Seven Failure Points When Engineering a Retrieval Augmented Generation System

разбор статьи
https://arxiv.org/abs/2401.05856

Дмитрий Колодезев @promsoft kolodezev.ru
2024.02.20 @ DS TALKS

RAG Failure Points

Seven Failure Points When Engineering a Retrieval Augmented Generation System

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, Mohamed Abdelrazek
{scott.barnett,stefanus.kurniawan,srikanth.thudumu,zach.brannelly,mohamed.abdelrazek}@deakin.edu.au
Applied Artificial Intelligence Institute
Geelong, Australia

ABSTRACT

Software engineers are increasingly adding semantic search capabilities to applications using a strategy known as Retrieval Augmented Generation (RAG). A RAG system involves finding documents that semantically match a query and then passing the documents to a large language model (LLM) such as ChatGPT to extract the right answer using an LLM. RAG systems aim to: a) reduce the problem of hallucinated responses from LLMs, b) link sources/references to generated responses, and c) remove the need for annotating documents with meta-data. However, RAG systems suffer from limitations inherent to information retrieval systems and from reliance

build new HCI solutions, complete complex tasks, summarise documents, answer questions in a given artefact(s), and generate new content. However, LLMs suffer from limitations when it comes to up-to-date knowledge or domain-specific knowledge currently captured in company's repositories. Two options to address this problem are: a) Finetuning LLMs (continue training an LLM using domain specific artifacts) which requires managing or serving a fine-tuned LLM; or b) use Retrieval-Augmented Generation (RAG) Systems that rely on LLMs for generation of answers using existing (extensible) knowledge artifacts. Both options have pros and cons related to privacy/security of data, scalability, cost, skills required,

Университет Deakin

- 14-й в австралийских рейтингах
- 233 QS 255 THE
- 60 000 студентов на 29 миллионов населения
- A2I2 — не самая большая часть
- Приличный региональный вуз (Воронеж?)
- https://a2i2.deakin.edu.au/about-us/

A2I2 прикладные исследования

- ML в баллистике: предсказание бронепробиваемости
  - Machine learning for predicting the outcome of terminal ballistics events
  - Ballistic limit predictions of non-identical layered targets
- Causal Inference
  - Causal Inference via Style Transfer for OOD Generalisation
  - Causality-aided Recommendation Systems
- Безопасность ML- систем
  - Prescriptive analytics with differential privacy
  - Black-box Few-shot Knowledge Distillation
- https://a2i2.deakin.edu.au/publications/

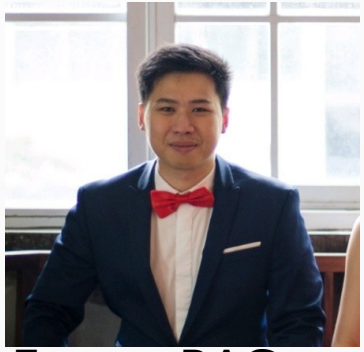
Авторы

- Scott Barnett
- Srikanth Thudumu
- Stefanus Kurniawan
- Zach Brannelly
- Mohamed Abdelrazek

Заказчик ?

Завлаб ?

Делал RAG



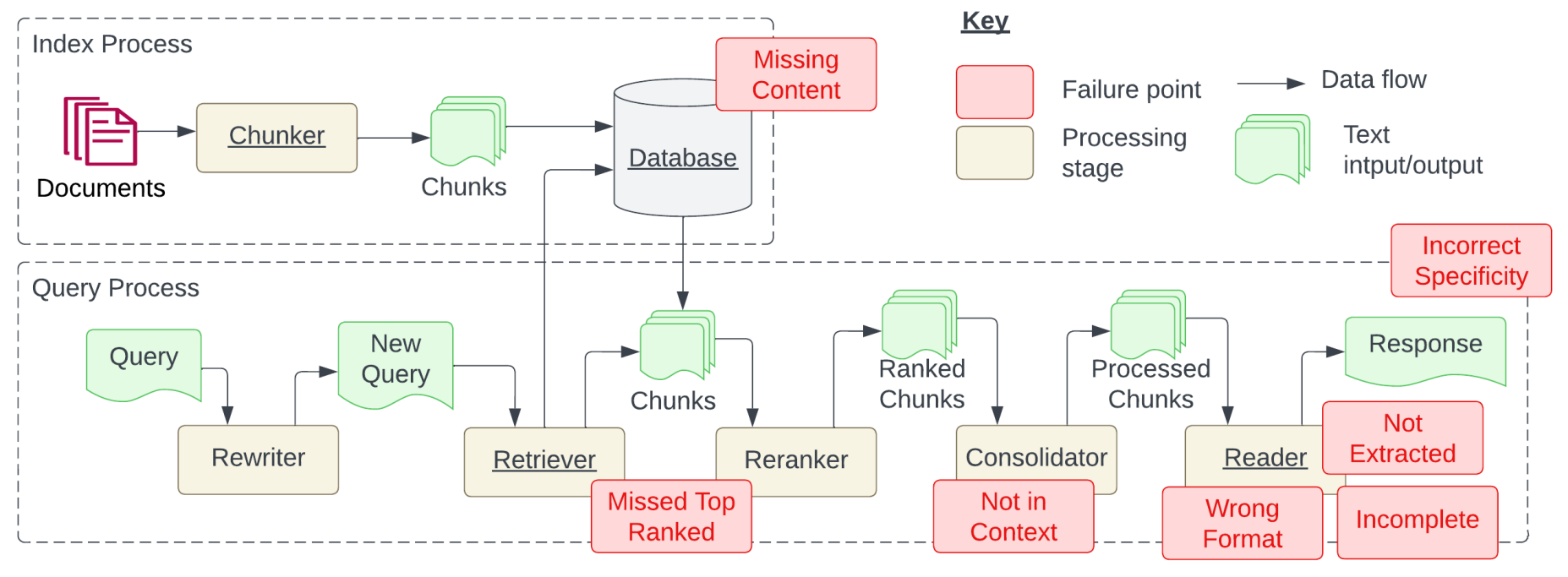
Делал RAG

Проф *

О чем это вообще

- Хотим, чтобы LLM отвечала на вопросы
- Хотим, чтобы знала доменную область
  - Доучивать дорого и долго
- Хотим дешево добавлять новые знания
  - Делаем поисковую систему по текстам
  - Добавляем кусочки текстов в промпт
  - Получаем Генерацию, Дополненную Поиском
    Retrieval Augmented Generation (RAG)

RAG в разрезе



Large Language Models for Information Retrieval: A Survey

Процесс построения RAG

- предобработка артефактов доменной области
- извлечение и нарезка текста на кусочки
- сохранение в хранилище,
  обычно с векторным индексом на эмбеддингах
- построение системы подбора нужных кусочков
- ранжирование и отбор найденного
- добавление в промпт LLM

Вопросы и выводы

- What are the failure points that occur when engineering a RAG system?
- What are the key considerations when engineering a RAG system?

- Validation of a RAG system is only feasible during operation.
- The robustness of a RAG system evolves rather than designed in at the start.

Какие системы строили

- **Cognitive Reviewer** is currently used by PhD students from Deakin University to support their literature reviews — в проде
- **The AI Tutor** is a RAG system where students ask questions about the unit and answers are sourced from the learning content — в проде
- **Biomedical Question and Answer** RAG system using the BioASQ dataset — для экспериментов

https://figshare.com/s/fbf7805b5f20d7f7e356

Небольшие датасеты

Case Study	Domain	Doc Types	Dataset Size	RAG Stages	Sample Questions
Cognitive Reviewer*	Research	PDFs	(Any size)	Chunker, Rewriter, Retriever, Reader	What are the key points covered in this paper?
AI Tutor*	Education	Videos, HTML, PDF	38	Chunker, Rewriter, Retriever, Reader	What were the topics covered in week 6?
BioASQ	Biomedical	Scientific PDFs	4017	Chunker, Retriever, Reader	Define pseudotumor cerebri. How is it treated?

Table 1: A summary of the RAG case studies presented in this paper. Case studies marked with a * are running systems currently in use.

FP1 Missing Content

- На некоторые вопросы нет ответа в наших документах.
- Хотелось бы, чтобы система сказала «не знаю»
- Если контент «где-то рядом»,
  система что-то придумает

Q: What is the aim of the Human Chromosome-centric Proteome Project (C-HPP)?
A: The given pieces of context do not provide information about the aim of the Human Chromosome-centric Proteome Project (C-HPP).
Expl: We don't have the original document where this question is answered from indexed in the system, that's why the system can't find it.

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

FP2 Missed the Top Ranked

- Ответ был в найденных «кусочках»,
  но мы взяли K самых релевантных,
  и он в них не попал.

Q: Which are the thyroid hormone analogs utilized in human studies?
A: The review mentions several thyroid hormone analogs that have been studied in humans, including T3, T4, and the thyronamines (T1AM and T0AM).
Ex: TRIAC and TETRAC are two different thyroid hormone analogs utilized in human studies
Expl: First chunk of the source document is ranked 15 in the retrieved chunks, but it didn't make it to context because we only use the top 4.

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

FP3 Not in Context

- Нужная информация была найдена, но была отброшена в процессе консолидации контекста. В результате модель её не видела
- Бывает, когда много «кусочков» найдено, и их надо как-то предобработать перед тем, как передавать в промпт

Q: Which are the main functions of the human HuR (ELAVL1) protein in fibroblasts?
A: I don't know the specific functions of the human HuR (ELAVL1) protein in fibroblasts.
Expl: We have 27 documents indexed in the system, but not a single chunk make it to the context.

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

FP4 Not Extracted

- Нужная информация была в датасете,
  но поиск ее не нашел
- Когда много шума или противоречий в данных

Q: List genes that have been found mutated in CMT1A (Charcot-Marie-Tooth disease type 1 A).
A: The gene that has been found mutated in CMT1A is PMP22 (Peripheral Myelin Protein 22).
Expl: MPZ is mentioned in the context but not extracted, only PMP22 is extracted

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

FP5 Wrong Format

- Модель ответила, но не в том формате, в котором мы хотели
- Например, просили список — вернули текст, который надо как-то разобрать

Q: List the two most important synaptic markers.
A: The two most important synaptic markers mentioned in the given context are PSD95 and synaptotagmin.
Expl: Expect list, but given a sentence.

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

см. Instructor

FP6 Incorrect Specificity

- Ответ получен, но либо слишком общий, либо, напротив, рассматривает только очень специфичный случай
- Когда человек не понимает, как спросить
- Когда помогаем студентам разобраться

Q: Is Sarcolipin a regulatory/inhibitory protein of the Calcium ATPase SERCA?
A: Yes, Sarcolipin is a regulatory protein of the Calcium ATPase SERCA. It binds to SERCA and interacts with it throughout the kinetic cycle.
Expl: Doesn't mentioned Ca(2+)-ATPase (Serca) pump, but instead a mixed combination of `ATPase Serca`.

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

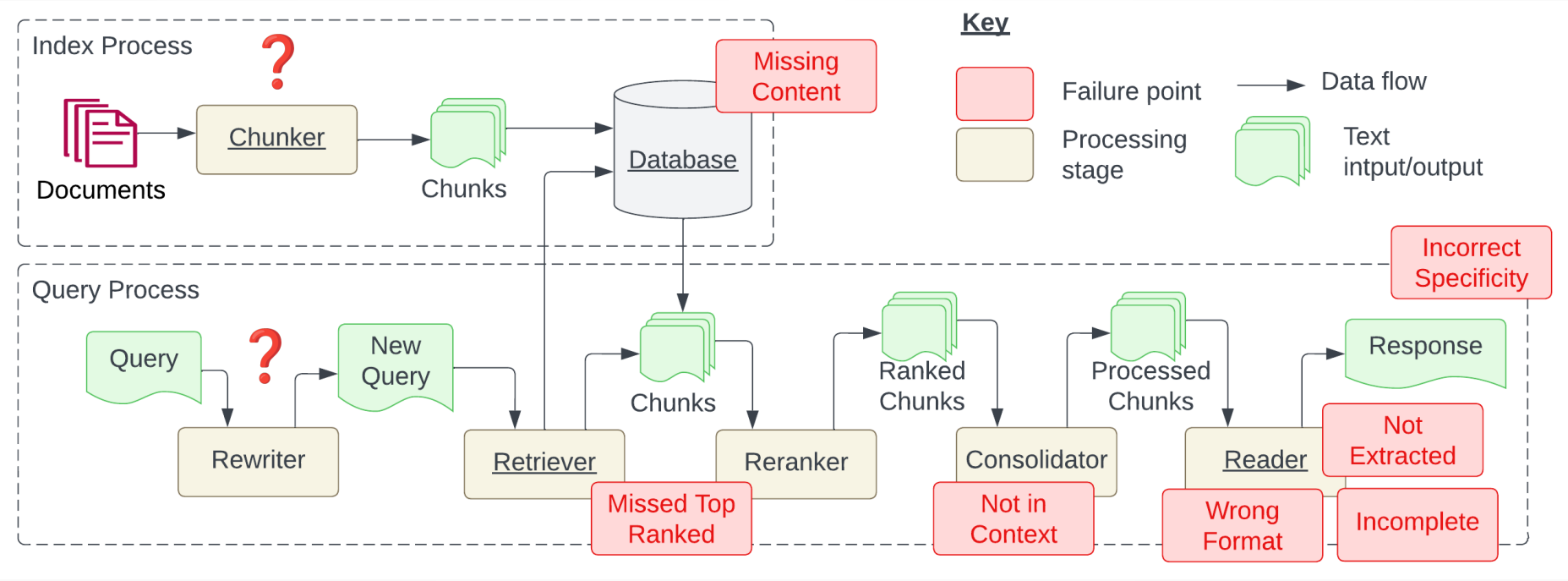
FP7 Incomplete answers

- Неполный ответ
- Частая причина — сложносочиненный вопрос.
- Будь проще — и LLM тебя поймет

Q: Which hormone abnormalities are common in Williams syndrome
A: The hormone abnormalities commonly found in Williams syndrome are dysregulation of oxytocin and vasopressin.
Expl: Incomplete because it is only saying 2 out of 8 from the expert answer and this could also be classified as "incorrect specifity" because it mentions "dysregulation" instead of "elevated".

https://figshare.com/s/fbf7805b5f20d7f7e356?file=43203591

Болевые точки на карте



+ нарезка через смысловые границы, «колдунщики»

И вот что мы поняли

FP	Lesson	Description	Case Studies
FP4	Larger context get better results (Context refers to a particular setting or situation in which the content occurs)	A larger context enabled more accurate responses (8K vs 4K). Contrary to prior work with GPT-3.5 [13]	AI Tutor
FP1	Semantic caching drives cost and latency down	RAG systems struggle with concurrent users due to rate limits and the cost of LLMs. Prepopulate the semantic cache with frequently asked questions [1].	AI Tutor
FP5-7	Jailbreaks bypass the RAG system and hit the safety training.	Research suggests fine-tuning LLMs reverses safety training [11], test all fine-tuned LLMs for RAG system.	AI Tutor
FP2, FP4	Adding meta-data improves retrieval.	Adding the file name and chunk number into the retrieved context helped the reader extract the required information. Useful for chat dialogue.	AI Tutor
FP2, FP4-7	Open source embedding models perform better for small text.	Opensource sentence embedding models performed as well as closed source alternatives on small text.	BioASQ, AI Tutor
FP2-7	RAG systems require continuous calibration.	RAG systems receive unknown input at runtime requiring constant monitoring.	AI Tutor, BioASQ
FP1, FP2	Implement a RAG pipeline for configuration.	A RAG system requires calibrating chunk size, embedding strategy, chunking strategy, retrieval strategy, consolidation strategy, context size, and prompts.	Cognitive Reviewer, AI Tutor, BioASQ
FP2, FP4	RAG pipelines created by assembling bespoke solutions are suboptima.	End-to-end training enhances domain adaptation in RAG systems [18].	BioASQ, AI Tutor
FP2-7	Testing performance characteristics are only possible at runtime.	Offline evaluation techniques such as G-Evals [14] look promising but are premised on having access to labelled question and answer pairs.	Cognitive Reviewer, AI Tutor

Table 2: The lessons learned from the three case studies with key takeaways for future RAG implementations

Chunking and Embeddings

- Нарезка на кусочки сложнее, чем кажется.
  - эвристики (параграфы, пунктуация)
  - семантика самого текста
  - очень не хватает способов измерить качество нарезки
- Эмбеддинги для поиска
  - качество подбора зависит от доменной области
- Предобработка запросов
  - сильно улучшает качество RAG,
    особенно для негативных и общих вопросов

см. Corrective Retrieval Augmented Generation

RAG vs Finetuning

- RAG удобнее и вроде бы лучше, но неустойчиво
- Базовые модели становятся лучше
- По мере развития моделей нужно пересматривать преимущества RAG и дообучения, особенно:
  - Точность
  - Задержку
  - Эксплуатационную стоимость
  - Устойчивость

Testing and Monitoring

- Для RAG еще не сформировались хорошие инженерные практики
- Тестирование / генерация тестовых сценариев затруднены (непонятно, где брать)
- Нужно адаптировать практики из других областей
- Нужно учиться мониторить такие системы

См. huggingface cookbook

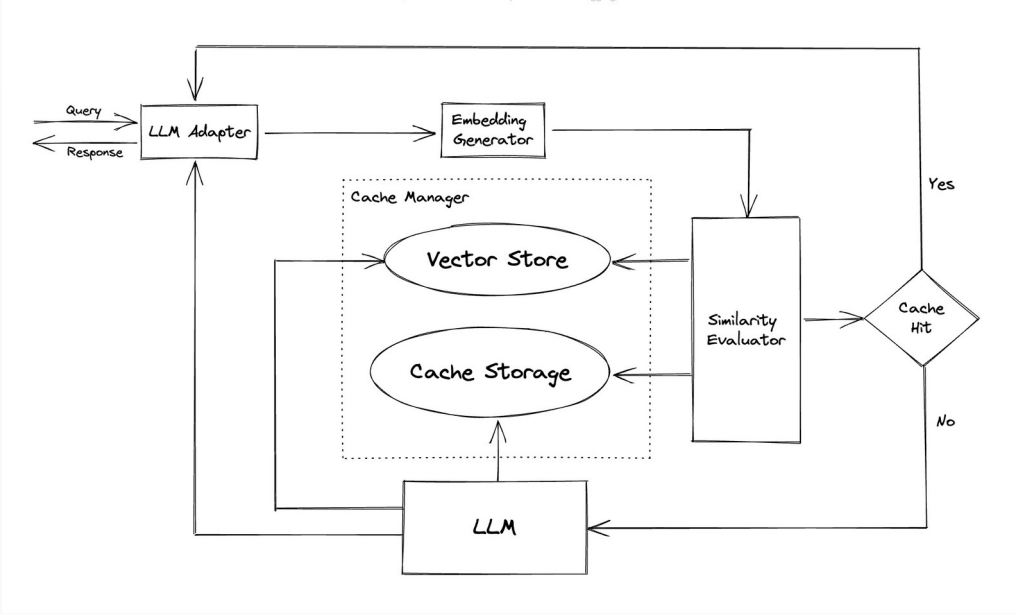
Интересное: OpenAI Evals

- https://github.com/openai/evals

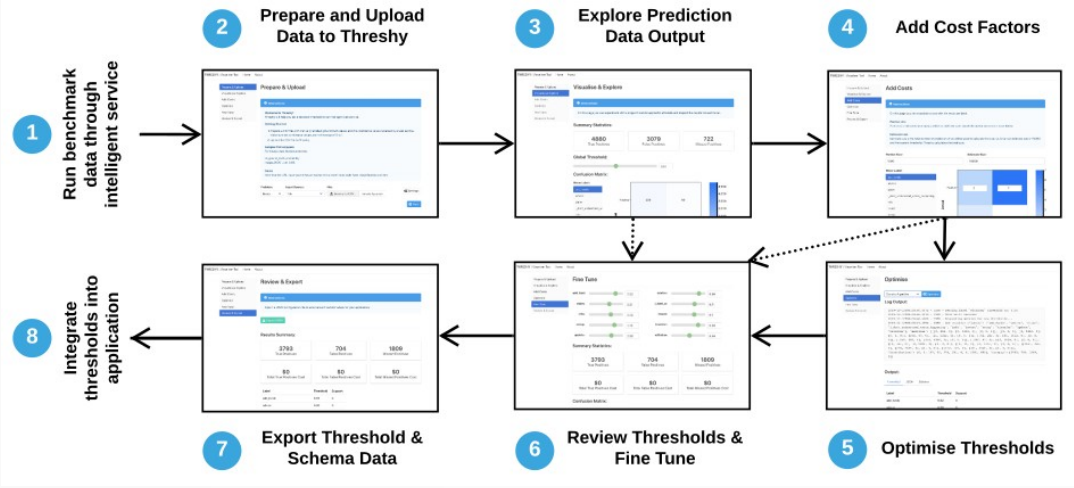
  {"**input**": [
  {"**role**": "**system**", "**content**": "Please note: In the following EXERCISE, it is important that you only respond with a single line in the format (x, y). Imagine you are standing in a 2D coordinate grid at (0, 0) where coordinates are represented like (x, y). You are currently facing the positive y direction."},
  {"**role**": "**user**", "**content**": "EXERCISE: If you take 5 steps forward, then turn 90 degrees left, then take 2 steps forward, then turn 90 degrees left, then take 1 step backward, then turn 90 degrees left, then take two steps backward, what coordinate are you at?"}
  ], "**ideal**": "(-4, 6)"}

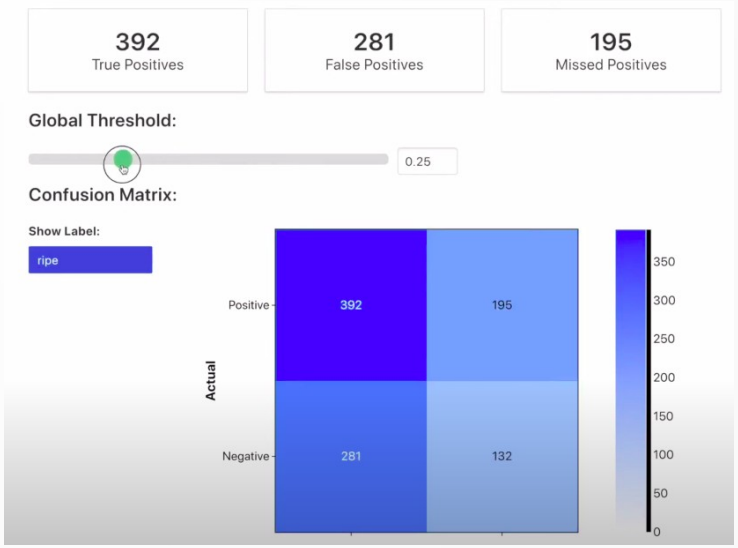
Интересное: GPT Cache

- https://openreview.net/pdf?id=ivwM8NwM4Z
- https://github.com/zilliztech/GPTCache



Threshy — подбор порога ;-)





Applied Artificial Intelligence Institute (A2I2)

13 followers · Australia · https://a2i2.ai

https://github.com/a2i2/threshy

Воронежский государственный университет

3 followers · Russian Federation · https://vsuet.ru · uit@vsuet.ru