Generative AI

Case Interview for ADC

July 22nd 2024

Felix Schilling, PhD

This presentation

- Amazon Review Trends POC (20 min)
  - Goal of the Project
  - Project Plan
  - POC
- Retrieval-Augmented Generation (RAG) (10 min)

The Goal

Provide **useful insights** from customer feedback to stakeholders in Amazon's health and personal care sector.

Project plan

- Workshop w. Stakeholder [Done]
- POC Pipeline
  - Batch Process, API, Cleaning
  - Model Development & Deployment
  - Accessible Visualization and Reporting using dashboards
- Iterative improvements of POC based on stakeholder feedback loop.

The big Picture I:

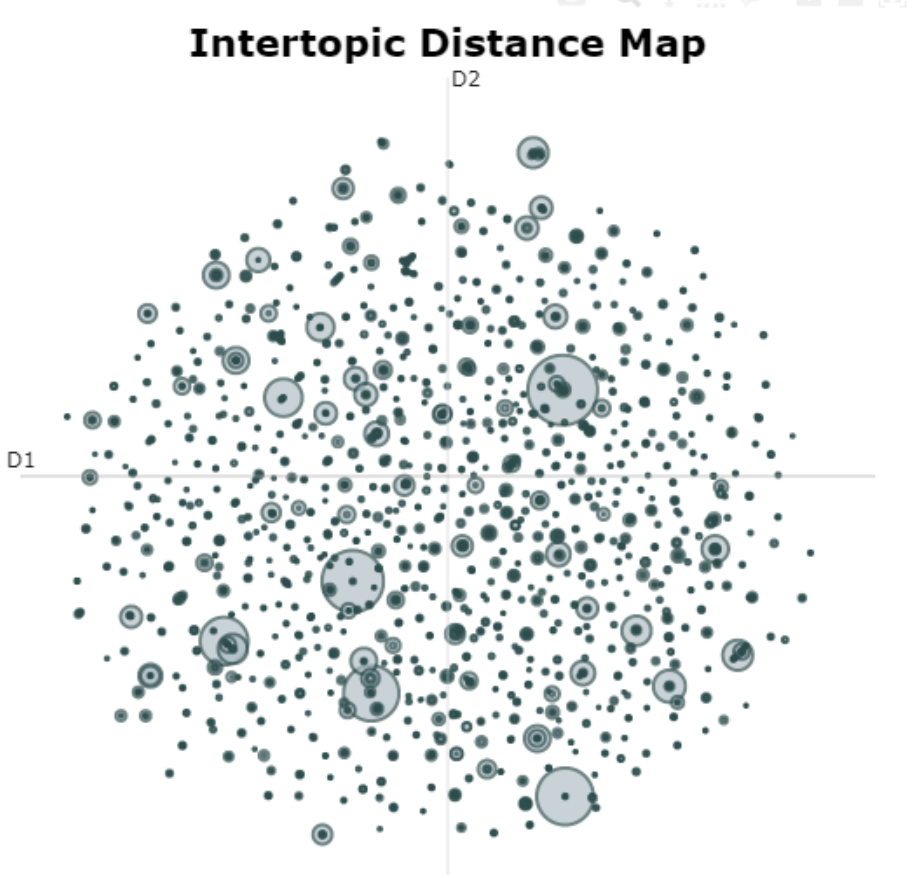
- Overall good quality of the reviews.
- Bots/Spam seems to be filtered out.
- **Assumption:** Reviews are "honest".
    - We can implement what they say to improve products.
    - We can test this assumption with the POC.

The big Picture II:

Topics & Trends are too complex to discuss them as a whole.

1. Text dimension (structured and unstructured): > 3000 Topics with BERT
2. Time dimension: >22 years
3. ASIN codes: >60k unique products

Let's focus on micro level cases for specific stakeholders.



Recap: Initial Stakeholder Consultation

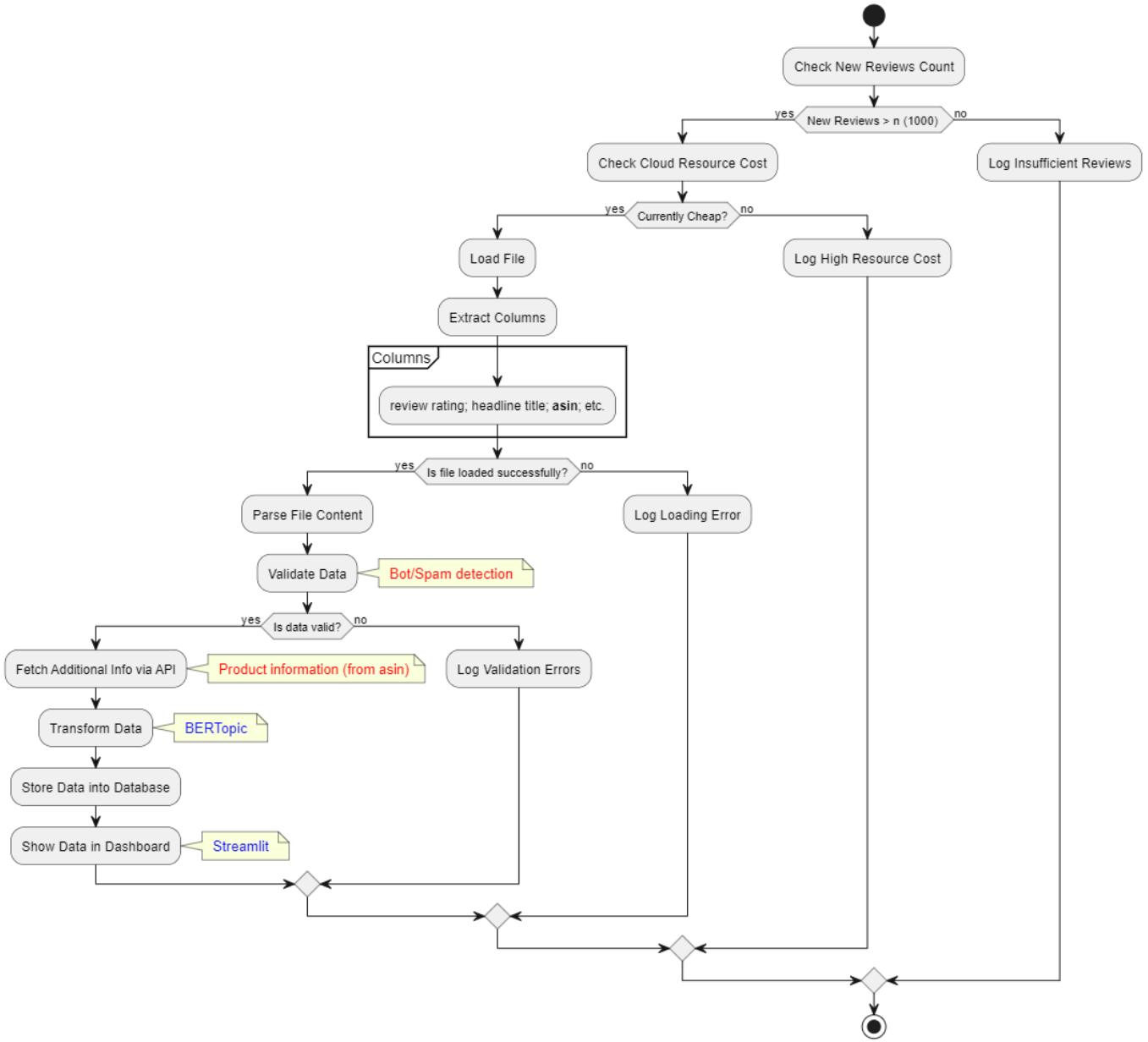
Let's define role, needs and metrics of our stakeholders.

Q: What's the challenge and metric defines the solution?

Stakeholder Roles Examples

1. Product Owner
    - Needs to know what people do and do not like about a specific product.
    - Interested in the product lifecycle since release.
    - Metrics:
        - Time between product release and reviews.
        - How the review sentiment changes over time.
2. Marketing Strategist
    - Interested in products that work via indirect consumer marketing.
    - Focuses on purchase decisions made within a household or close relationship.
    - Metric: Product reviews that mention family members.
    - Example: A spouse or partner buys your shower gel.

Pipeline (Batch Processing, API Call & NLP)



Model Development

BERTtopic.

- Find topics and trends with an unsupervised Algorithm.
- AI (ChatGPT) to label Topic Representation.

Link to ASIN and Time variables before hand.

- Select only the **relevant** reviews for each task.
- Increases accuracy and reduces runtime.

Dashboard

Accessible "micro insights" for each stakeholder.

- How do the metrics that matter for my task look like?
- How do my actions relate to this ?

Link to the app (Localhost!)

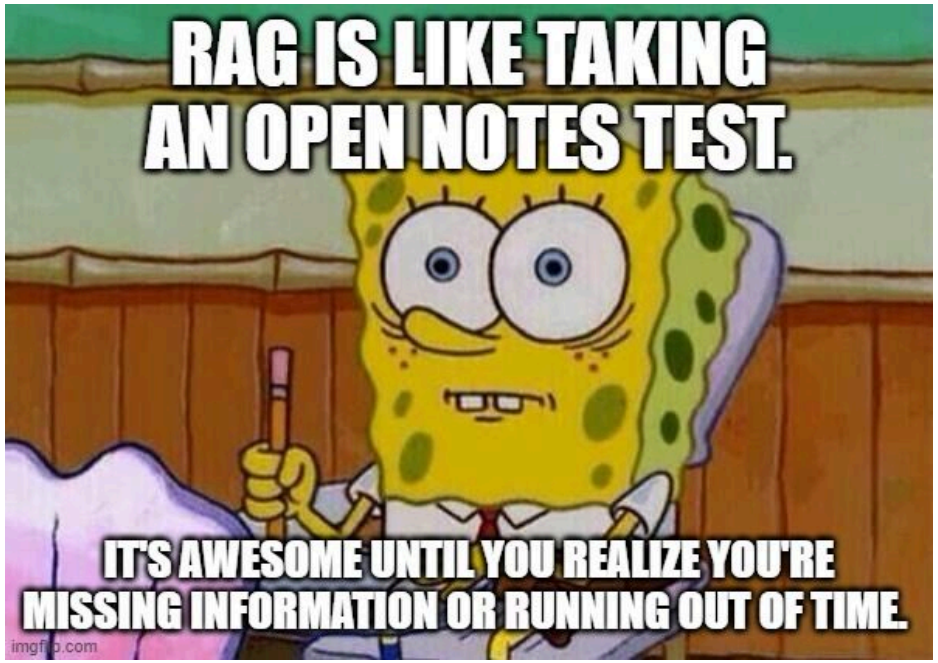
RAG: Retrieval-Augmented Generation

The Goal

- Feeding LLMs domain specific knowledge
- Better answers
- References
- Publishing date of reference /information

Main Challenge

1. Performance.
2. Getting the right data.
3. Knowing which information is (still) relevant.



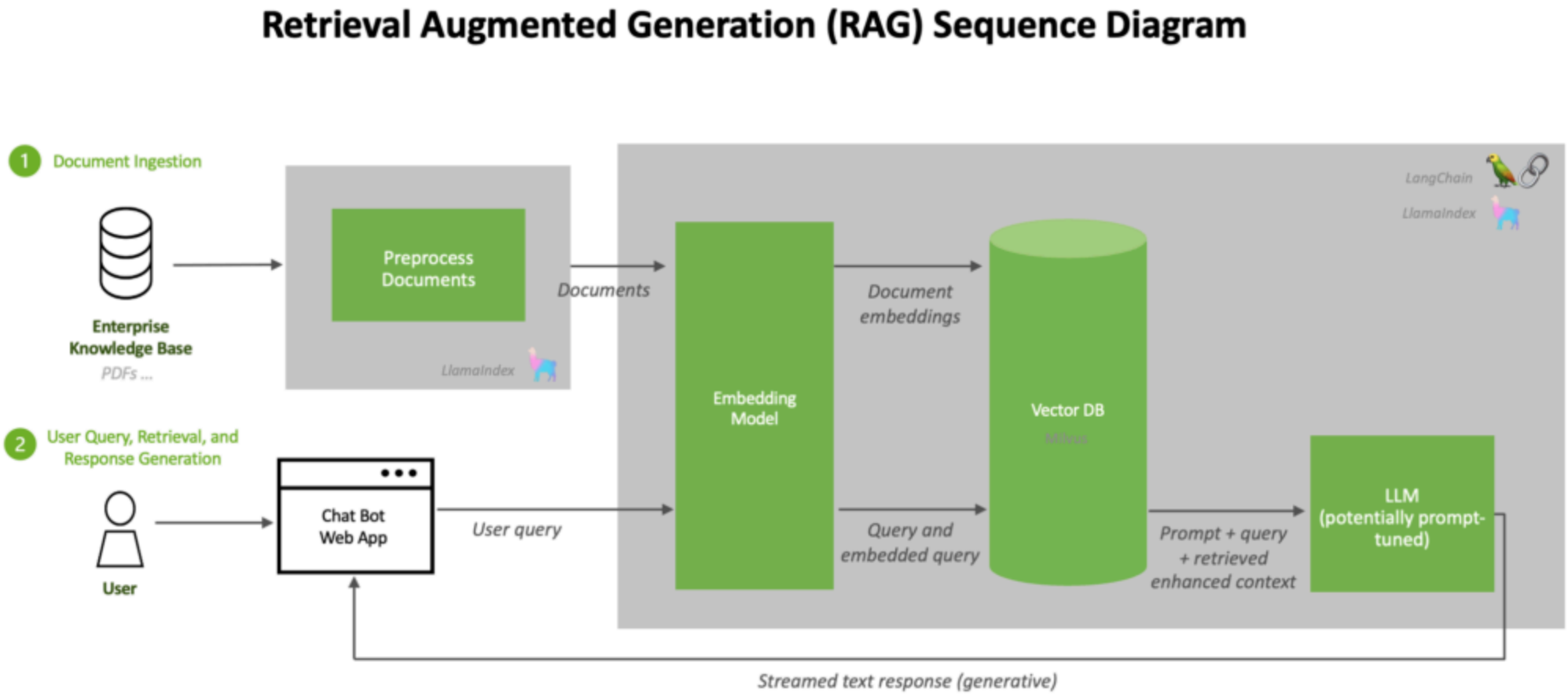
Hypothesis

- Best practices & existing tech stack to minimize the uptake in tech burden.
- Knowledge of domain specific experts.
- They understand the problem
- They know what data exists
- They know what changed over the last years.

Input quality of the data is crucial.

- (Semi-) structured data might be easier to junk.

RAG Pipeline



from Nvidida Blog

Choice of Generative Model

- **Model Selection**: Compare models from LMSYS Chatbot Arena Leaderboard for reference.
- **A/B Testing** to find the best performing model for specific use case.

Challenges, Considerations, and Evaluation Metrics

- **Deployment Options**: Local vs. Cloud
  - **Local**: Better data privacy but higher setup costs.
  - **Cloud**: Easier access to powerful models but involves ongoing costs.
- **Performance**: Evaluate speed and accuracy.
- **Contextual Relevance**: Ensure the model understands the specific use case.
- **Data Privacy**: Secure sensitive customer data.

Evaluation Metrics:

- Accuracy
- Response Time
- Customer Satisfaction (Experts!)

Ethical Considerations

- **Bias Mitigation**: Ensure the model is free from biases.
- **Transparency**: Inform users about AI-generated responses.
- **Data Privacy**: Protect user data in compliance with regulations.

Conclusion

- Consultants should :
  - Collaborate with domain experts
  - Close alignment with existing cloud infrastructure.
- Garbage in, garbage out (GIGO) applies to RAGs aswell.

Thanks!

- felix.s.schilling@gmail.com
- schillingerkurs.github.io