# TRINITY

# Companion – User feedback analysis

Refinement plan to incorporate user inputs and other Advance features

February 2, 2024

# Agenda Slide

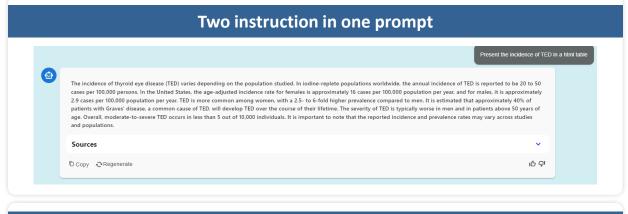
#### **User feedback consolidation**

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

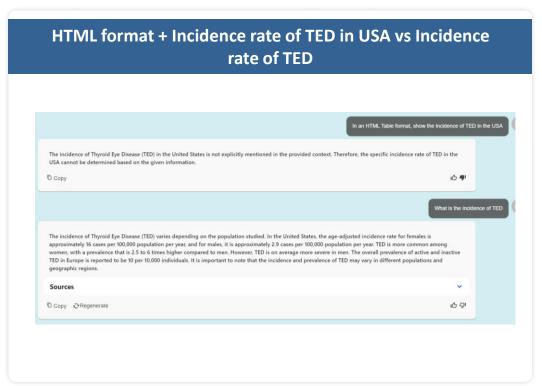


## User Feedback Snippets











# 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

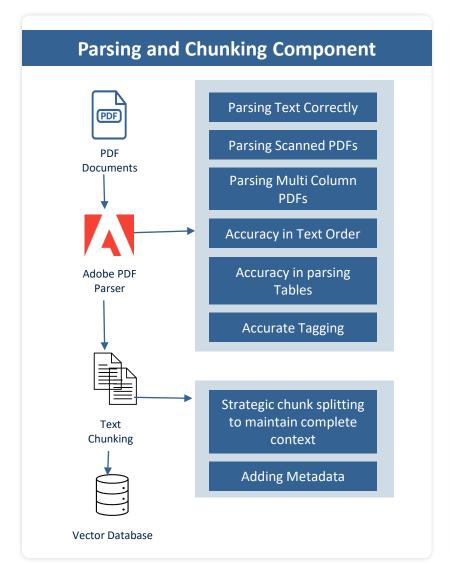
- User feedback consolidation
- User Feedback Summary and other Advance features

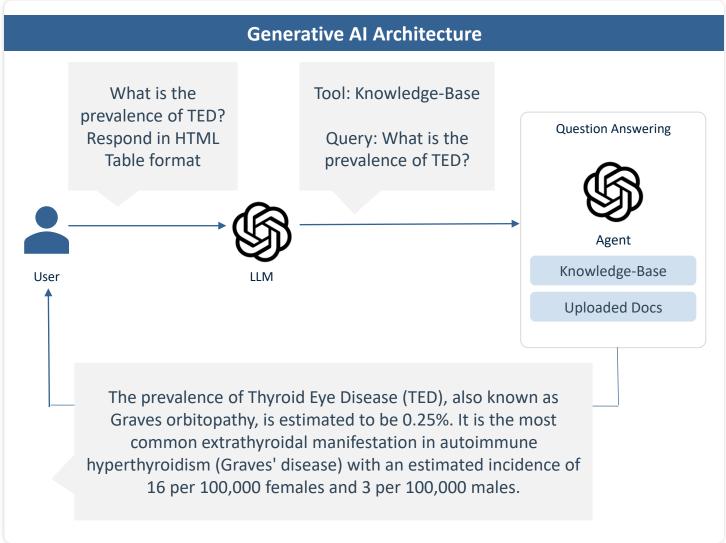
# **Current system Architecture**

- Application Refinement Plan
- Effort estimation



#### **Current Architecture**





# 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 GenAl Traceability – An alternative to Langsmith

- Capability to monitor all user activities, including GenAl 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**

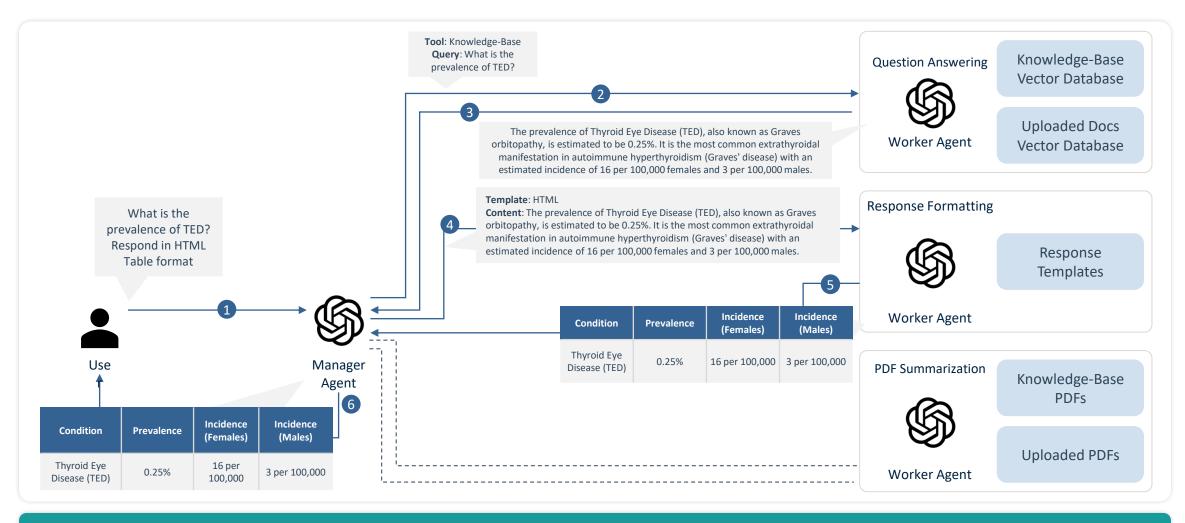
**Rationale**: A superior model with enhanced reasoning and improved query/answer abilities compared to GPT-3.5. Additionally, GPT-4 Turbo is more cost-effective than GPT 4.

#### 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.

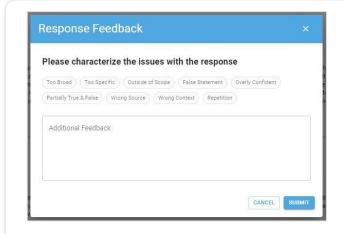
Note: The drawback of implementing this is the potential for increased latency and possible expenses due to LLM calls, depending on the complexity of the queries

#### Proposed System architecture

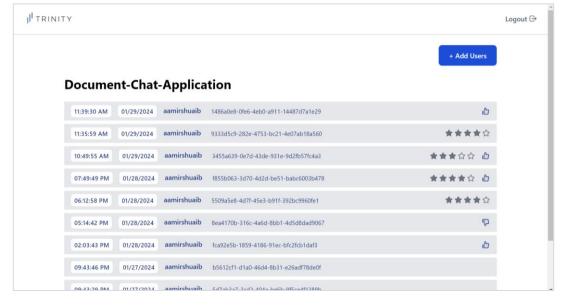


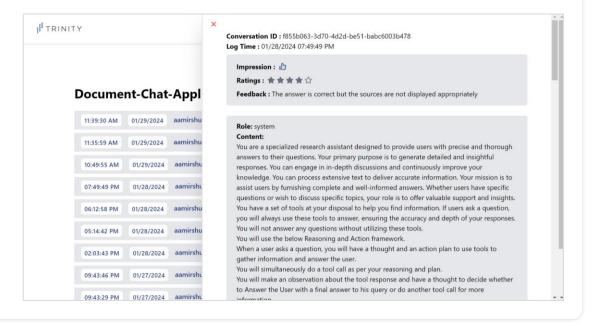
Note: The drawback of implementing this is the potential for increased latency and possible expenses due to LLM calls, depending on the complexity of the queries

## User Feedback Tracking

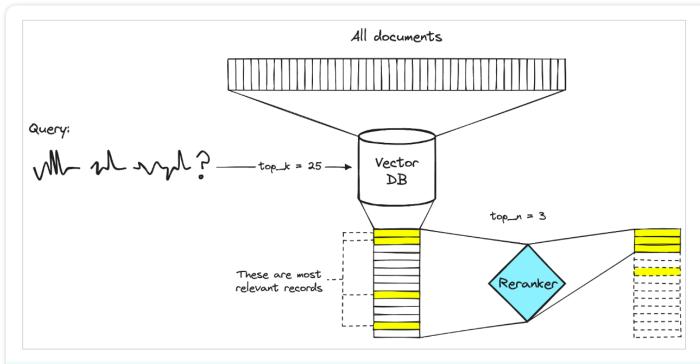


- Embedded Feedback collection component within the Companion Application
- Introducing a new application for continuous tracking.
- The Application tracks Gen-Al Operations/Flow and User Feedback
- This will help in improved assessment of the flagged/disliked responses





## Re-Ranking



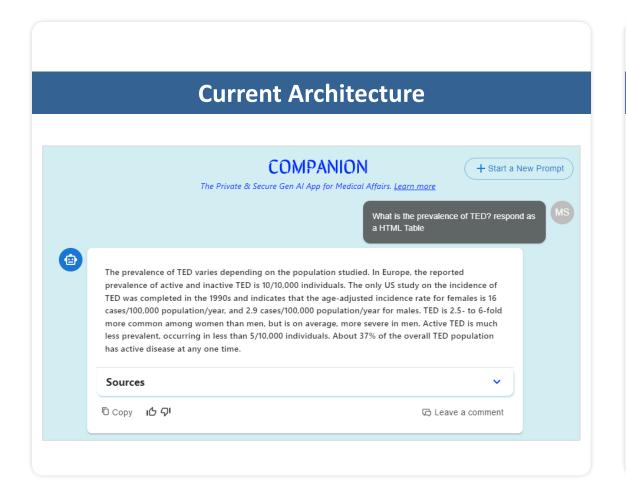
Re-ranking in Retrieval-Augmented Generation (RAG) applications is essential due to its ability to improve the quality of generated responses

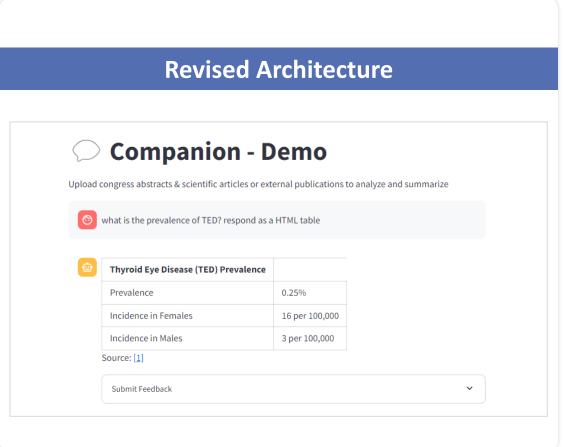
#### **Advantages**

- 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

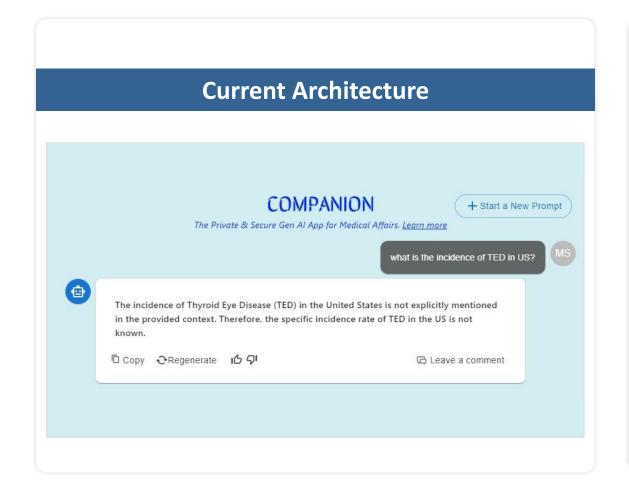


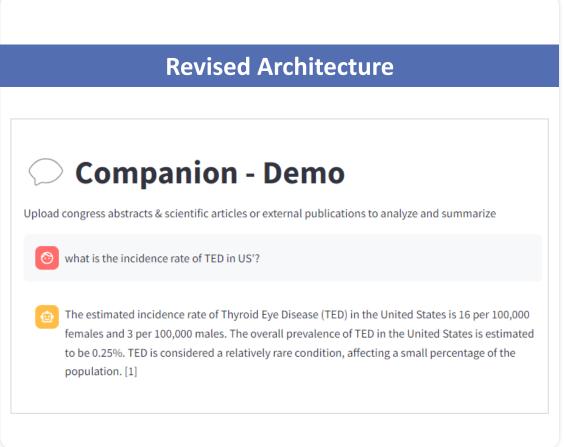
## Illustration – Companion response quality Comparison w.r.t Architecture



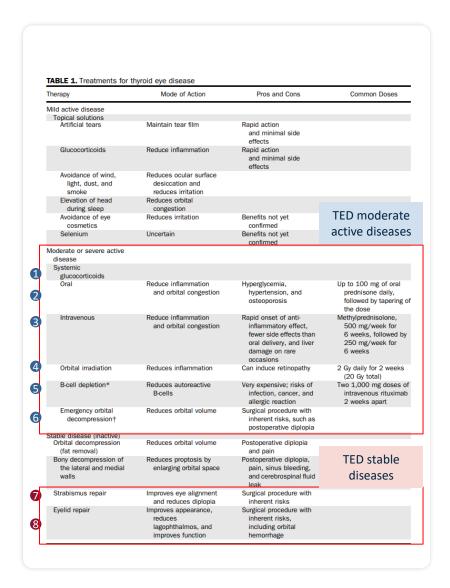


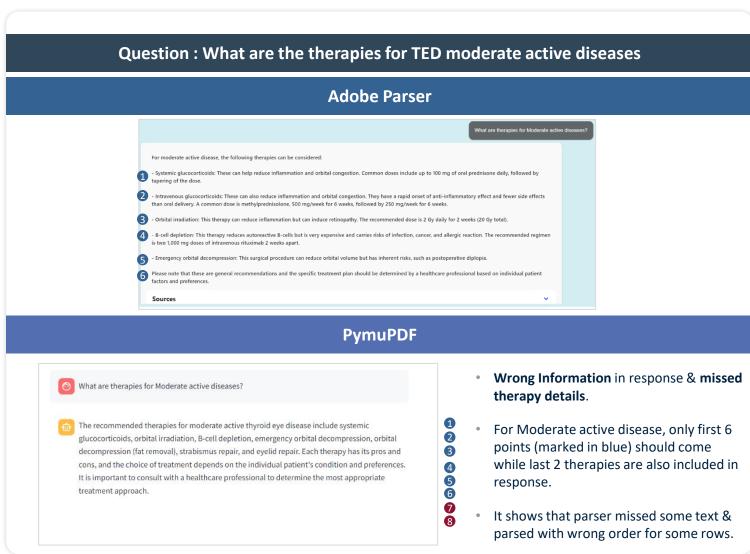
## Illustration – Companion response quality Comparison w.r.t Architecture



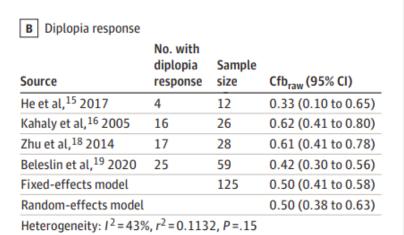


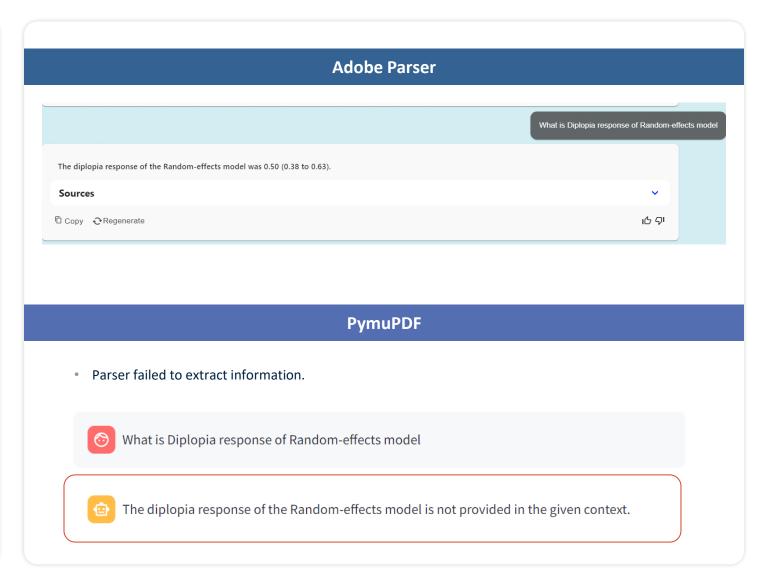
## Examples – Companion response quality Comparison w.r.t PDF parser





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





# 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 GenAl 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



# PDF parser

#### Comparison

Criteria	PyMuPDF	MuPDF Adobe Extract API (Services API)	
Туре	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 <b>or</b> VIP version \$11,000 for 200,000 transactions per year.	



## Paragraph Order and Segmentation

onfirm the clinical relevance of this finding.

inclusion and exclusion criteria, data were pooled to obtain

treatment arms with 84 randomized patients and 87

randomized patients for teprotumumab and placebo.

lished literature assessing the most commonly recommended

dose of IVMP among patients with moderate to severe active

search (October 5, 2020) using a search strategy that in-

cluded key terms and controlled vocabulary (eg, "intrave-

"Graves' ophthalmopathy") (search strategy presented in

eAppendix 1 in the Supplement). Results were filtered to in-

established to capture any recent studies until April 1, 2021.

study inclusion was based on PICOS (population, interven-

tion, comparator, outcomes, and study design) criteria estab-

1 of the 2 outcomes of interest (ie, change from baseline in pro-

ptosis in millimeters and/or Bahn-Gorman diplopia score)

were included.10 Two reviewers (R.A.Q. and R.B.) indepen-

dently reviewed each title and abstract to identify eligible

studies. Full texts of eligible studies were also examined fo inclusion criteria and then reviewed to catalog the results.

Original Investigation Research

points Proptosis and Diplopia Response in Moderate to Severe Thyroid Eve Disease hypoid eve disease (TED), or Graves ophthalmopathy. is an autoimmune disorder characterized by progressive inflammation and damage to orbital and ocular estion Is teprotumumab more efficacious than intravenous tissues.1,2 Age-adjusted prevalence in the US is estimated at methylprednisolone (IVMP) for proptosis and diplopia? 0.25%.3 Thyroid eye disease causes expansion of retrondings This meta-analysis and matching-adjusted indirect orbital fat and extraocular muscle, thought to be mediated mnarison showed an association with small improvements in primarily by the upregulation of the insulin like growth facoptosis from baseline for IVMP vs placebo (-0.16 mm); tor 1 receptor on orbital fibroblasts. Patients may develop ociated proptosis improvements were statistically significantl considerable disfiguring facial changes owing to proptosis, greater with teprotumumab vs IVMP (treatment difference disabling diplopia, and in severe cases, vision loss.1 -2.31 mm). For diplopia response, IVMP was not favored over placebo while teprotumumab was favored over IVMP. currently there are limited noninvasive treatment options that improve proptosis and diplopia. The Para ning Improvements in proptosis and diplopia with IVMP vs European Group on Graves' Orbitopathy (EUG placebo may be small/not clinically relevant; in this meta-analysis lines recommend a cumulative dosage of 4.5 to eprotumumab was associated with greater improvements in venous methylpredgisologe (IVMP) over 12 weeks for mos ptosis and diplopia vs IVMP, but clinical trials are needed to

clinically meaningful in prior TED clinical trials. On January 21, 2020, teprotumumab became the first US | Unterature Review for IVMP Food and Drug Administration-approved treatment for TED. 5,6 A literature review was conducted to identify existing pub-Teprotumumab, a fully human, monoclonal antibody, inhibits insulin like growth factor 1 receptor activity and reduces downstream pathogenic signaling in TED. A total of 2 placebo- TED. 9 PubMed and Embase were searched for relevant RCTs controlled, double-masked, randomized clinical trials (RCTs) and observational studies from database inception to date of of patients with moderate to severe TED demonstrated that teprotumumab was associated with clinically significant reductions in inflammation, proptosis, and diplopia over | nous steroid," "Graves' orbitopathy," "thyroid eye disease," 24 weeks,7,8

tients with moderate to severe active TED.4 Although da

demonstrate that IVMP is associated with reduced inflamma

tion, the dose, timing of administration, and duration of

therapy vary in the literature, making it challenging to com-

pare the clinical results, particularly on the progressive out-

comes of proptosts and diplopta. A 2-mm reduction in propto-

sts and a 1-grade improvement in diplopta have been considered

To our knowledge, there are currently no studies directly comparing the efficacy of the most recommended dose of | clude only studies conducted in humans. Regular alerts were IVMP with teprotumumab or placebo; as such, matchingadjusted indirect comparisons (MAICs) simulating direct comparisons between treatments can be used to estimate | Screening and Selection Criteria comparative treatment effects. The objectives of this study are to (1) to evaluate improvements in proptosis and diplopia with the most recommended treatment regimen of IVMP as lished a priori. Briefly, only studies including patients with reported in the literature and (2) to compare these results | moderate to severe active TED receiving treatment with IVMP teprotumumab and placebo in patients with moderate at a dosage of 4.5 g to 5 g over 12 weeks and reporting at least

evere active TED using MAICs.

#### Method

pata sources included deidentified patient-level data for tepro-

#### Adobe Parser

```
ISON OUTPUT
    > P[7]: Currently there are limited noninvasive treatment opti...
         Path: "//Document/P[7]'
    > P[8]: On January 21, 2020, teprotumumab became the first US ...
         Path: "//Document/P[8]"
    > P[9]: To our knowledge, there are currently no studies direc...
         Path: "//Document/P[9]"
    > H1: Methods
         Path: "//Document/H1"
    > H2: Patients Receiving Teprotumumab and Placebo
    P[10]: Data sources included deidentified patient-level data...
         Path: "//Document/P[10]"
    ∨P:
         Bounds: [323.76, 706.48, 367.20, 718.19]
       > Font:
         HasClip: false
        Page:
         Path: "//Document/Aside[2]/P"
         Text: "Key Points "
       > attributes:
    > P[2]: Question Is teprotumumab more efficacious than intrave...
         Path: "//Document/Aside[2]/P[2]"
    > P[3]: Findings This meta-analysis and matching-adjusted indi...
         Path: "//Document/Aside[2]/P[3]"
    > P[4]: Meaning Improvements in proptosis and diplopia with IV...
          Path: "//Document/Aside[2]/P[4]"
    > P[11]: inclusion and exclusion criteria, data were pooled to...
         Path: "//Document/P[11]"
   > H2[2]: Literature Review for IVMP
         Path: "//Document/H2[2]"
    > P[12]: A literature review was conducted to identify existin...
         Path: "//Document/P[12]"
    > H2[3]: Screening and Selection Criteria
         Path: "//Document/H2[3]"
    > P[13]: Study inclusion was based on PICOS (population, inter...
         Path: "//Document/P[13]"
```

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

#### PvMuPDF Parser

```
RCTs\nand observational studies from database inception to date of\nsearch (October 5, 2020) using a search strategy that
in-\ncluded key terms and controlled vocabulary (eg, \u201cintrave-\nnous steroid,\u201d \u201cGraves\u2019 orbitopathy,\u201d
 \u201cthyroid eye disease,\u201d\n\u201cGraves\u2019 ophthalmopathy\u201d) (search strategy presented in\neAppendix 1 in the
Supplement). Results were filtered to in-\nclude only studies conducted in humans. Regular alerts were\nestablished to capture any
311.78192138671875.
488.15826416015625.
539.4005126953125,
622.5857543945312.
 "Screening and Selection Criteria\nStudy inclusion was based on PICOS (population, interven-\ntion, comparator, outcomes, and study
design) criteria estab-\nlished a priori. Briefly, only studies including patients with\nmoderate to severe active TED receiving
treatment with IVMP\nat a dosage of 4.5 g to 5 g over 12 weeks and reporting at least\n1 of the 2 outcomes of interest (ie, change
from baseline in pro-\nptosis in millimeters and/or Bahn-Gorman diplopia score)\nwere included.10 Two reviewers (R.A.Q. and R.B.)
 indepen-\ndently reviewed each title and abstract to identify eligible\nstudies. Full texts of eligible studies were also examined
for\ninclusion criteria and then reviewed to catalog the results.\n",
311.7825927734375.
637.3286743164062,
539.3622436523438.
725.8577270507812
 "Data Extraction\nData were extracted by a single reviewer (R.A.O.) and verified\nfor accuracy by a second reviewer (R.B.). Data
extraction was\ncompleted using a standardized form and included study\ncharacteristics (eg, authors, study design), eligibility
 criteria\n(ie, inclusion and exclusion criteria), patient baseline charac-\nteristics (eg, sample sizes, sex, age, smoking status)
and trial\noutcomes (eg, change from baseline in proptosis).\n",
 323.7561950683594
76.07433319091797.
364.9618225097656
85.05453491210938.
 "Key Points\n",
323.7561950683594,
91.281982421875.
515.3381958007812.
 "Question Is teprotumumab more efficacious than intravenous\nmethylprednisolone (IVMP) for proptosis and diplopia?\n",
```

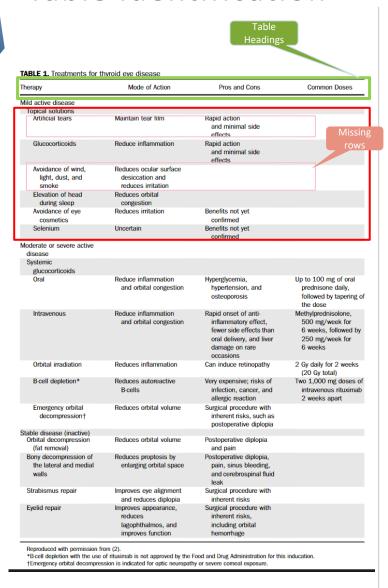
Tables and images, similar to key points, are typically placed at the end of the page, potentially Adobe Acrobat disrupting the following sequence.



tumumab or placebo from the phase 2 (NCTO1868997) and 3 Data Extraction Data were extracted by a single reviewer (R.A.Q.) and verified (NCTO3298867) trials and published aggregate-level data for IVMP (4.5-5 g over 12 weeks). Data for patients receiving for accuracy by a second reviewer (R.B.). Data extraction was teprotumumab or placebo were obtained from 2 published completed using a standardized form and included study trials; a phase 2 trial that included 43 patients and 45 patients characteristics (eg. authors, study design), eligibility criteria in the teprotumumab and placebo groups, respectively, and a (ie, inclusion and exclusion criteria), patient baseline characphase 3 trial that included 41 patients and 42 patients in teristics (eg, sample sizes, sex, age, smoking status), and trial the teprotumumab and placebo groups.7,8 Given the similar outcomes (eg, change from baseline in proptosis).

Ensure that the sequence of paragraphs, such as

#### Table Identification



Adobe Parser

```
JSON OUTPUT
     ∨ P:
        > Font:
          Lang: "en"
          Path: "//Document/Table/TR/TH/P"
          Text: "Therapy "
        > attributes:
     > TH[2]:
                  "//Document/Table/TR/TH[2]"
     > P: Mode of Action
                  "//Document/Table/TR/TH[2]/P"
     > TH[3]:
                  "//Document/Table/TR/TH[3]"
     > P: Pros and Cons
                  "//Document/Table/TR/TH[3]/P"
     > TH[4]:
                  "//Document/Table/TR/TH[4]"
     > P: Common Doses
                  "//Document/Table/TR/TH[4]/P"
     > TD:
                  "//Document/Table/TR[2]/TD"
Privacy Terms of Use Cookie preferences Do not sell or share my personal information
```

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\nfl\nrmed', '', '']]

[['Systemic\nglucocorticoids', '', '', '', '', '']]

[['Intravenous', '', 'Reduce in\nfl\nammation\nand orbital congestion', '', 'Rapid onset of anti-\nin\nfl\nammatory eff

[['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', '', 'P

[['Eyelid repair', '', 'Improves appearance,\nreduces\nlagophthalmos, and\nimproves function', '', 'Surgical procedur
```

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

```
"HasClip": false,
"Lang": "en",
"Page": 1,

"Path": "//Document/P[13]",
"Text": "This study adhered to the tenets of the Declaration of Helsinki, was performed in accordance with the Health Insurance Portability and Accountability Act and was approved by our institutional review board. ",

"TextSize": 9.962600708007812,
"attributes": {

    "LineHeight": 12.375,
    "SpaceAfter": 11.875,
    "TextAlign": "Justify"
},
"elementId": 23
```

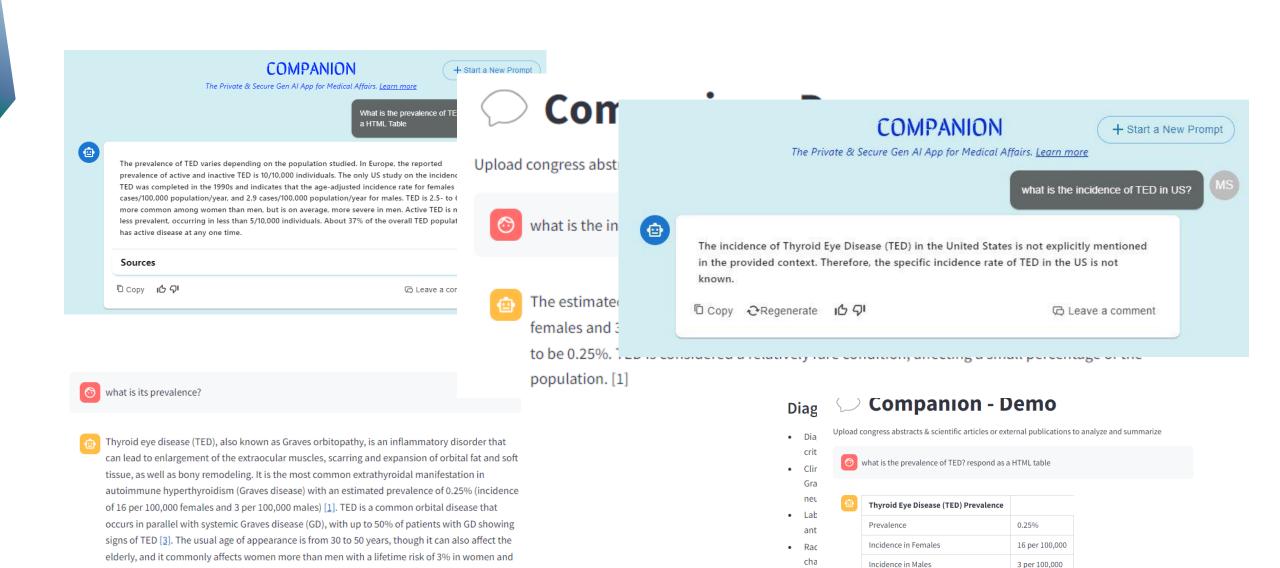
```
"HasClip": false,
         "alt_family_name": "Adv OTf 943 3e 2d",
                                                                                  Heading Tag
                                                            "Lang": "en",
        "embedded": true,
                                                            "Page" · 1
        "encoding": "WinAnsiEncoding",
                                                            "Path": "//Document/H2",
        "family name": "Adv OTf 943 3e 2d",
                                                            "Text": "Patients ",
        "font_type": "Type1",
                                                             Text51ze": 10.958999633/89062,
        "italic": false,
                                                            "attributes": {
        "monospaced": false,
                                                                "LineHeight": 13.125,
        "name": "HAGOIL+AdvOTf9433e2d",
                                                                "SpaceAfter": 12.375
        "subset": true,
        "weight": 400
                                                            "elementId": 24
    "HasClip": false,
    "Lang": "en",
                                                                                            Heading Tag
    "Page": 3.
                                                            "Path": "//Document/H1",
    "Path": "//Document/Table[2]/TR[11]/TD[6]/P",
                                                           "Text": "Methods ",
    "Text": "3 ",
     "TextSize": 8.46820068359375,
    "attributes": {
        "LineHeight": 10.125
                                                            Footnote Tag
"Path": "//Document/Footnote",
"Text": "@ The Author(s), under exclusive licence to The Royal College of Ophthalmologists 2020 "
```

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

#### PyMuPDF Parser

```
'number': 12,
'type': 0,
'bbox': (178.5826416015625,
499.0291442871094,
544.26220703125.
'lines': [{|'spans': [{'size': 8.468199729919434,
   'flags': 4,
    'font': 'AdvOTf9433e2d',
    'color': 0,
    'ascender': 0.894999980926513/,
    'descender': -0.20800000429153442,
    'text': 'Case',
    'origin': (178.5826416015625, 506.6081848144531),
    'bbox': (178.5826416015625,
    499.0291442871094,
    194.88392639160156,
    508.36956787109375)}],
  'wmode': 0,
  'dir': (1.0, 0.0),
  'bbox': (178.5826416015625,
  499.0291442871094,
  194.88392639160156,
  508.36956787109375)
  'spans': [{'size': 8.468199729919434,
   'flags': 4,
    'font': 'AdvOTf9433e2d',
   'color': 0,
    'descender': -0.20800000429153442,
    'text': 'Gender',
    'origin': (283.52142333984375, 506.6081848144531)
    'bbox': (283.52142333984375,
    499.0291442871094,
    308.1977233886719,
    508.36956787109375)}],
  'wmode': 0,
  'dir': (1.0, 0.0),
  'bbox': (283.52142333984375,
  499.0291442871094,
  308.1977233886719,
  508.36956787109375)}
```

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



Source: [1]

Submit Feedback

Trea

Ask me questic

اا

0.5% in men [2].

Submit Feedback

## 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 GenAl 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

#### Al Model

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



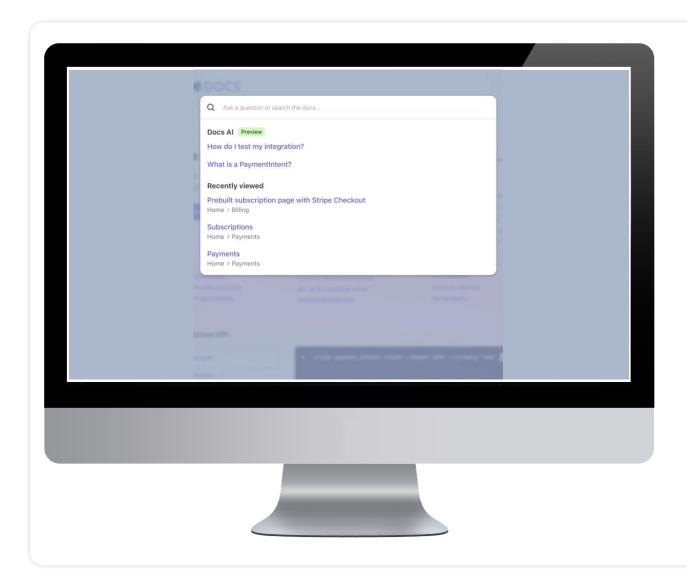
## Examples - PDF parser (Parsing Comparison)

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