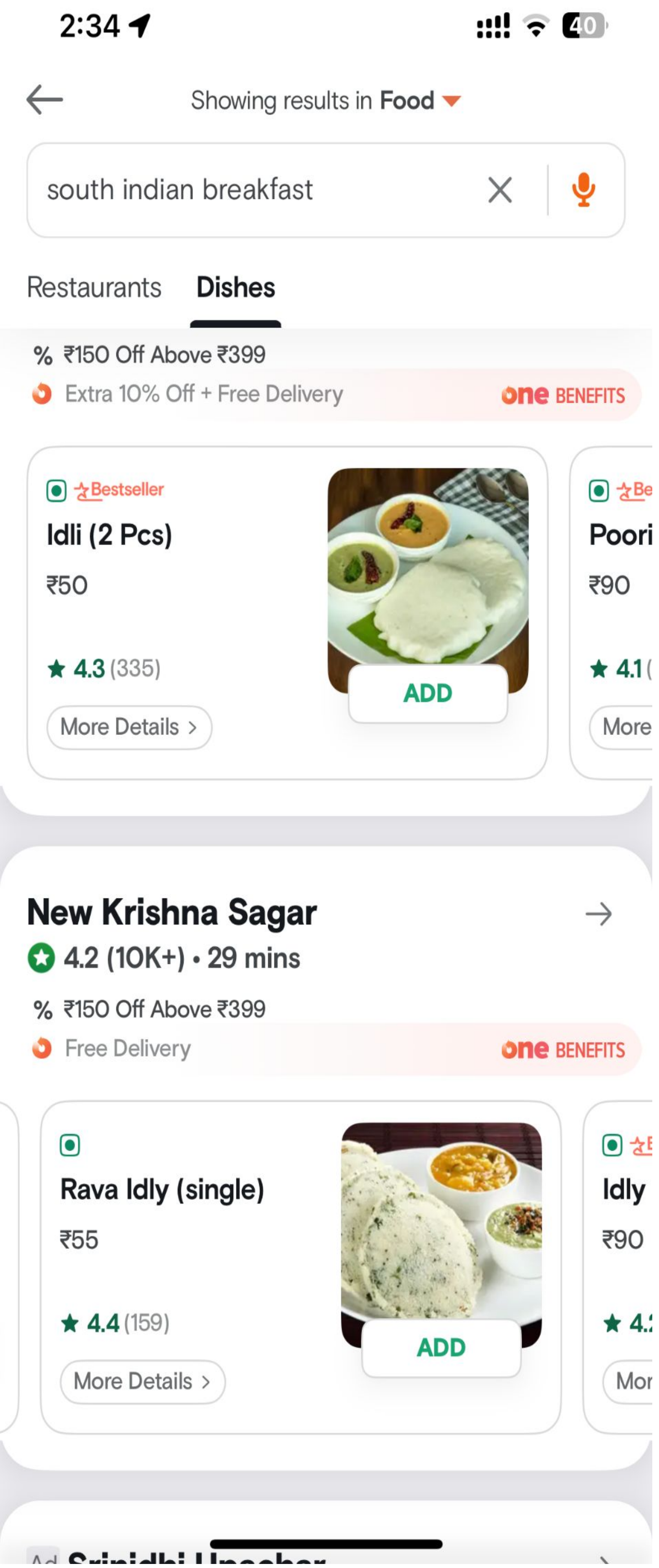
Language models

Context aware embeddings across data sources

1. Item + Meta data descriptions
2. Vectorization (>768 dimensions)
3. Vector similarity engine

Constraints: Latency (P95 <70-100ms), CPU inference, Swiggy domain vocabulary



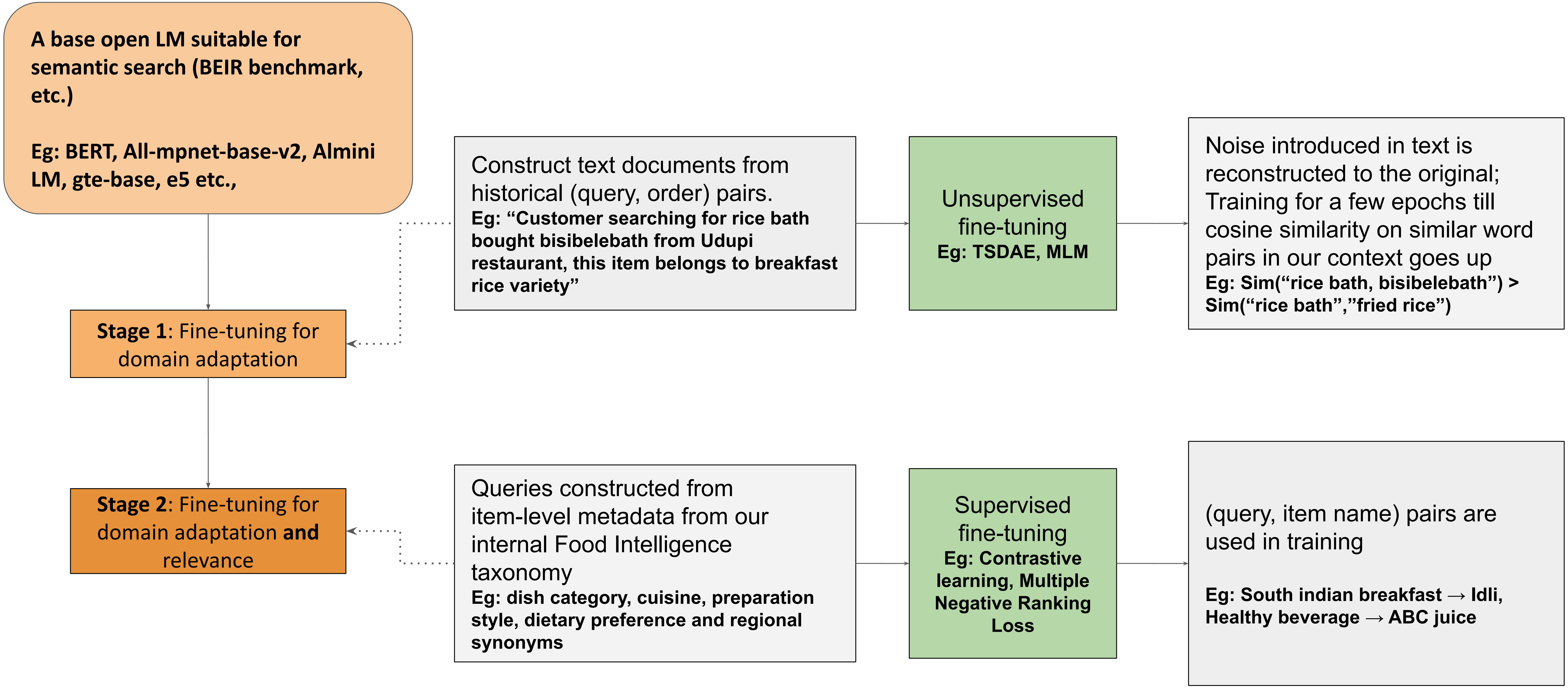


Under A/B on 10% of Search traffic in BLR

Discovery & Search track: Enhancing our Search

The ML pipeline: fine-tuned LM based on all-mpnet-base-v2

(100M+ params)

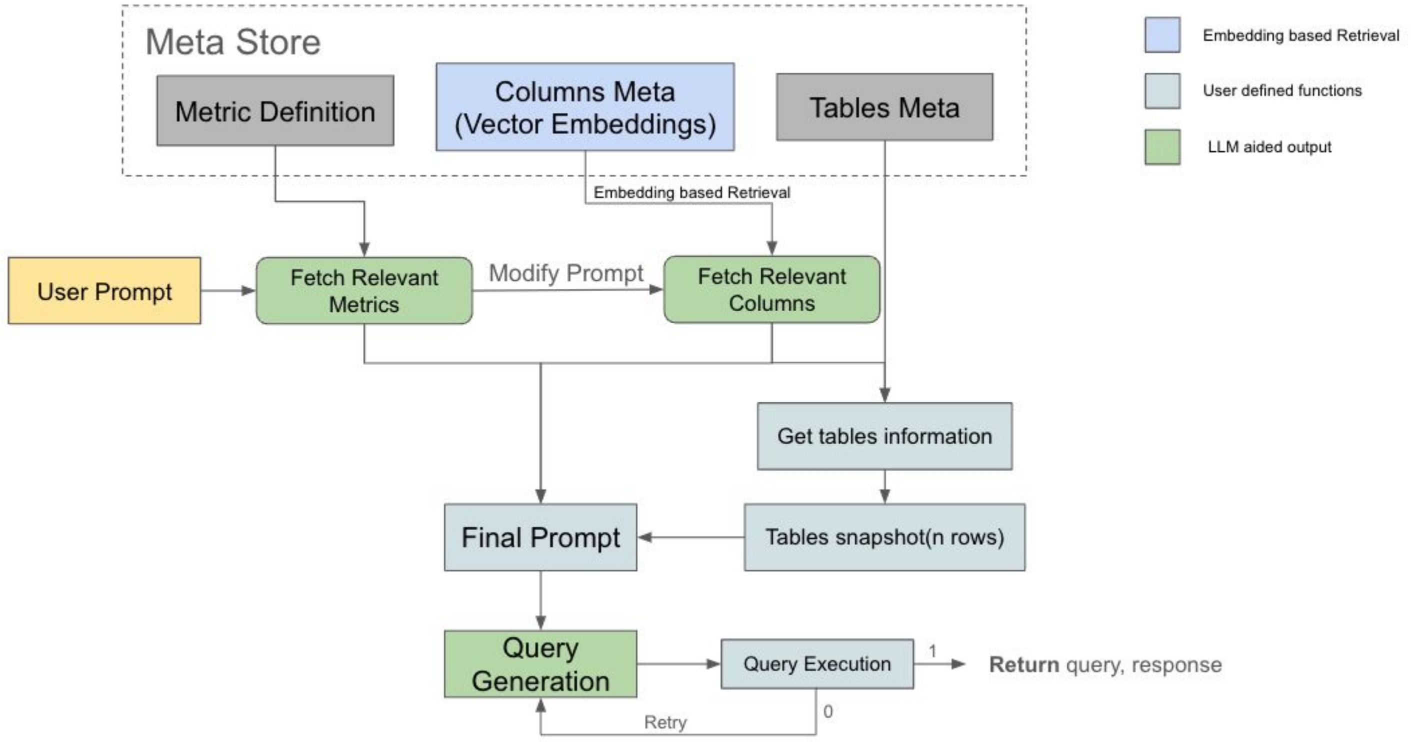


Productivity track: Conversational Analytics

Text2Text
Summarization, Q&A on KB
(GPT 3.5)

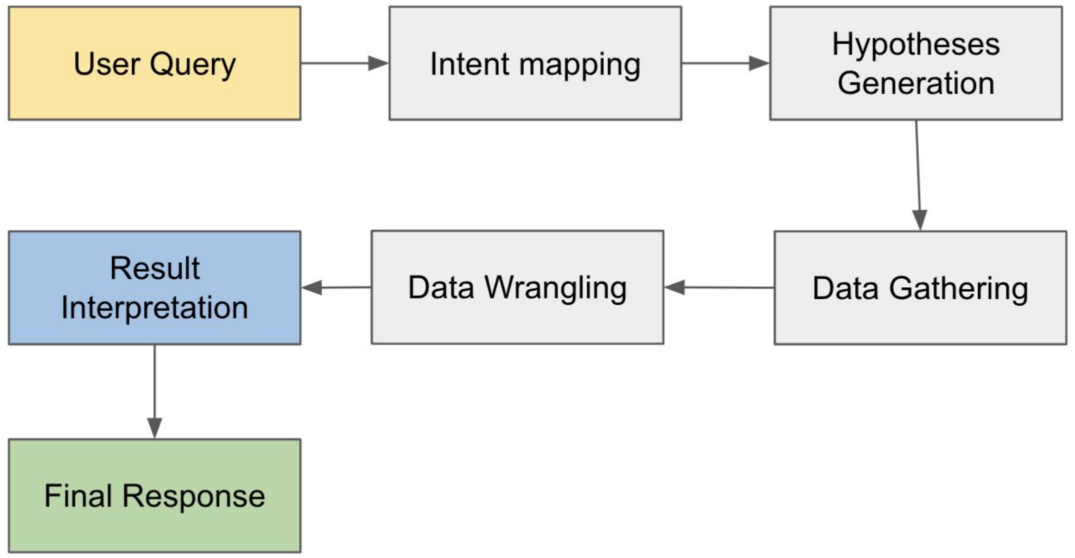
Rolling out to a few thousand restaurants

Text2SQL (GPT 4)



70-90% accurate in tests; In beta with select domains--trending closer to 50%

Text2Insight (GPT 4)

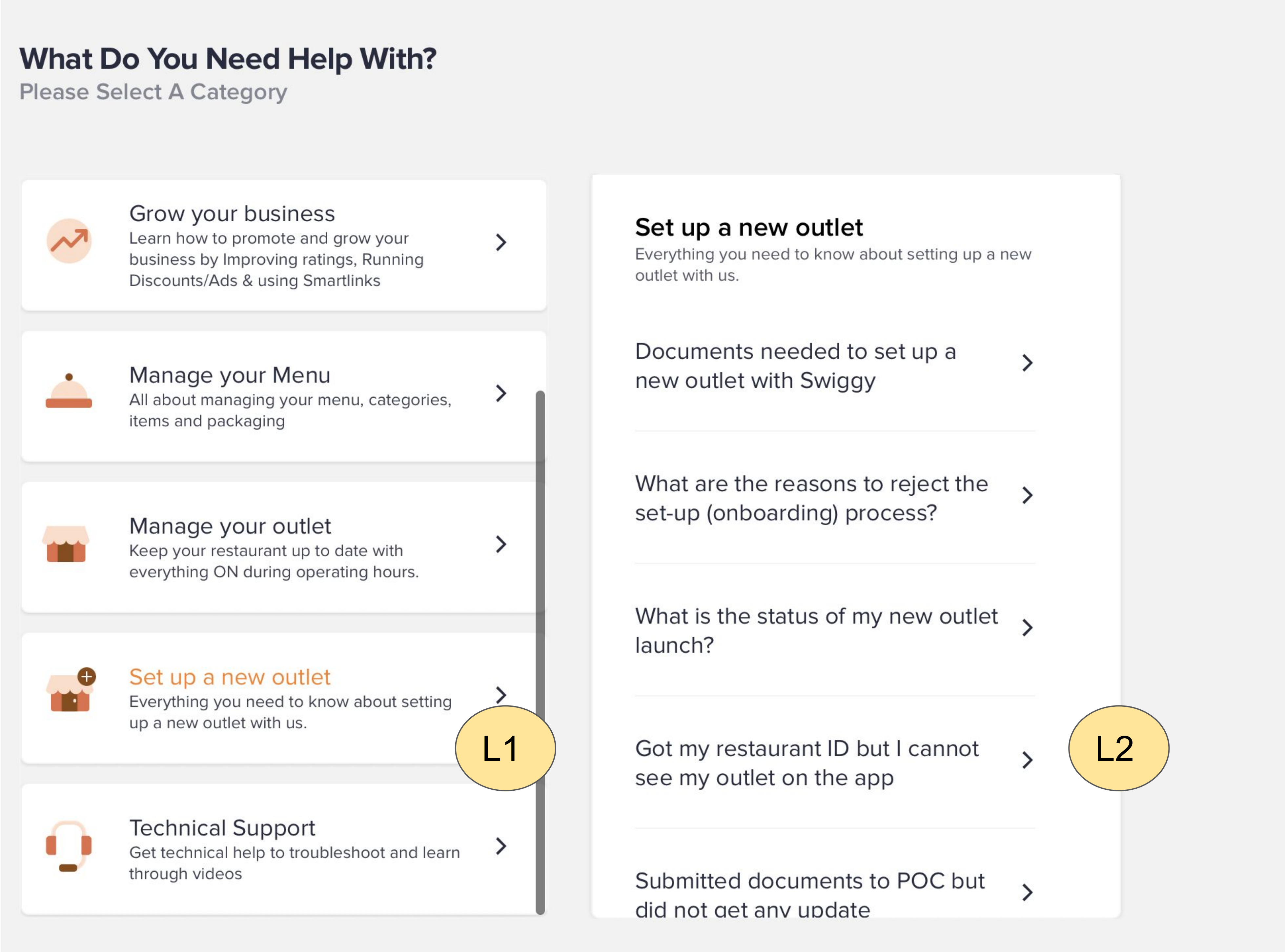


Data/ Biz context is the long pole; Tech is doable

Productivity track: Conversational Analytics

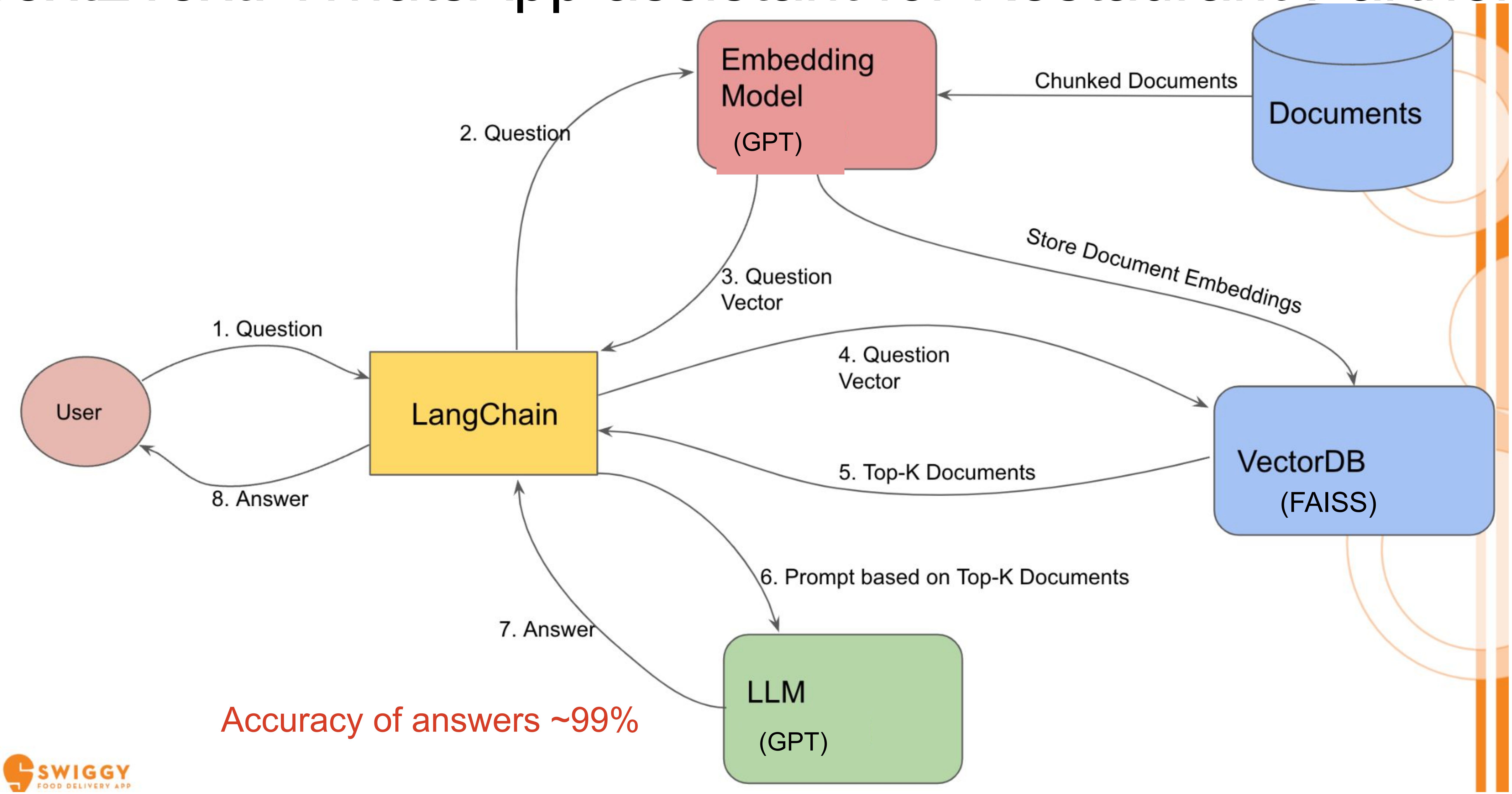
Text2Text: WhatsApp assistant for Restaurant Partners

- Restaurant owners have questions like ‘How do I setup a new outlet’, ‘How do I mark an item out of stock’, ‘How do claims affect my ratings’, etc.
- Navigating through a maze of dispositions on Help Center is not easy
- Conversational Q&A can help improve experience, reduce TAT and load on support teams



Productivity track: Conversational Analytics

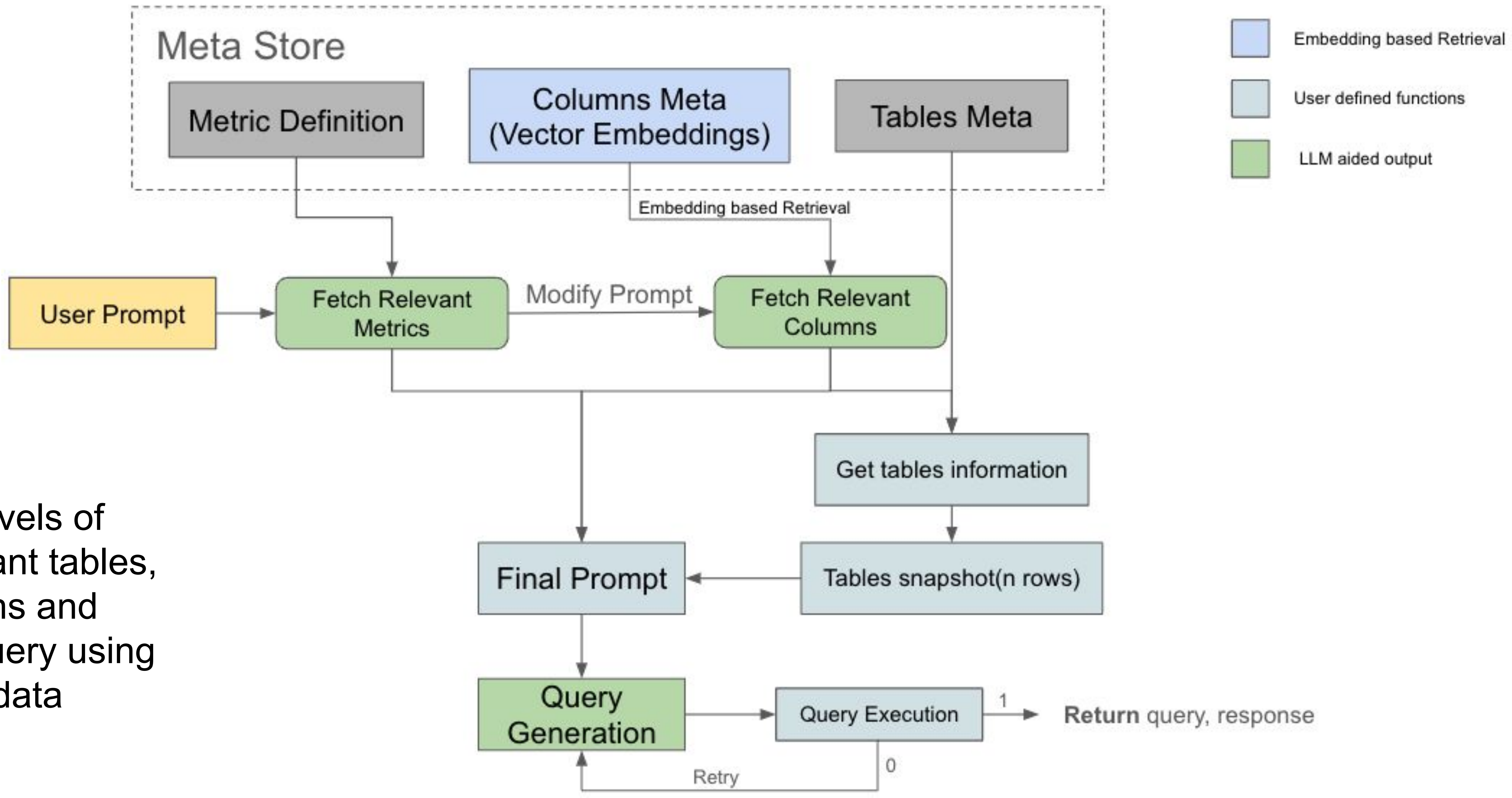
Text2Text: WhatsApp assistant for Restaurant Partners



Accuracy of answers ~99%

Productivity track: Conversational Analytics

Text2SQL: Intermediate-level SQL based self-serve via Slack

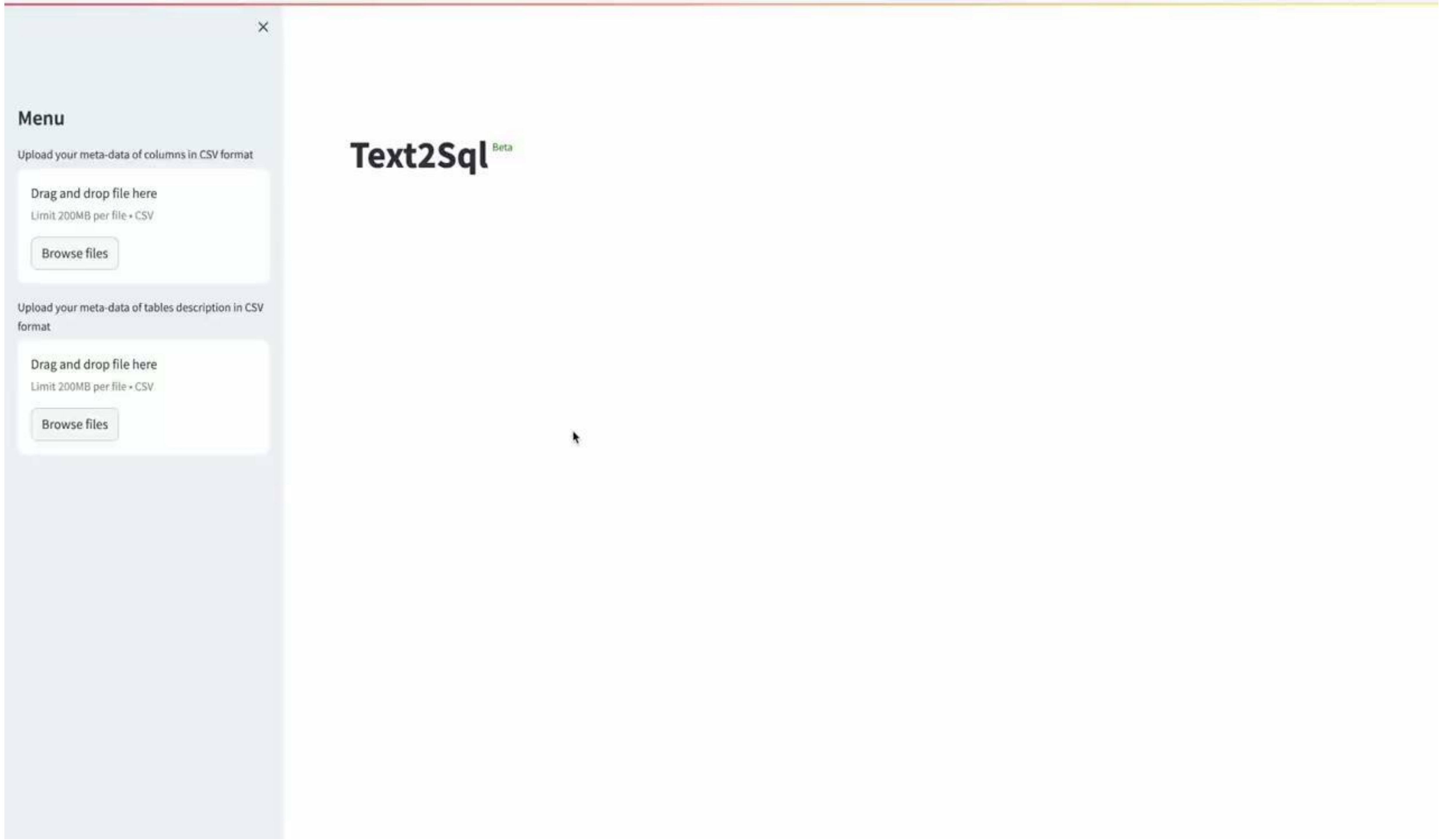


RAG with multiple levels of retrieval: fetch relevant tables, then relevant columns and construct the final query using meta of the fetched data

Productivity track: Conversational Analytics

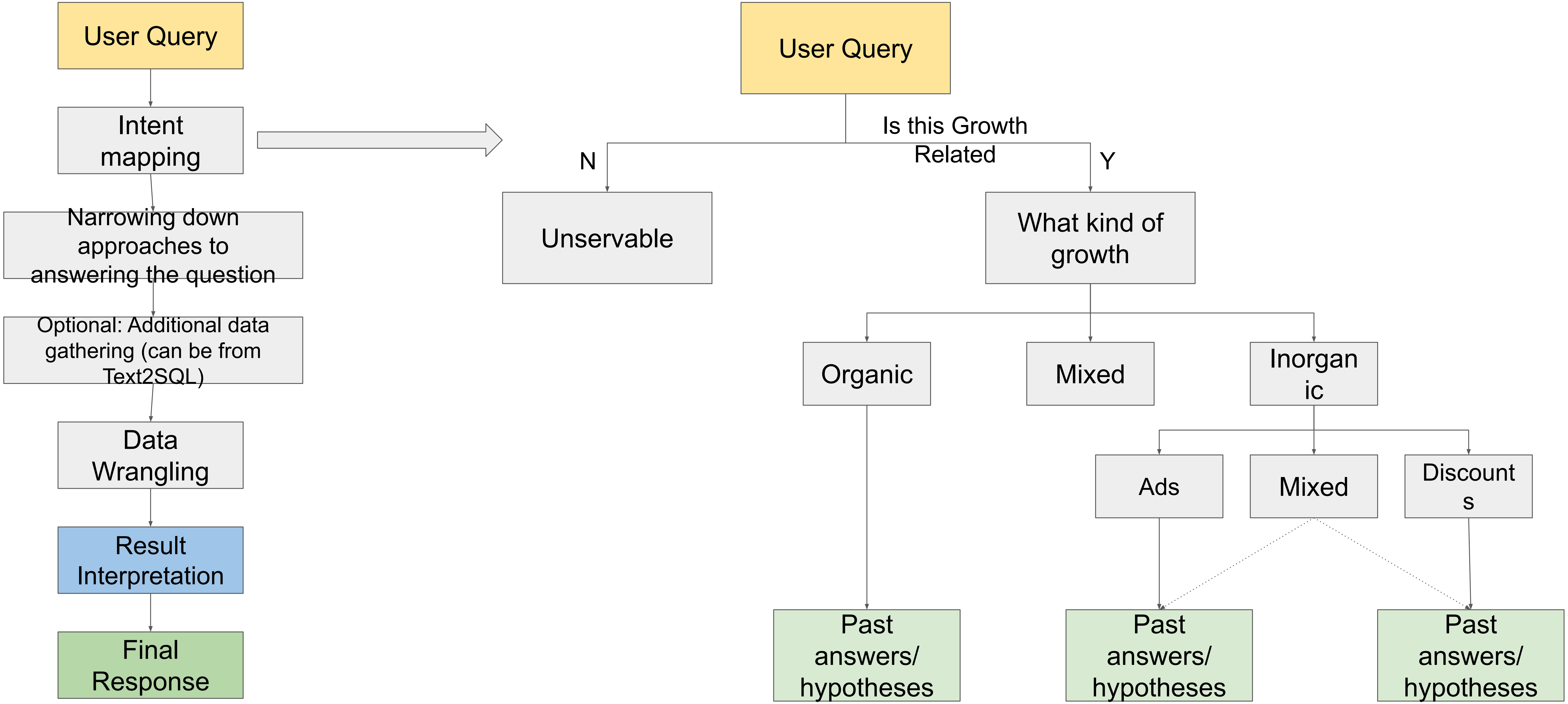
Text2SQL: A quick demo

Task	Accuracy
Simple logic queries	90%
Complex logic queries (multiple joins, JSON processing)	70%



Productivity track: Conversational Analytics

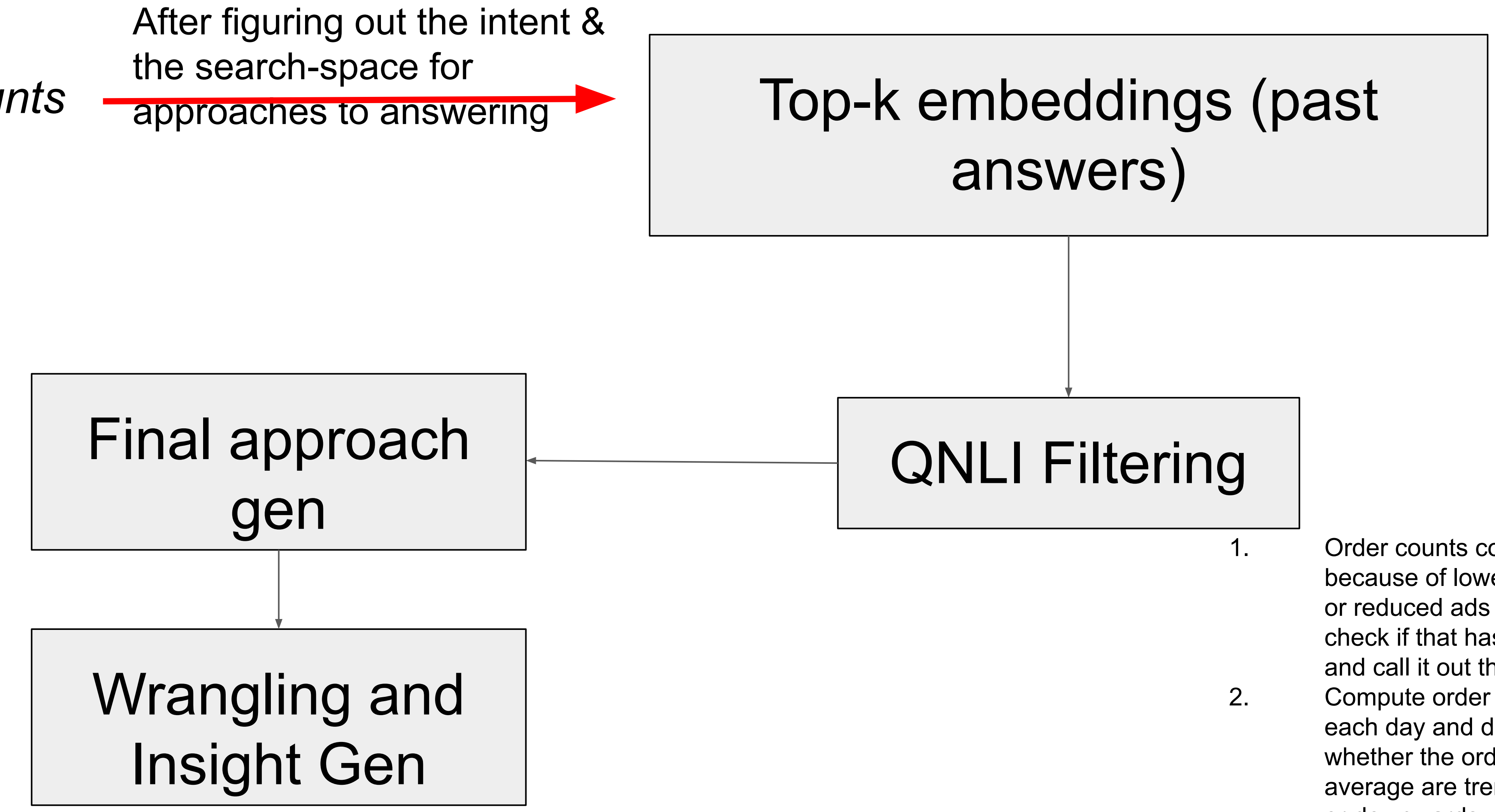
Text2Insight: Relies heavily on codifying Biz hypotheses



Productivity track: Conversational Analytics

Text2Insight: From hypotheses to answers

What are some of the reasons my order counts been ***<u>increasing</u>*** *day over day?*



['Order counts could drop because of lowered discounts or reduced ads spending, check if that has happened and call it out the user.',
"Order counts are also a function of days of the week, compare the order counts with the same days of the week in the previous week. If that's comparable,that is within 30% margin, then there is no actual drop and call that out.",
'To answer any questions regarding AOV values, using the bill_amount column to aggregate averages.',
'Drop in average order values could be due to increased average discounts that might reduce the overall order value.',
'Drop in average order values could be due to increased sales proportion of a cheaper item or decreased sales proportion of an expensive item in the current duration as opposed to a few days before which can be the reference duration. Compare the sales the top selling item and their respective prices between time periods and if that indicates a drop in AOV, call that out.',
'Ads performance could see a dip in case there are other restaurants that are similar to yours in terms of cuisines offered that have increased their ads spending in recent time.',
'Compute order counts for each day and determine whether the orders counts on average are trending upwards or downwards. Also call out if order volume has dropped significantly(>50% of average daily counts) for any particular days as an anomalous day in terms of order volume.']

1. Order counts could drop because of lowered discounts or reduced ads spending, check if that has happened and call it out the user.
2. Compute order counts for each day and determine whether the orders counts on average are trending upwards or downwards. Also call out if order volume has dropped significantly(>50% of average daily counts) for any particular days as an anomalous day in terms of order volume.

Order counts could increase because of higher discounts or increased ads spending, check if that has happened and call it out to the user. Also, compute order counts for each day and determine whether the orders counts on average are trending upwards. If there is a significant increase (>50% of average daily counts) for any particular days, call it out as an exceptional day in terms of order volume.

The days with order counts significantly higher than the average (more than 1.5 times the average) are:

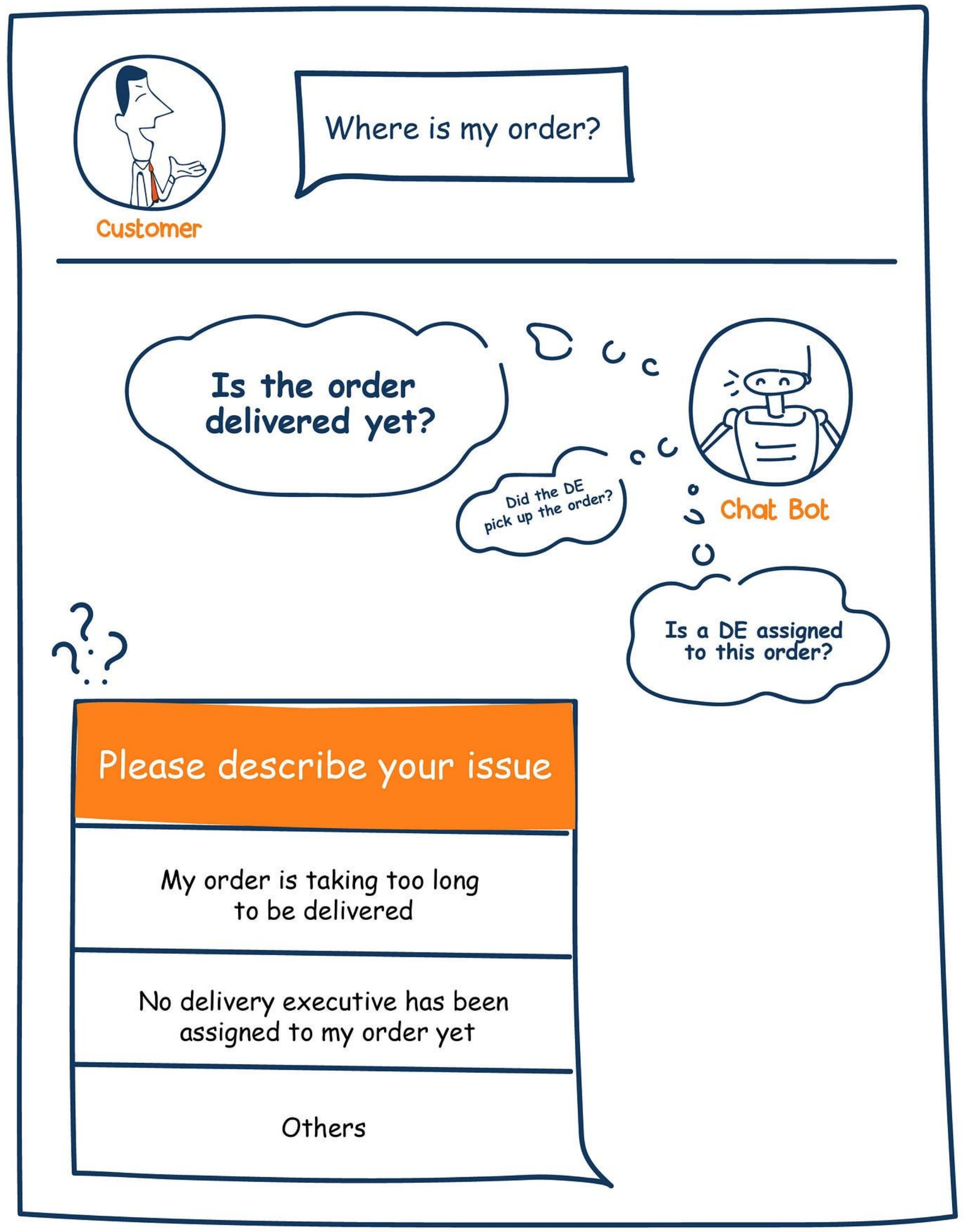
```
2023-06-11: 911 orders
2023-06-18: 1078 orders
2023-06-25: 1039 orders
2023-07-02: 1012 orders
```

These days can be considered as exceptional in terms of order volume.

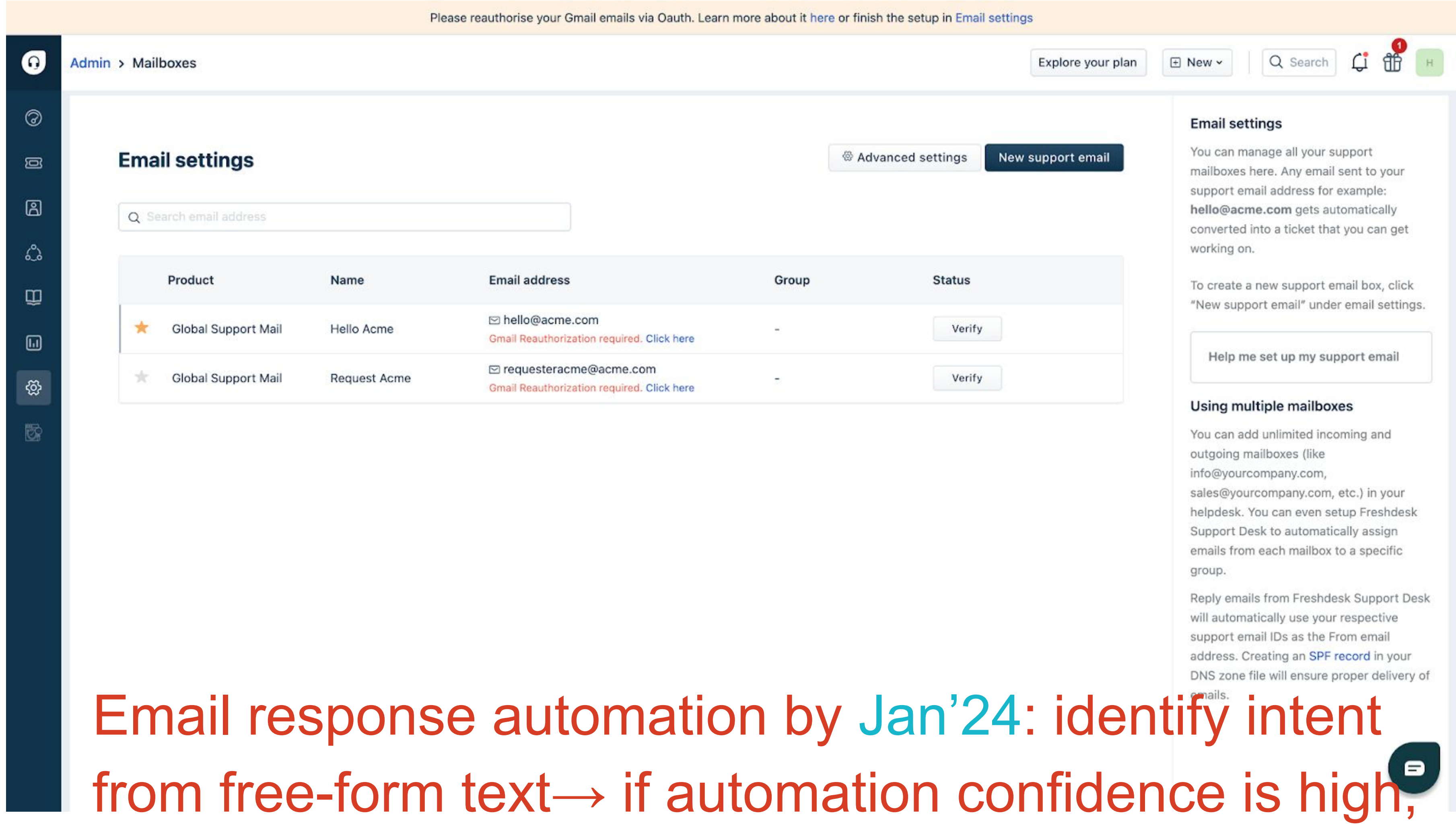
In conclusion, the increase in order counts could be due to the higher discounts given on certain days. The days with significantly higher order counts also coincide with the days with higher discounts. This suggests that discounts could be driving the increase in order counts.

Automation track

Customer Service Chatbot



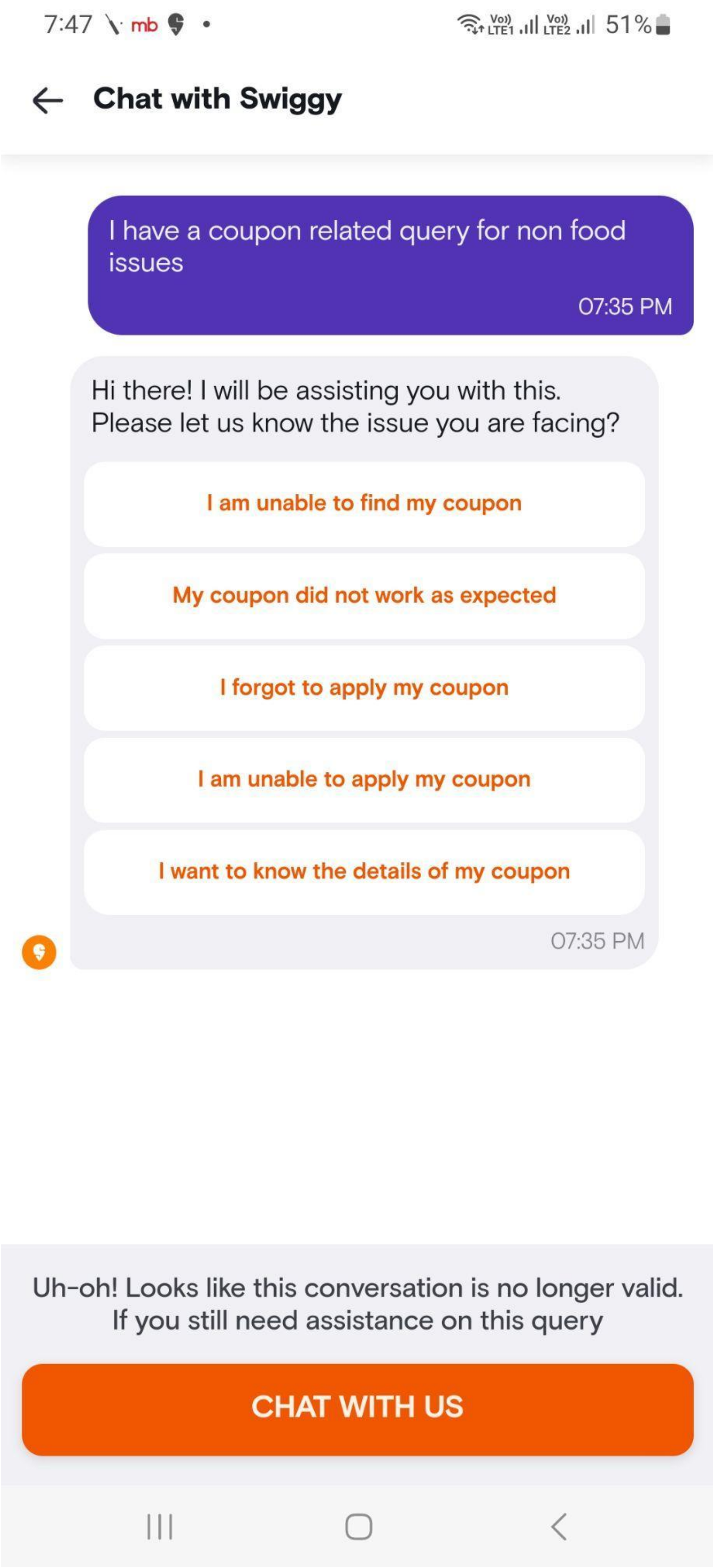
A small % is pushed to email



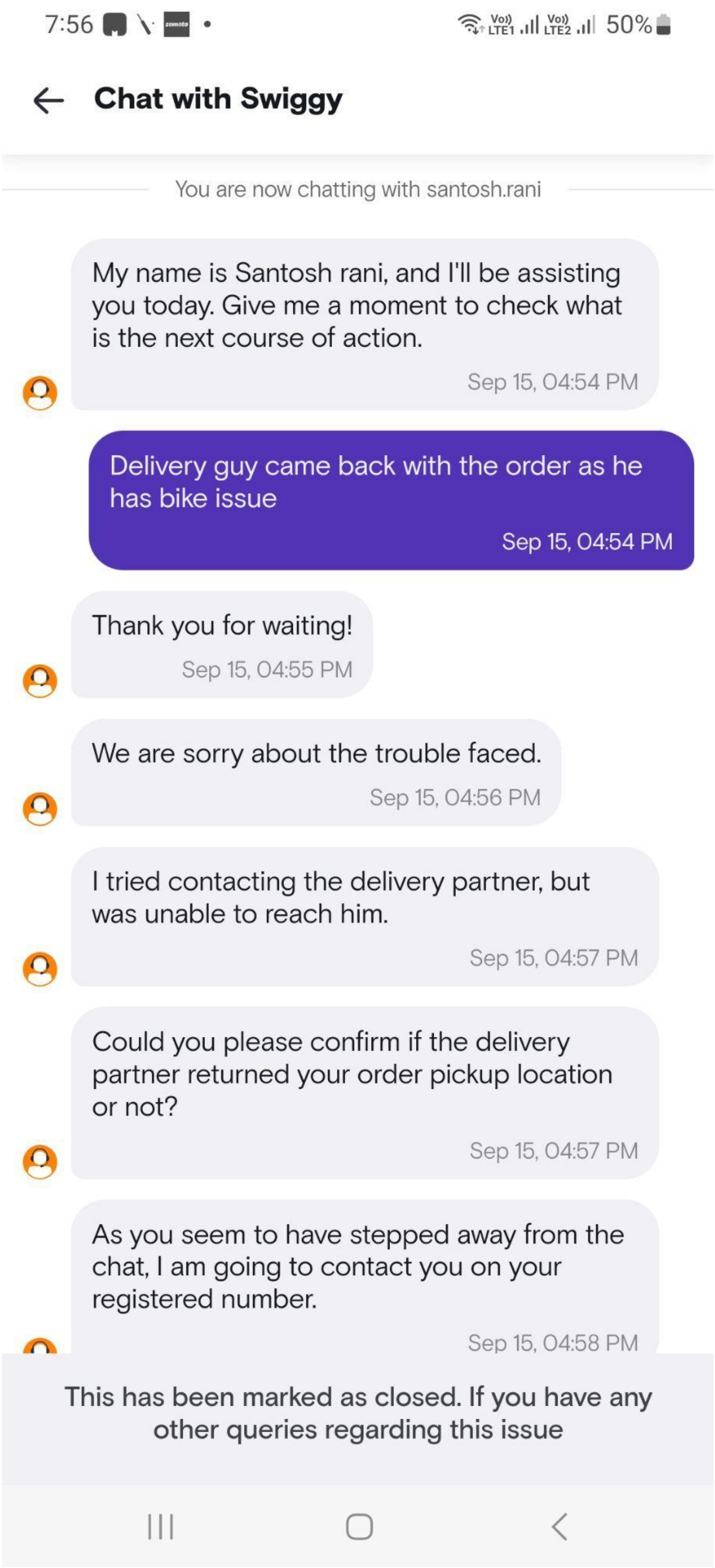
Automation of dispositions that do NOT require a lot of context

Email response automation by Jan’24: identify intent from free-form text→ if automation confidence is high, execute SOP, generate empathetic response

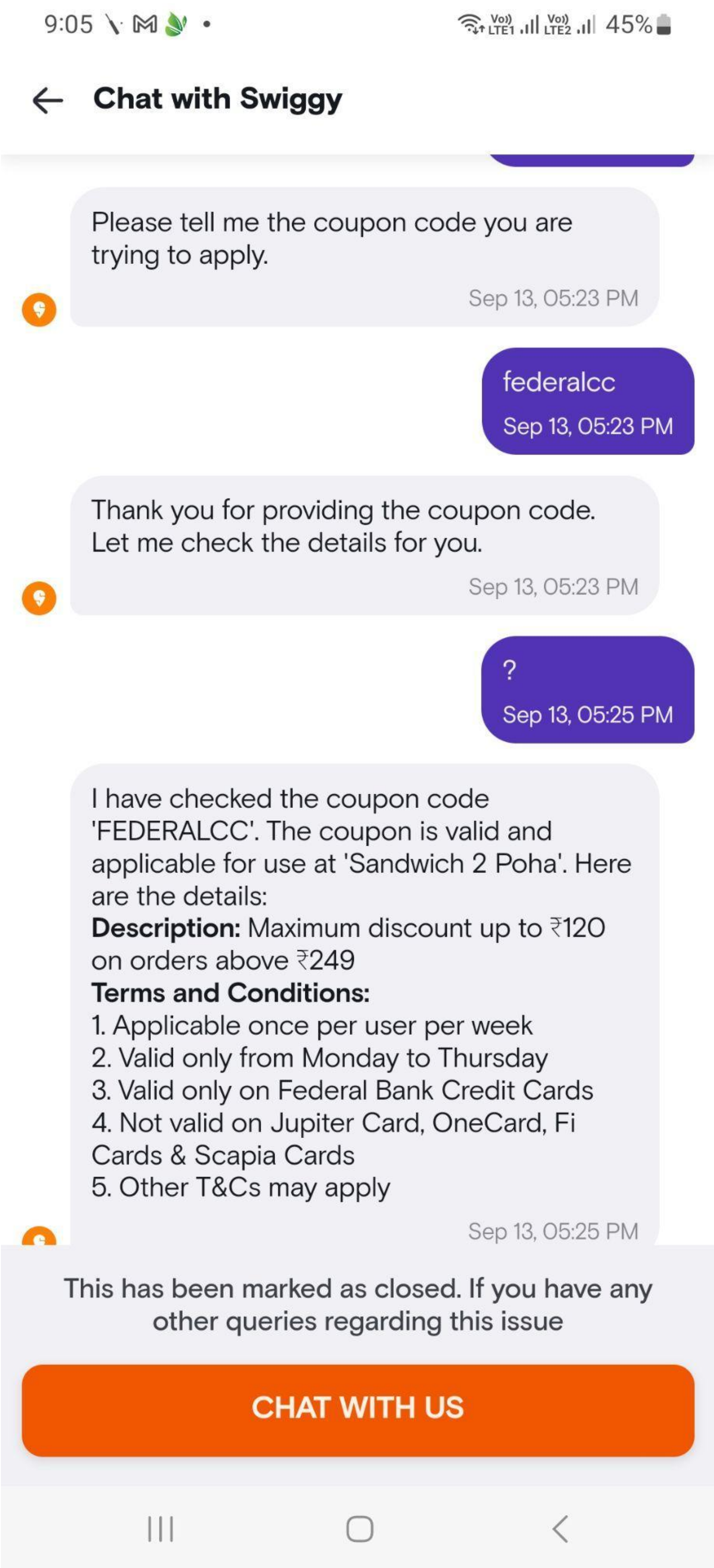
Automation track: LLM customer-service bot



Decision-tree based bot



Human agent conversation



LLM powered bot

Automation track: LLM customer-service bot

Key challenges

- Dynamic data
- Steerability/driving towards closure
- Jailbreaking/hallucinations are expensive
- Free form text related issues
    - Out of order input
    - Conversational memory
    - Context shift

Automation track: LLM customer-service bot
Key learnings

- Data intensive use cases
  - Latency + token costs
  - Remedies
    - Context compression using LLM
    - Vectorization/RAG (Retrieval Augmented Generation) pipeline
    - Streaming API for gathering LLM responses

- Jailbreaking
  - System message to limit the context
  - Moving system message to the end
  - Few shot prompting

- Deploying for customers
  - VPC with firewalls and DDoS protection
  - Tenant isolation

Some learnings so far

- It took us about 3-4 months of wandering to land on potentially high ROI items. Until then we ended up flailing quite a bit
  - Conserve your bandwidth-- limit how many inbounds you entertain (both internal and external)
  - There’s a delta in effort/quality between a cool hackathon project/ external demo and it’s production version. Set/ manage expectations

- For cases which are not real-time, hard to beat GPT out of the box
  - Cost <> ROI trade-off
  - Hallucination was a real problem; Lots of internal-user testing & guardrailing
  - Using GPT directly from OpenAI quickly led to governance difficulties. Moved to Azure AI
  - This also means fine-tuning is usually a last resort/ expensive choice

- We didn’t see a lot of pull for customer-facing conversational interfaces (‘chatbot for food ordering / table reservations’)

Some learnings so far

- Be dispassionate and replace if something better comes along
  - Replaced Image Blending with Img2Img despite having spent a decent amount of time on the former
- Data and metadata are paramount. Especially for conversational analytics use-cases, we had to source and codify a lot of context / knowledge sitting in Business/ Ops folks' heads
  - 'Copilot for X' use-cases are tempting but quickly run into data difficulties
- Be pragmatic and be patient
  - Not easy to find/ land slam-dunk use-cases
  - Relatively easier to find automation tasks (i.e., efficiency improvement/ cost reduction) vs. top-line ones
  - Human-in-the-loop/ Ops validation is almost always necessary

THANK YOU!

Don’t forget to:

Swiggy today!

Get your Swiggy One!