TRINITY

Companion – User feedback analysis

Refinement plan to incorporate user inputs and other Advance features

February 2 , 2024

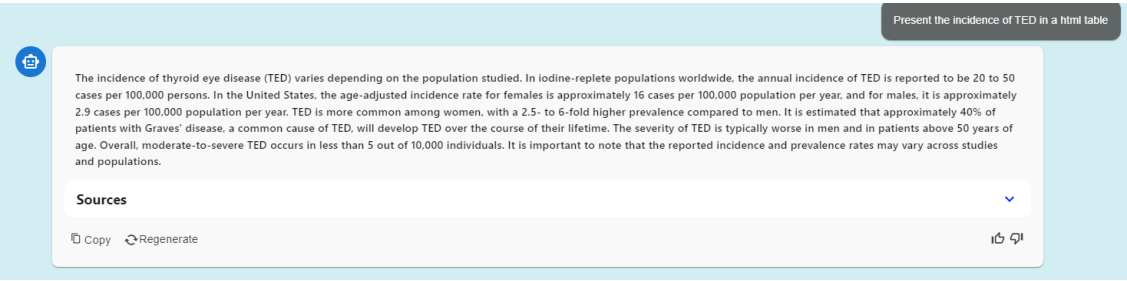
Agenda Slide

User feedback consolidation

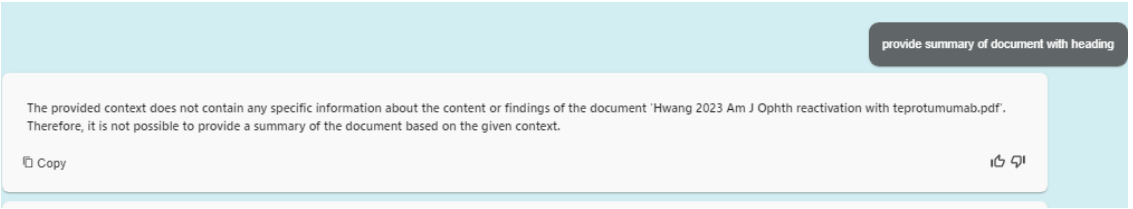
- User Feedback Summary and other Advance features
- Current system Architecture
- Application Refinement Plan
- Effort estimation

User Feedback Snippets

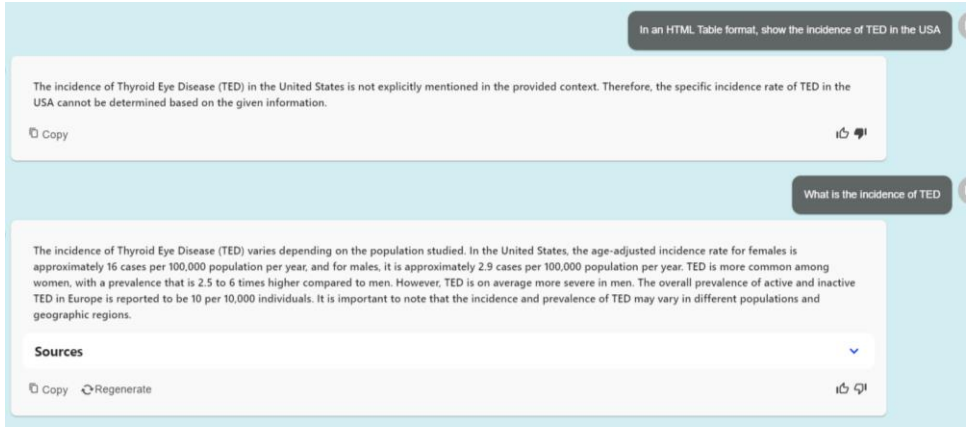
Two instruction in one prompt



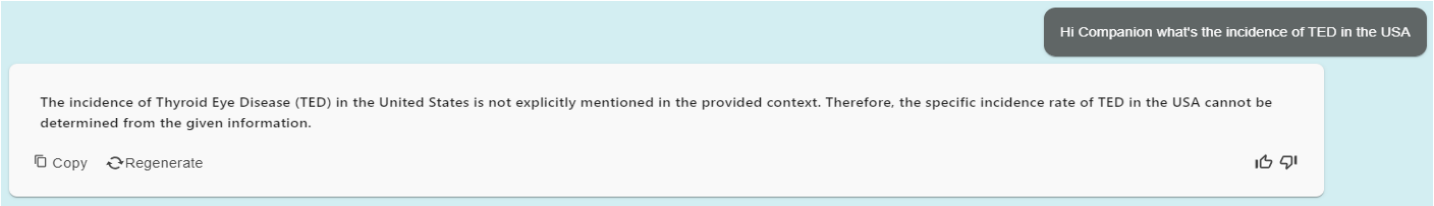
Summary in query prompt



HTML format + Incidence rate of TED in USA vs Incidence rate of TED



Inconsistency in Quality of Response



Agenda Slide

- User feedback consolidation
- **User Feedback Summary and other Advance features**
- Current system Architecture
- Application Refinement Plan
- Effort estimation

User feedback summary and Advance feature's theme

Feedback Digest

Summarizing the uploaded doc using query prompt

Summarizing a document from knowledge base using query prompt

Providing Multiple instructions in single prompt for e.g., what is the prevalence of TED? Respond in HTML table format

Quality of response (Failing to identify elements of a query, and inconsistency in response)

Other Advance Features

Reading Tables and Images

Response in predefined format/template (e.g., SRL format)

Word document scenarios – PPT creation, Upload PPT, Improvising document

Footnote Citations in the response

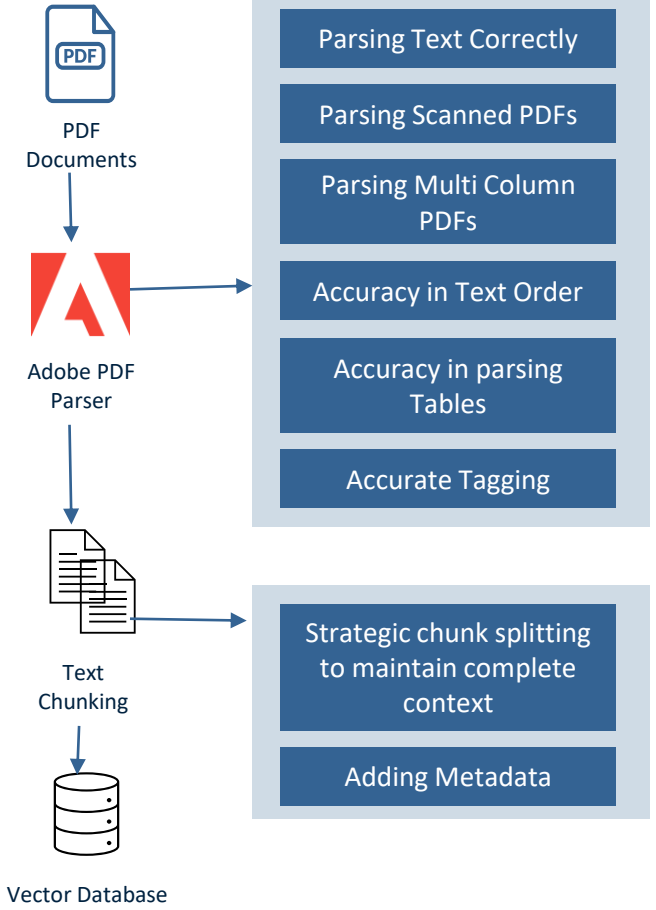
Comparing documents (e.g., Knowledgebase vs uploaded documents/2 uploaded docs)

Agenda Slide

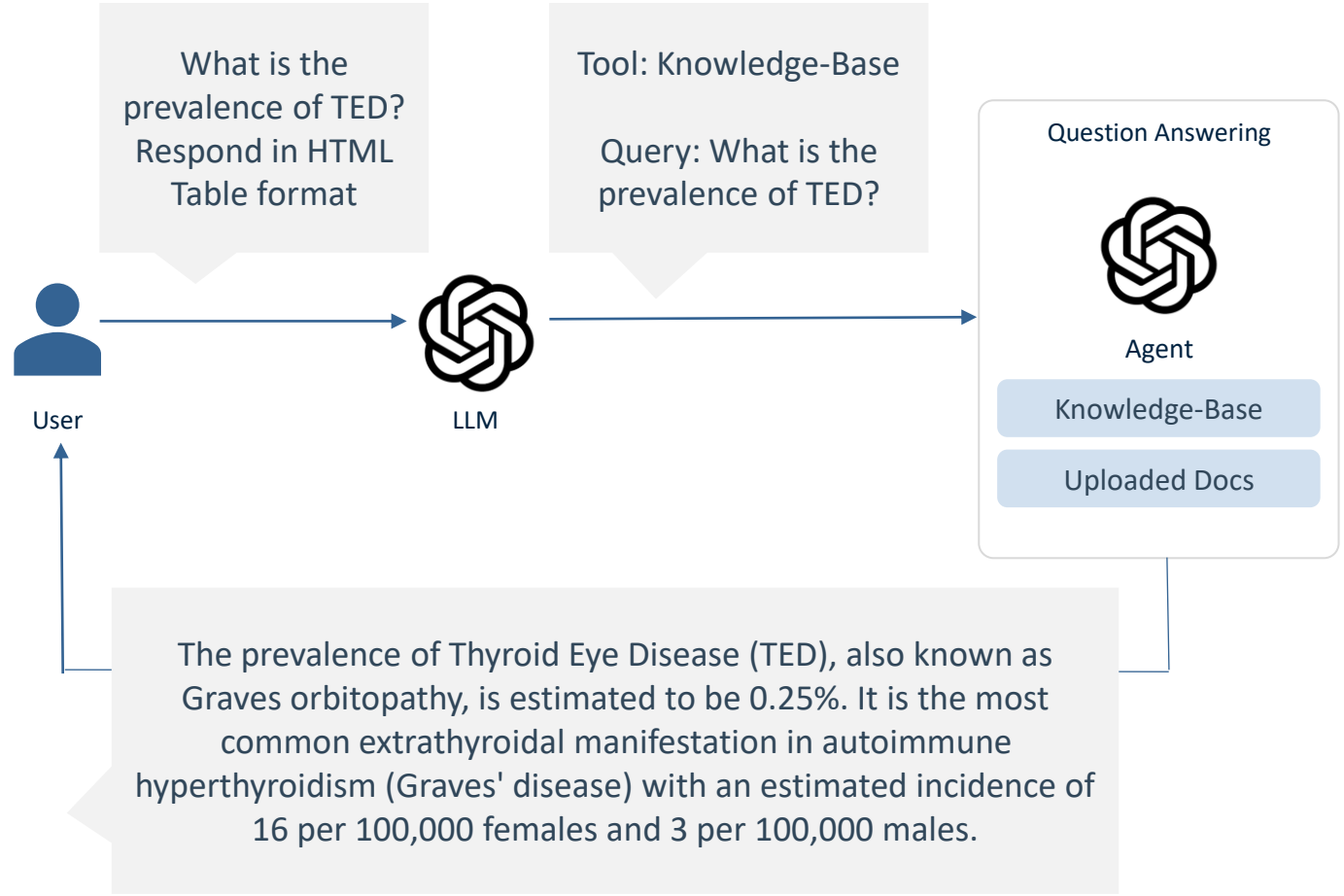
- User feedback consolidation
- User Feedback Summary and other Advance features
- **Current system Architecture**
- Application Refinement Plan
- Effort estimation

Current Architecture

Parsing and Chunking Component



Generative AI Architecture



Agenda Slide

- User feedback consolidation
- User Feedback Summary and other Advance features
- Current system Architecture
- **Application Refinement Plan**
- Effort estimation

Refinements plan

Based on User feedback analysis and Observations

Mandatory Major Refinements

Robust PDF Parsing Capability

- Enhance PDF parsing capabilities to handle tables, text, text order, images, and key points in complex documents. This should be on par with the Adobe PDF Parser.
- **Rationale:** Accurate storage of PDF information is a vital step in generating a desirable response to user queries.

LLM Architectural Modifications

- Replacement of langchain with a custom framework
- **Rationale:** Langchain lacks flexibility for advanced functionalities implementation such as table reading, creative responses, source footnotes, and customized template responses. Additionally, it struggles with handling multiple queries simultaneously.
- Adoption of a Manager-Worker Nodes Model with a Multi-Agent Framework – Like Boosting LLM framework
  - Implementation of the ReACT (Reasoning and Action) Framework as a prompting framework
  - **Rationale:** Enables processing multiple queries in a single prompt, breaks down tasks into sub-tasks, facilitates the implementation of a chain of thoughts, and enhances the precision and relevance of responses.

Build an Interface for GenAI Traceability – An alternative to Langsmith

- Capability to monitor all user activities, including GenAI ops, Debugging, Evaluating, Monitoring, Usage metrics.
- **Rationale**: Aiding in active feedback review and analysis to enhance application performance as Langsmith can't be used in current setup due to security reasons

Additional Optional Changes

Integration of GPT-4 Turbo

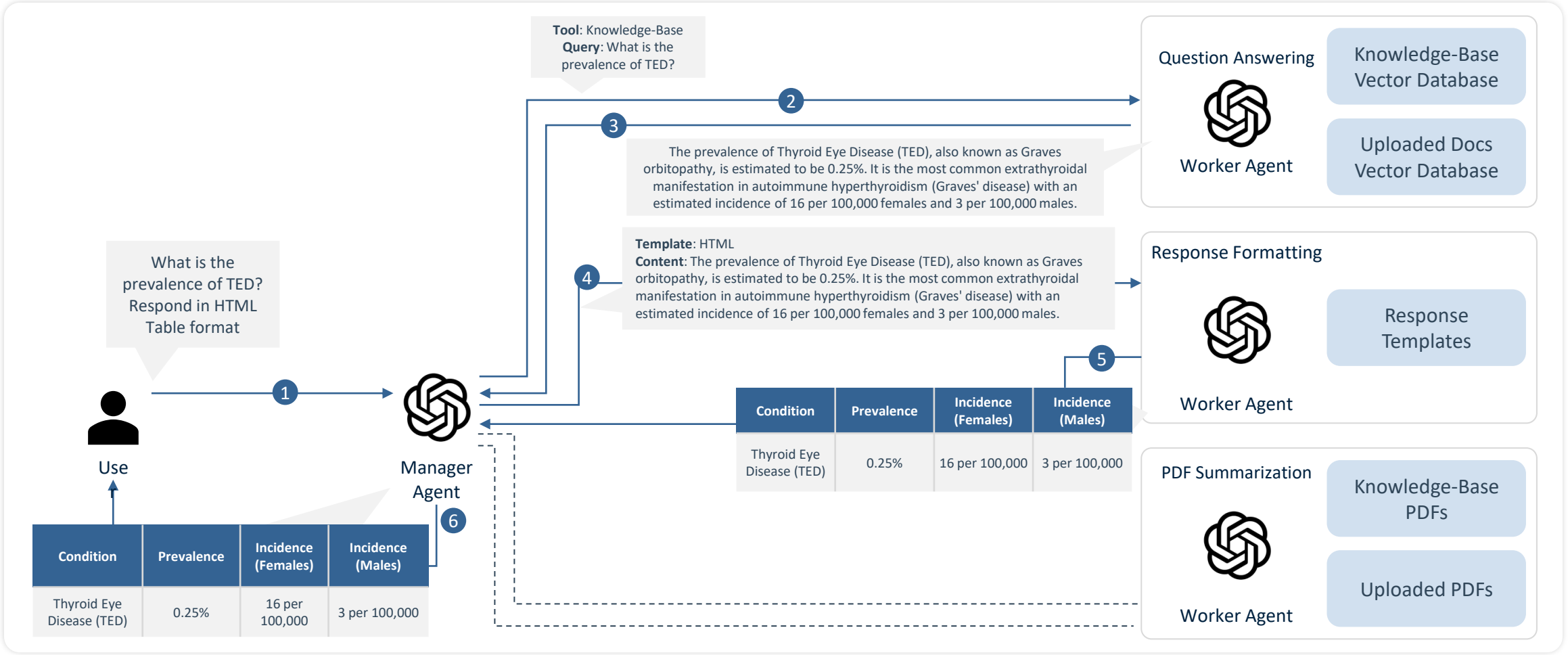
- **Rationale :** A superior model with enhanced reasoning and improved query/answer abilities compared to GPT-3.5. Additionally, GPT-4 Turbo is more cost-effective than GPT 4.

Introduction of a Reranker during response generation

- Implements chunk reordering based on relevance to queries
- **Rationale:** Improve the quality of response, more precise and desirable.

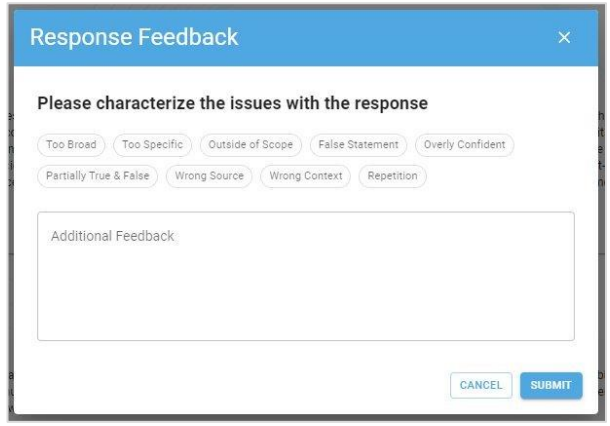
Note: The drawback of implementing this is the potential for increased latency and possible expenses due to LLM calls, depending on the complexity of the queries

Proposed System architecture

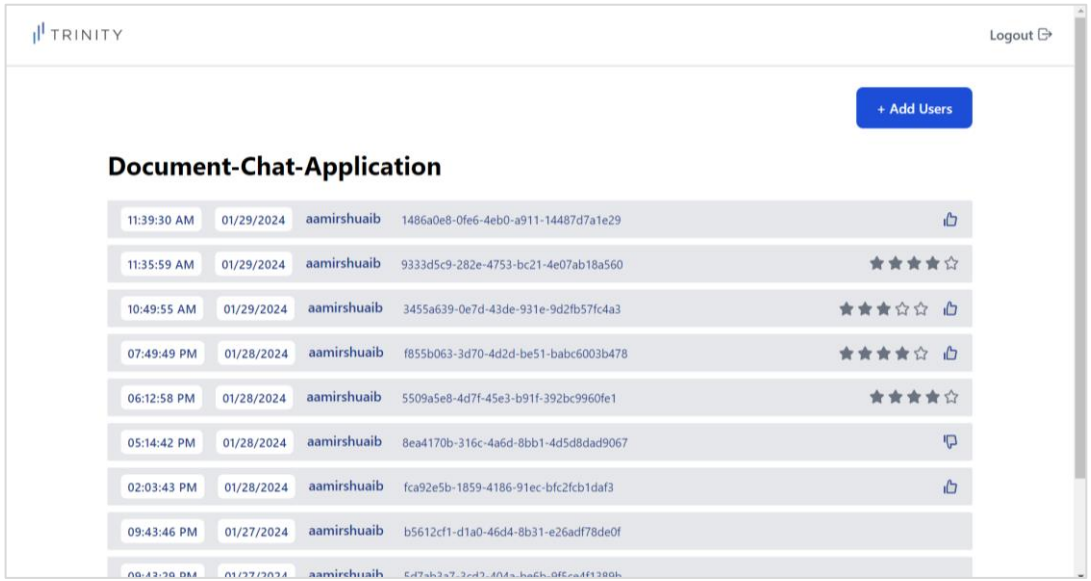


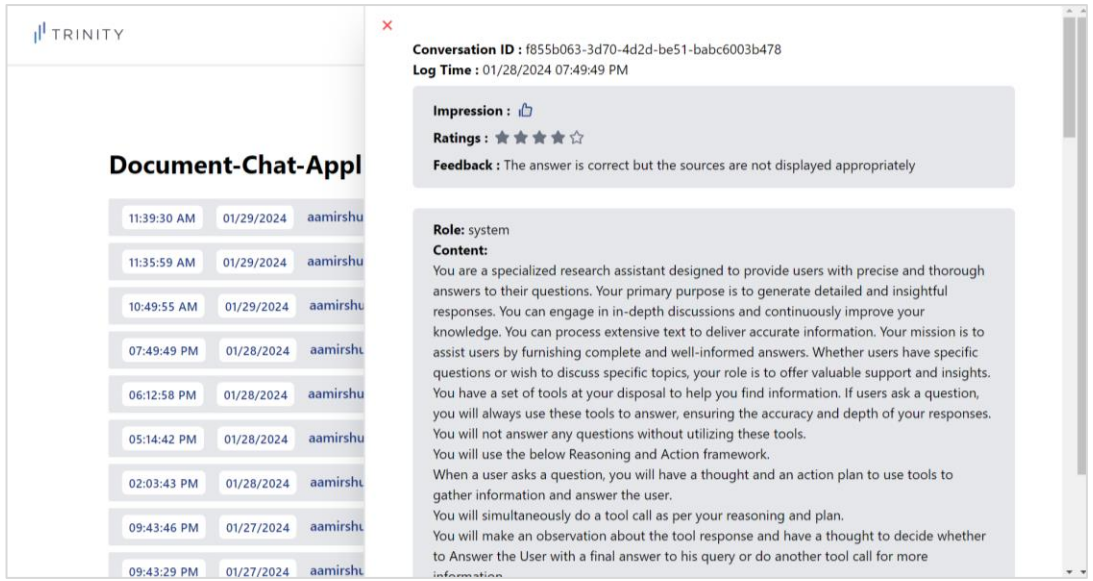
Note: The drawback of implementing this is the potential for increased latency and possible expenses due to LLM calls, depending on the complexity of the queries

User Feedback Tracking

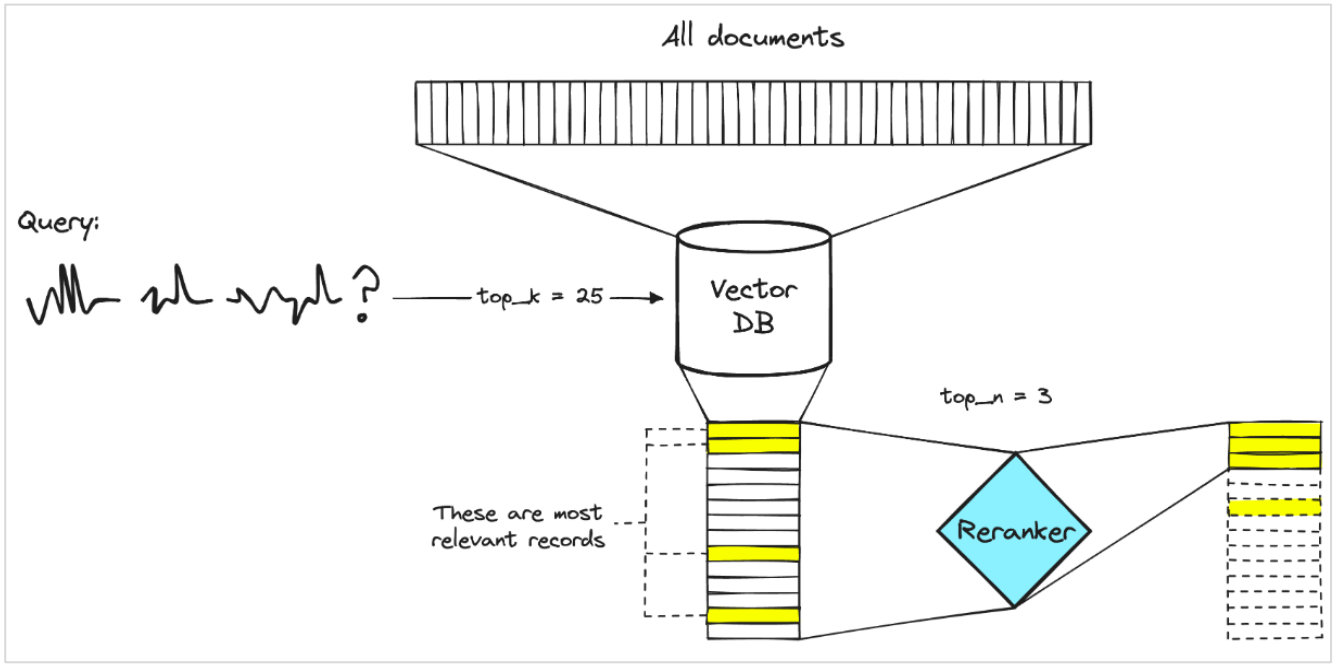


- Embedded Feedback collection component within the Companion Application
- Introducing a new application for continuous tracking.
- The Application tracks Gen-AI Operations/Flow and User Feedback
- This will help in improved assessment of the flagged/disliked responses





Re-Ranking



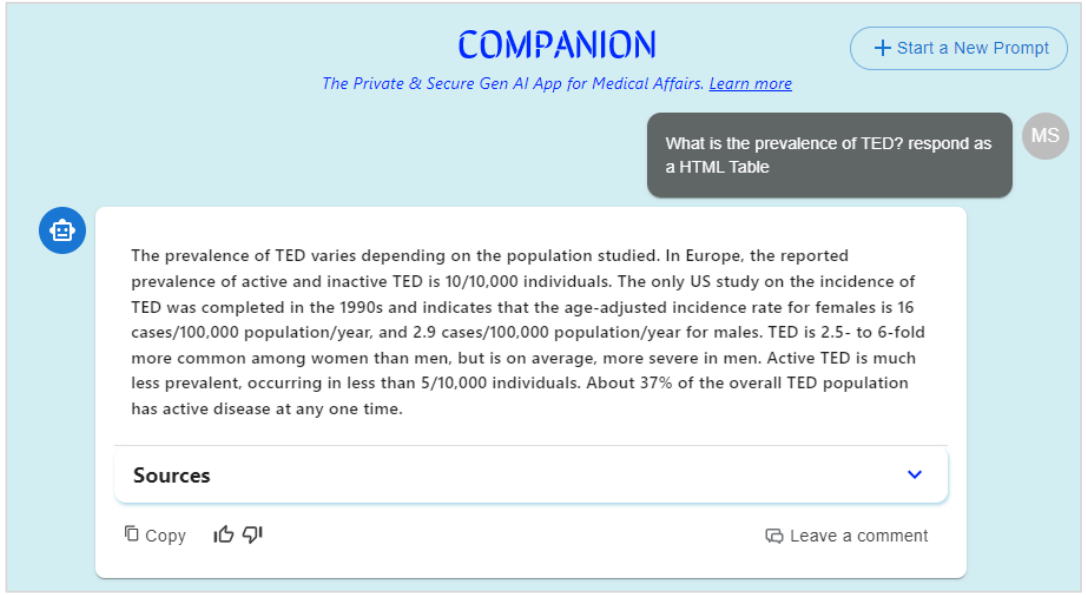
Re-ranking in Retrieval-Augmented Generation (RAG) applications is essential due to its ability to improve the quality of generated responses

Advantages

- Improved Answer Quality: The re-ranker can refine the initial retrieval set by reordering the documents based on their relevance to the query. This helps in generating better responses.
- Contextual Understanding: Re-ranking considers the context of the question and the document, leading to more accurate and relevant responses

Illustration – Companion response quality Comparison w.r.t Architecture

Current Architecture



Revised Architecture

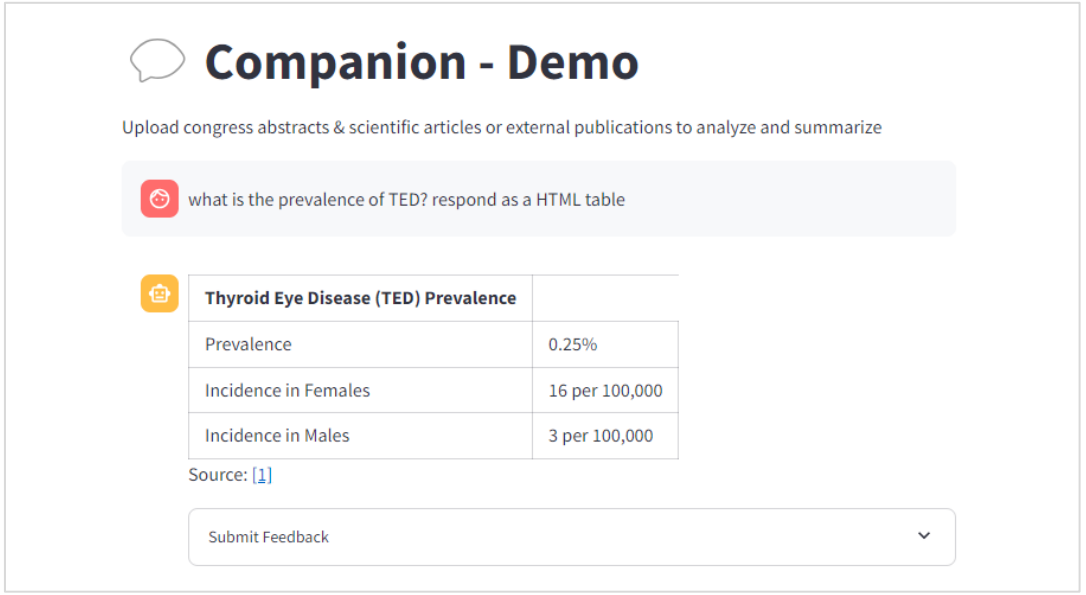
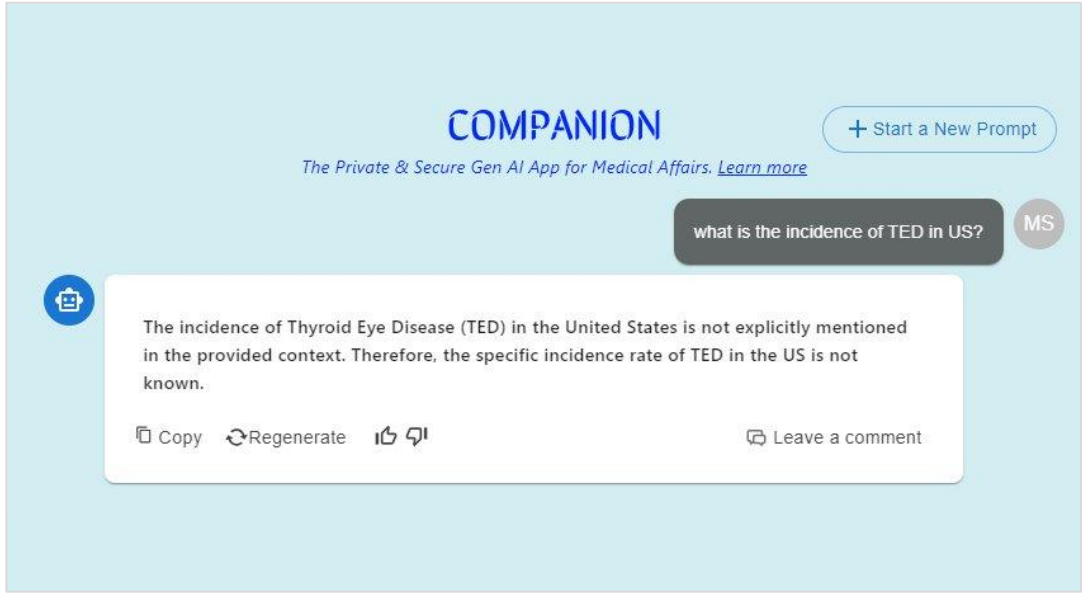
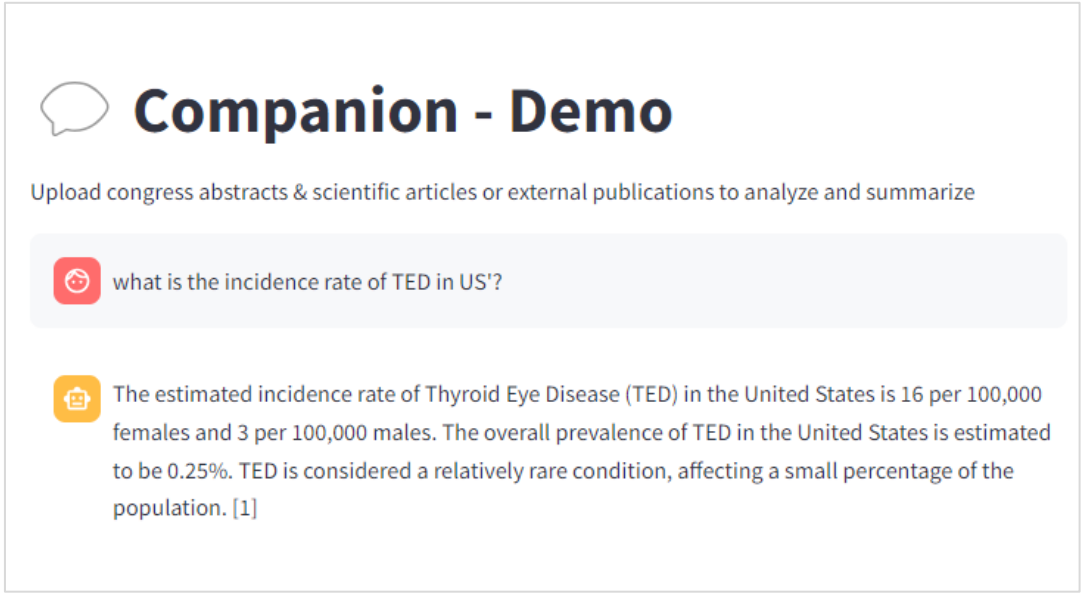


Illustration – Companion response quality Comparison w.r.t Architecture

Current Architecture



Revised Architecture



Examples – Companion response quality Comparison w.r.t PDF parser

TABLE 1. Treatments for thyroid eye disease

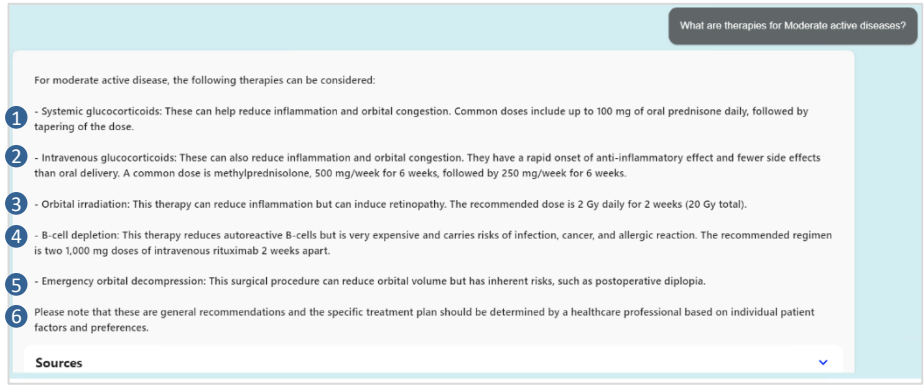
Therapy	Mode of Action	Pros and Cons	Common Doses
Mild active disease			
Topical solutions			
Artificial tears	Maintain tear film	Rapid action and minimal side effects	
Glucocorticoids	Reduce inflammation	Rapid action and minimal side effects	
Avoidance of wind, light, dust, and smoke	Reduces ocular surface desiccation and reduces irritation		
Elevation of head during sleep	Reduces orbital congestion		
Avoidance of eye cosmetics	Reduces irritation	Benefits not yet confirmed	
Selenium	Uncertain	Benefits not yet confirmed	
Moderate or severe active disease			
1 Systemic glucocorticoids			
2 Oral	Reduce inflammation and orbital congestion	Hyperglycemia, hypertension, and osteoporosis	Up to 100 mg of oral prednisone daily, followed by tapering of the dose
3 Intravenous	Reduce inflammation and orbital congestion	Rapid onset of anti-inflammatory effect, fewer side effects than oral delivery, and liver damage on rare occasions	Methylprednisolone, 500 mg/week for 6 weeks, followed by 250 mg/week for 6 weeks
4 Orbital irradiation	Reduces inflammation	Can induce retinopathy	2 Gy daily for 2 weeks (20 Gy total)
5 B-cell depletion*	Reduces autoreactive B-cells	Very expensive; risks of infection, cancer, and allergic reaction	Two 1,000 mg doses of intravenous rituximab 2 weeks apart
6 Emergency orbital decompression†	Reduces orbital volume	Surgical procedure with inherent risks, such as postoperative diplopia	
Stable disease (inactive)			
Orbital decompression (fat removal)	Reduces orbital volume	Postoperative diplopia and pain	
Bony decompression of the lateral and medial walls	Reduces proptosis by enlarging orbital space	Postoperative diplopia, pain, sinus bleeding, and cerebrospinal fluid leak	
7 Strabismus repair	Improves eye alignment and reduces diplopia	Surgical procedure with inherent risks	
8 Eyelid repair	Improves appearance, reduces lagophthalmos, and improves function	Surgical procedure with inherent risks, including orbital hemorrhage	

TED moderate active diseases

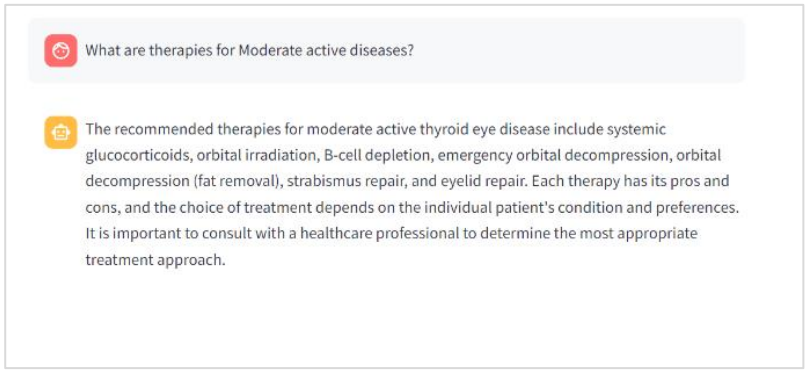
TED stable diseases

Question : What are the therapies for TED moderate active diseases

Adobe Parser



PymuPDF



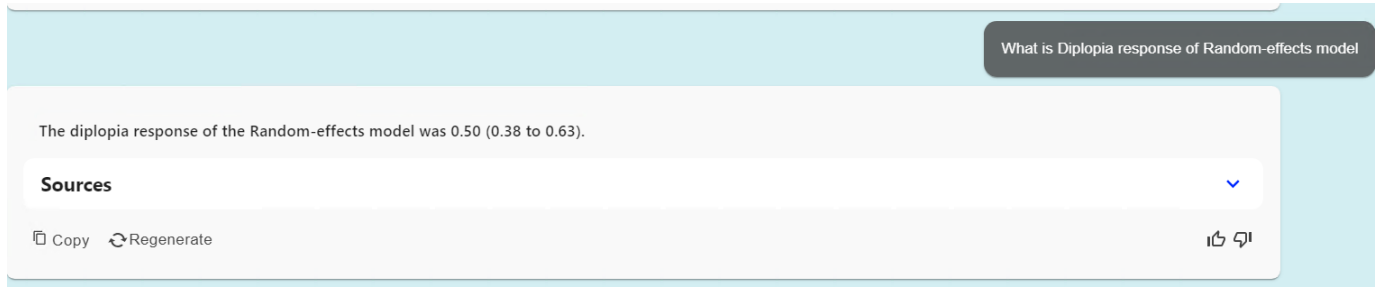
- **Wrong Information** in response & **missed therapy details**.
- For Moderate active disease, only first 6 points (marked in blue) should come while last 2 therapies are also included in response.
- It shows that parser missed some text & parsed with wrong order for some rows.

Examples – Companion response quality Comparison w.r.t PDF parser

B Diplopia response

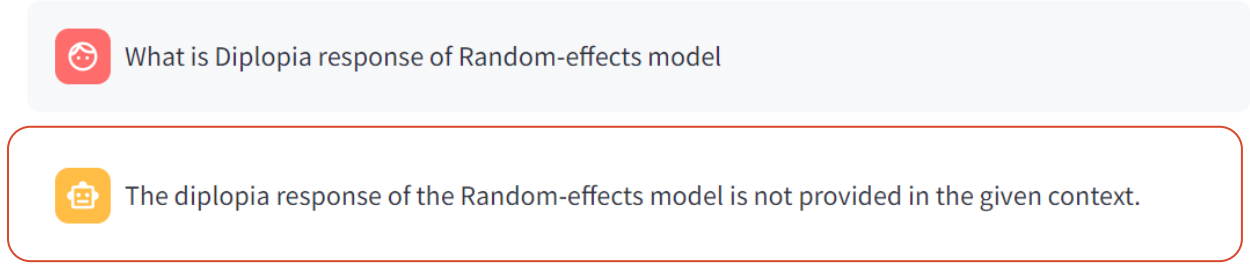
Source	No. with diplopia response	Sample size	Cfb_{raw} (95% CI)
He et al,[15] 2017	4	12	0.33 (0.10 to 0.65)
Kahaly et al,[16] 2005	16	26	0.62 (0.41 to 0.80)
Zhu et al,[18] 2014	17	28	0.61 (0.41 to 0.78)
Beleslin et al,[19] 2020	25	59	0.42 (0.30 to 0.56)
Fixed-effects model		125	0.50 (0.41 to 0.58)
Random-effects model			0.50 (0.38 to 0.63)
Heterogeneity: $I^2 = 43\%$, $r^2 = 0.1132$, $P = .15$			

Adobe Parser



PymuPDF

- Parser failed to extract information.



Agenda Slide

- User feedback consolidation
- User Feedback Summary and other Advance features
- Current system Architecture
- Application Refinement Plan
- **Effort estimation**

Preliminary Effort Projection

	Refinements	Time	Resource
Mandatory Major Refinements	Robust PDF Parsing Capability ★	2 weeks	1.5
	LLM Architectural Modifications ★	2 weeks	1.5
	Build an Interface for GenAI Traceability – An alternative to Langsmith ★	2 week	1
Additional Optional Changes	Integration of GPT-4 Turbo	1 week	1
	Introduction of a Reranker during response generation	1 week	1

Thank you

PDF parser

Comparison

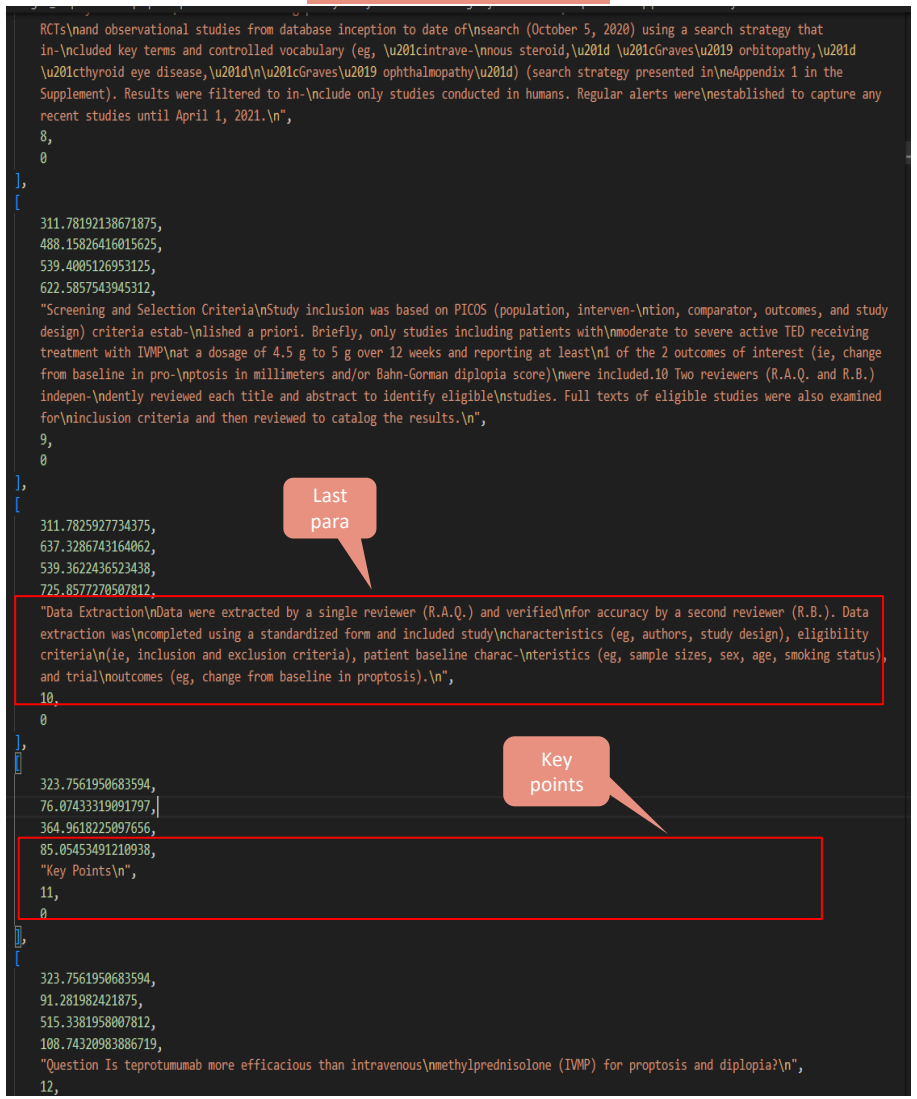
Criteria	PyMuPDF	Adobe Extract API (Services API)
Type	Python Library	Cloud API
Free Tier	Yes (Open-Source Python Library)	Yes (500 document transactions per month i.e., 2000 pages approx.)
Text Extraction Accuracy & Reading Multi Column	Above average – Able to read multi-column docs, divide pages into sections/blocks.	Most Accurate
OCR Capability / Scanned Documents	No, Need to integrate separately (EasyOCR / OCRmyPDF)	Yes
Paragraph Order	Paragraphs/Blocks not numbered. Continuity in between pages not maintained (majorly due to images & tables)	Numbered Elements. Order between paragraphs Maintained. Helps in creating complete sectional chunks.
Table Identification & Extraction	Improper, able to extract tables with boundaries defined by lines and boxes	Better (Mostly accurate, provides output in csv format)
Document Structure Identification / Tagging	No, Font Information is available	Yes (Helps in text cleaning like removing footnotes resulting in better chunks)
Image Identification & Extraction	Yes	Yes
Pricing	Free	$25K for 500K Transaction per year **or** VIP version $11,000 for 200,000 transactions per year.

Paragraph Order and Segmentation

Adobe Parser

- Ensure that the sequence of paragraphs, such as paragraphs 10 followed by paragraph 11, is maintained without any disruption.
- Key points is tagged differently hence help to identify right sequence.

PyMuPDF Parser



- Tables and images, similar to key points, are typically placed at the end of the page, potentially disrupting the following sequence.

Adobe Acrobat Document

Table Identification

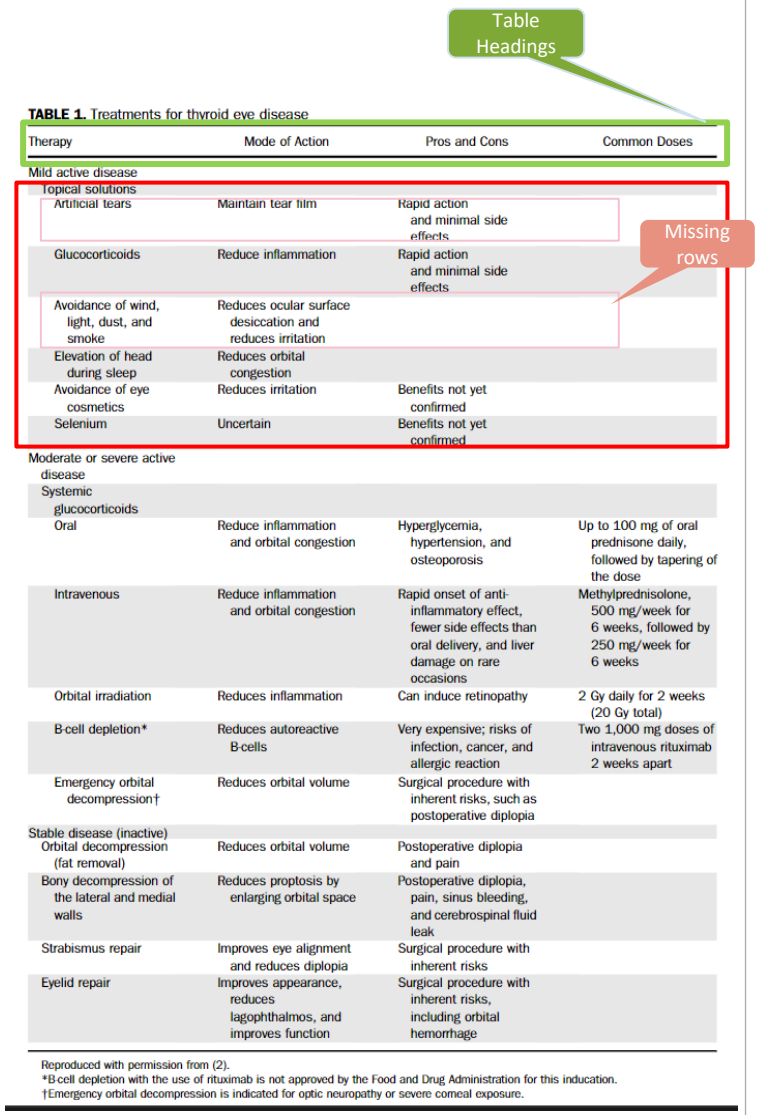


TABLE 1. Treatments for thyroid eye disease

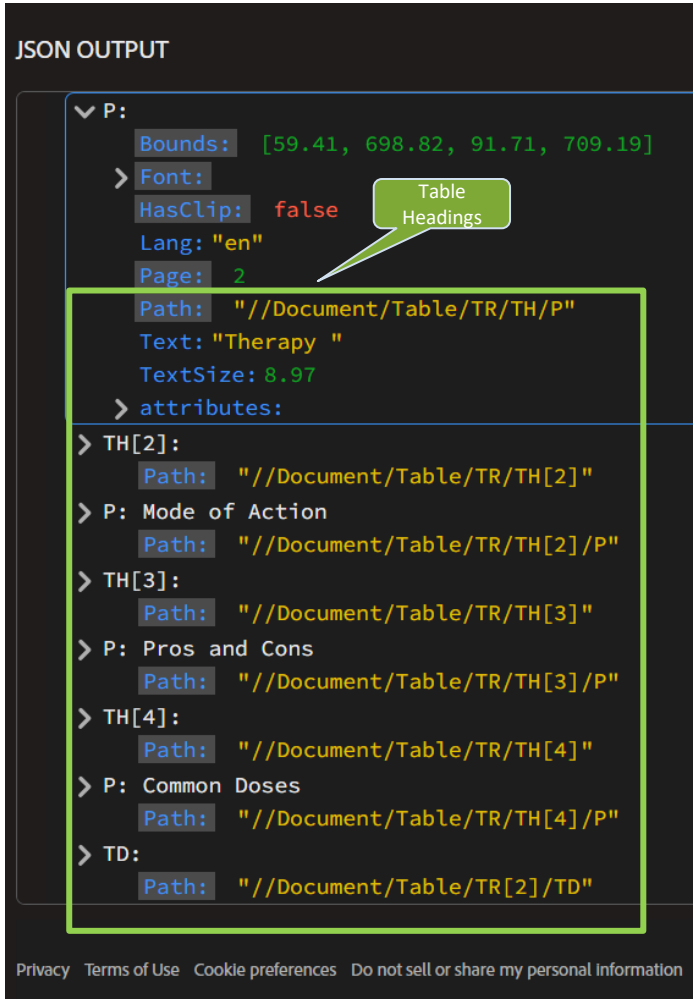
Therapy	Mode of Action	Pros and Cons	Common Doses
Mild active disease			
Topical solutions			
Artificial tears	Maintain tear film	Rapid action and minimal side effects	
Glucocorticoids	Reduce inflammation	Rapid action and minimal side effects	
Avoidance of wind, light, dust, and smoke	Reduces ocular surface desiccation and reduces irritation		
Elevation of head during sleep	Reduces orbital congestion		
Avoidance of eye cosmetics	Reduces irritation	Benefits not yet confirmed	
Selenium	Uncertain	Benefits not yet confirmed	
Moderate or severe active disease			
Systemic glucocorticoids			
Oral	Reduce inflammation and orbital congestion	Hyperglycemia, hypertension, and osteoporosis	Up to 100 mg of oral prednisone daily, followed by tapering of the dose
Intravenous	Reduce inflammation and orbital congestion	Rapid onset of anti-inflammatory effect, fewer side effects than oral delivery, and liver damage on rare occasions	Methylprednisolone, 500 mg/week for 6 weeks, followed by 250 mg/week for 6 weeks
Orbital irradiation	Reduces inflammation	Can induce retinopathy	2 Gy daily for 2 weeks (20 Gy total)
B-cell depletion*	Reduces autoreactive B-cells	Very expensive; risks of infection, cancer, and allergic reaction	Two 1,000 mg doses of intravenous rituximab 2 weeks apart
Emergency orbital decompression†	Reduces orbital volume	Surgical procedure with inherent risks, such as postoperative diplopia	
Stable disease (inactive)			
Orbital decompression (fat removal)	Reduces orbital volume	Postoperative diplopia and pain	
Bony decompression of the lateral and medial walls	Reduces proptosis by enlarging orbital space	Postoperative diplopia, pain, sinus bleeding, and cerebrospinal fluid leak	
Strabismus repair	Improves eye alignment and reduces diplopia	Surgical procedure with inherent risks	
Eyelid repair	Improves appearance, reduces lagophthalmos, and improves function	Surgical procedure with inherent risks, including orbital hemorrhage	

Reproduced with permission from (2).
*B-cell depletion with the use of rituximab is not approved by the Food and Drug Administration for this indication.
†Emergency orbital decompression is indicated for optic neuropathy or severe corneal exposure.

Table Headings

Missing rows

Adobe Parser



Accurate table identification along with row & column information

TH – Table Heading, TR – Table row, TD – Table data

PyMuPDF Parser

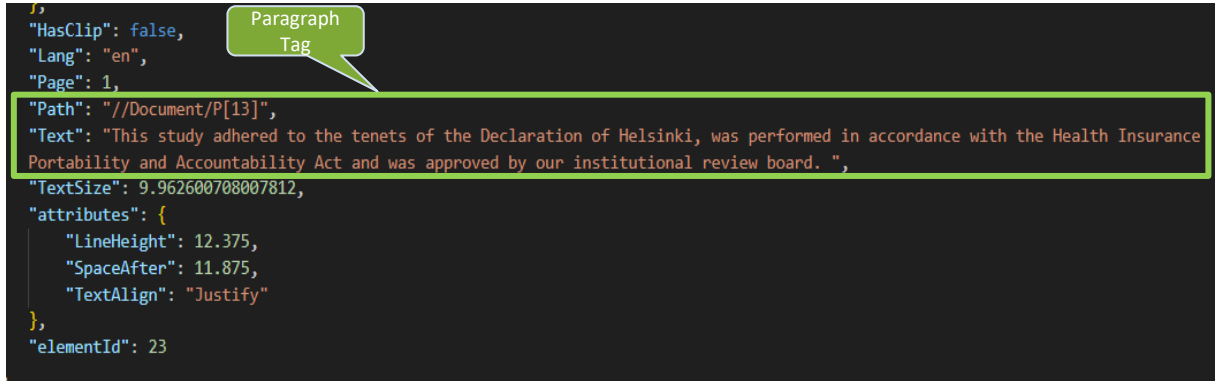
```
[['Topical solutions', '', '', '', '', '', '']]
[['Glucocorticoids', '', 'Reduce in\nfl\nammation', '', 'Rapid action\nand minimal side\neffects', '', '']]
[['Elevation of head\nduring sleep', '', 'Reduces orbital\ncongestion', '', '', '', '']]
[['Selenium', '', 'Uncertain', '', 'Bene\nfi\nts not yet\ncon\nfi\nrmed', '', '']]
[['Systemic\nglucocorticoids', '', '', '', '', '', '']]
[['Intravenous', '', 'Reduce in\nfl\nammation\nand orbital congestion', '', 'Rapid onset of anti-\nin\nfl\nammatory effe
[['B-cell depletion*', '', 'Reduces autoreactive\nB-cells', '', 'Very expensive; risks of\ninfection, cancer, and\nal
[['Stable disease (inactive)', '', '', '', '', '', '']]
[['Bony decompression of\nthe lateral and medial\nwalls', '', 'Reduces proptosis by\nenlarging orbital space', '', 'Po
[['Eyelid repair', '', 'Improves appearance,\nreduces\nlagophthalmos, and\nimproves function', '', 'Surgical procedure
```

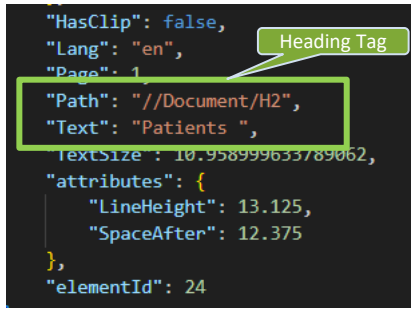
Not reading Table headings and missed rows

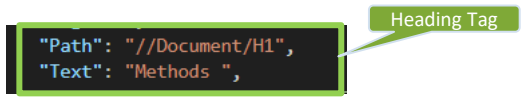
- Not able to identify table bounding boxes properly.
- Missed to extract some rows and headers.
- Tried setting strategy argument as ‘text’, didn’t worked as expected
- Works well only when table contains boundary (lines & boxes)

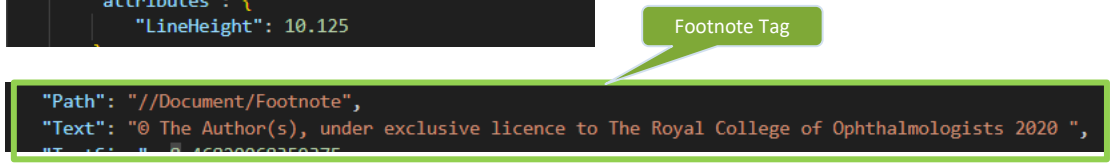
Tagging Identification

Adobe Parser



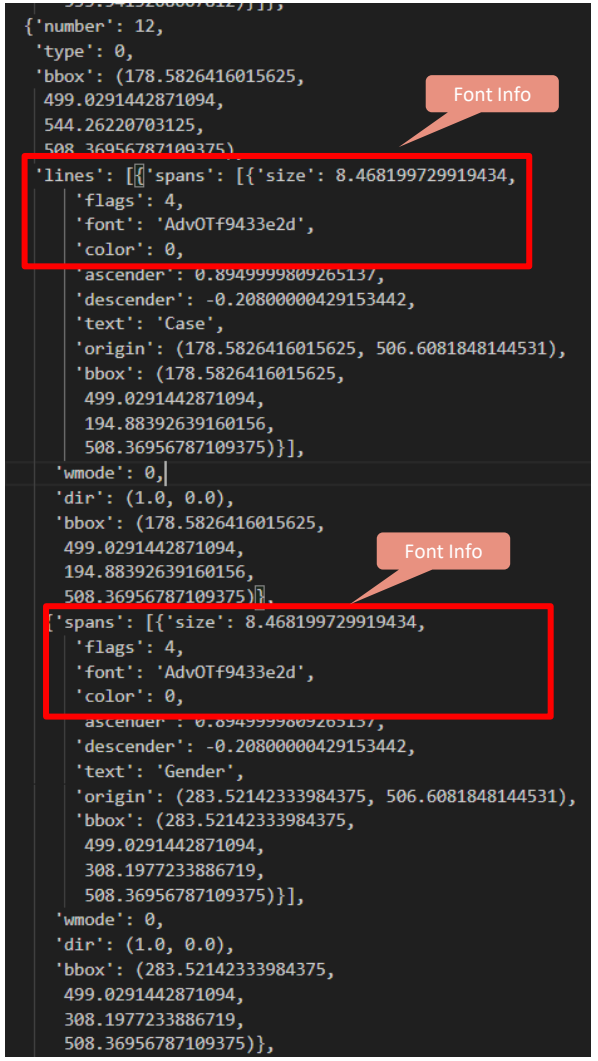




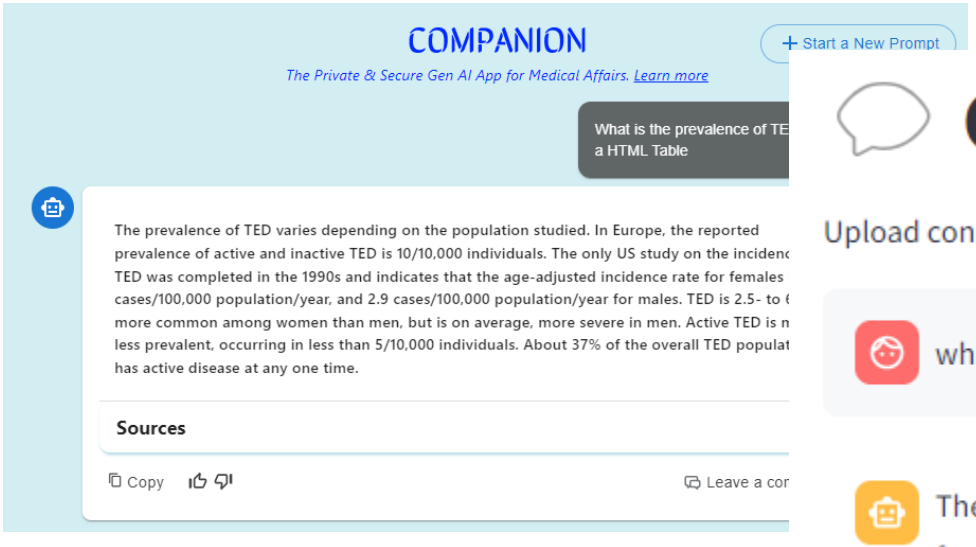


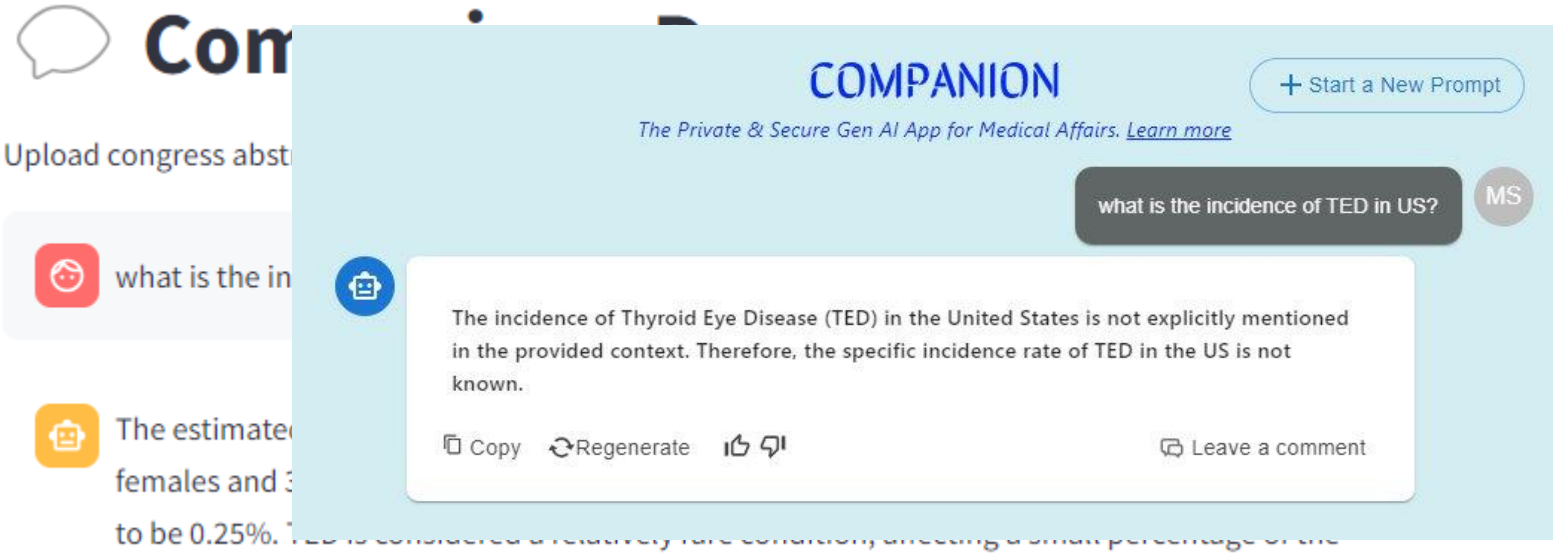
Parsed o/p is categorized by tags like Header, Paragraph, Footnote etc.

PyMuPDF Parser

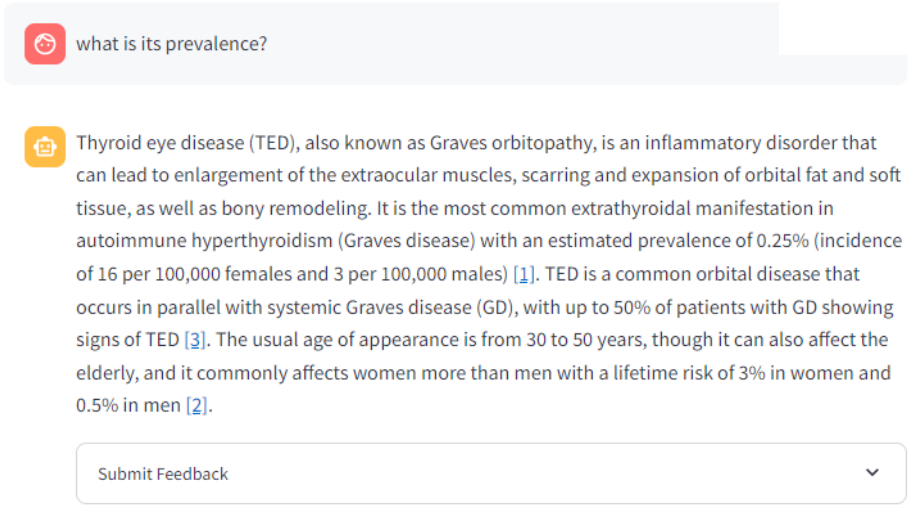


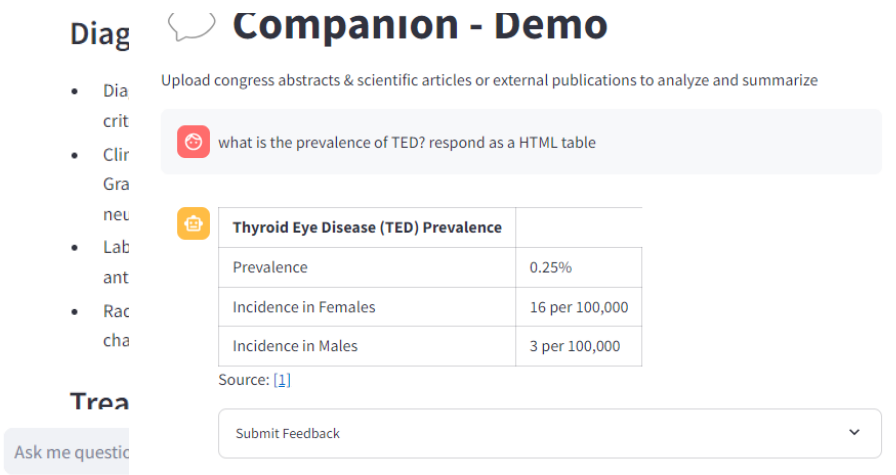
1. Doesn't categorize text by tag information
2. Additional Task – Tagging using given Font Information like font name and size.


COMPANION
Start a New Prompt
The Private & Secure Gen AI App for Medical Affairs. Learn more
What is the prevalence of TE
a HTML Table
The prevalence of TED varies depending on the population studied. In Europe, the reported prevalence of active and inactive TED is 10/10,000 individuals. The only US study on the incidenc TED was completed in the 1990s and indicates that the age-adjusted incidence rate for females cases/100,000 population/year, and 2.9 cases/100,000 population/year for males. TED is 2.5- to more common among women than men, but is on average, more severe in men. Active TED is n less prevalent, occurring in less than 5/10,000 individuals. About 37% of the overall TED populat has active disease at any one time.
Sources
Copy
Leave a cor


Com
Upload congress abst
what is the in
The estimate
females and 3
to be 0.25%.
population. [1]

COMPANION
Start a New Prompt
The Private & Secure Gen AI App for Medical Affairs. Learn more
what is the incidence of TED in US?
MS
The incidence of Thyroid Eye Disease (TED) in the United States is not explicitly mentioned in the provided context. Therefore, the specific incidence rate of TED in the US is not known.
Copy
Regenerate
Leave a comment


what is its prevalence?
Thyroid eye disease (TED), also known as Graves orbitopathy, is an inflammatory disorder that can lead to enlargement of the extraocular muscles, scarring and expansion of orbital fat and soft tissue, as well as bony remodeling. It is the most common extrathyroidal manifestation in autoimmune hyperthyroidism (Graves disease) with an estimated prevalence of 0.25% (incidence of 16 per 100,000 females and 3 per 100,000 males) [1]. TED is a common orbital disease that occurs in parallel with systemic Graves disease (GD), with up to 50% of patients with GD showing signs of TED [3]. The usual age of appearance is from 30 to 50 years, though it can also affect the elderly, and it commonly affects women more than men with a lifetime risk of 3% in women and 0.5% in men [2].
Submit Feedback


Companion - Demo
Upload congress abstracts & scientific articles or external publications to analyze and summarize
what is the prevalence of TED? respond as a HTML table
Thyroid Eye Disease (TED) Prevalence
Prevalence 0.25%
Incidence in Females 16 per 100,000
Incidence in Males 3 per 100,000
Source: [1]
Submit Feedback

Illustrative

Examples – Companion response quality Comparison

- Aamir - Provide some dummy examples & Expected difference in response.
- To present the actual comparison demonstration would require a significant implementation effort, which is equivalent to executing the actual refinement steps.

Proposed solution

Foundation

- Strong PDF parser – Table, text, text order, Images, key points, - complex documents.
- Architectural changes – Manager and Worker nodes
- Feedback tracking system – Alternative of Langsmith GenAI Tracing

Additional Add on Changes (Good to have)

- GPT 4 turbo
- Reranker

Current System architecture

PDF parser

- Reading document correctly is crucial
- Better the chunk – better the response quality

AI Model

- Agent
- Does n’t do thinking
- Multi – instructions doesn’t work
- Interpret the prompt, thinking and take action

Examples – PDF parser (Parsing Comparison)

Comparison

- Aamir - Provide some dummy examples & Expected difference in response.
- For actual comparison demonstration would take some time to implement it. Which will be equivalent to actual implementation of refinements steps

Enhancing User Experience and User Engagement

- Displaying Progress Bar
- Displaying live Steps taken by the Companion application to Generate the Response