TRINITY

Companion – Med Affair and Internal Feedback Analysis

Refinement plan to incorporate user inputs and other Advance features

April 18, 2024

Agenda

User feedback Consolidation

Causal Analysis on constructive feedback

Proposed solution

User Feedback Snippets

Positive Comments

It provided a good response for the implications on practice and future research needs.
– Margarita and Alnoor

This question is hard to gauge accuracy overall since I can't see which studies its citing, but the answer seems to be accurate. If citing multiple studies, it would be nice to show results from individual case reports in different sections and then a pooled section similar to how the phase 2/3 answer was structured.
– Margarita and Alnoor

Overall like the structure of the response, showing Primary/Secondary/Surrogate endpoints in different sections. I think it needs to be clearer of which study it is basing its response from.
– Margarita and Alnoor

This is a good response to providing an overview of the Ugradar paper. It provides the results in simple bullets with different sections. It makes it clear what the overall results were.
– Margarita and Alnoor

I like the structure of the response with 3 different sections providing information. This question is hard to gauge accuracy overall since I can't see which studies its citing, but the answer seems to be accurate.
– Margarita and Alnoor

Consideration Comments

Iterative process may be needed for gender differentiation and baseline risk factors for Thyroid disease - such as iodine and selenium deficiencies.
– Margarita and Alnoor

Presently we have challenges due to lexicon evolution and the fact that Thyroid Eye Disease has several lexicon words that share the same definitions (Thyroid Ophthalmopathy for example)- Consider discussing this delicate topic internally when we meet
– Margarita and Alnoor

This is an area where images may be useful
– Margarita and Alnoor

Consider educational process and iteration based on user understanding
– Margarita and Alnoor

The diagnosis should make reference to the guidelines as well and the possibility of atypical presentations
– Margarita and Alnoor

Rewrite query didn't show the response in desired format (**)
– Koushika

From TARLATAMAB use case, Use Image in the response to display related poster, flowchart and video etc.
– Delilah

Constructive Comments

This answer appears to be including both studies in the answer. This response is confusing because the two studies are so different. It would be better to separate the inclusion/ exclusion criteria per study since they are so different.
– Margarita and Alnoor

This answer appears to be including both studies in the answer. This response is confusing because the two studies are so different. It would be better to separate the inclusion/ exclusion criteria per study since they are so different. It almost makes it seem like there were patients on steroids in the clinical trial when there was not. *– Margarita and Alnoor*

Convert above response in html table format (The previous prompt was Rewrite this response with headings and bullet points formatting) The response looks like what's below. It doesn't look like a table.
– Nick

Model Hallucination – What is Green Tea, What is the Weather
– Nick

User latency
– Nick

User feedback summary and Advance feature's theme

Feedback Digest

Response Structure:
Organizing responses into sub-sections for clarity.

Specific Information:
Responses were somewhat generic and lacked more specific information.

Confusion from Mixed Results:
Responses sometimes mixed results from different studies or trials, leading to confusion.

Comprehensive Summaries:
Questions asking for summaries should be more comprehensive and detailed, containing larger amounts of information.

Iterative Approach:
Users suggested an iterative approach, where follow-up questions are asked to narrow down information.

Hallucinations:
Sometimes AI generated incorrect responses, providing information not present in the knowledge base.

Other Advance Features

Displaying Images in the responses

Employing Educational process by iterating based on user understanding

Include Video transcripts in the context

Agenda

User feedback Consolidation

Causal Analysis on constructive feedback

Proposed solution

Reasons for the Drawbacks

Causal Analysis from User Feedback

Token Limit Impact

Due to limited context, responses may lack comprehensiveness, specific information, and may mix results, affecting the quality of responses. Increasing context can improve response quality.

Response Structure	Specific Information	Confusion from Mixed Results	Comprehensive Summaries	Iterative Approach	Hallucinations

Multi tasking Capability

Models with lower multi-tasking capability may ignore some instructions, leading to issues like hallucinations. Higher multi-tasking capability improves adherence to instructions, reducing hallucinations and enhancing response quality.

Response Structure	Specific Information	Confusion from Mixed Results	Comprehensive Summaries	Iterative Approach	Hallucinations

Note: OpenAI's GPT-3.5-turbo model has a context window of 16,000 tokens, limiting the amount of information that can be provided to the model.

Token Limit Impact

Large PDF Documents

Proptosis and Diplopia Response in Moderate to Severe Thyroid Eye Disease — Original Investigation Research

Thyroid eye disease (TED), or Graves ophthalmopathy, is an autoimmune disorder characterized by progressive inflammation and damage to orbital and ocular tissues.[1,2] Age-adjusted prevalence in the US is estimated at 0.25%.[3] Thyroid eye disease causes expansion of retro-orbital fat and extraocular muscle, thought to be mediated primarily by the upregulation of the insulin like growth factor 1 receptor on orbital fibroblasts.[1] Patients may develop considerable disfiguring facial changes owing to proptosis, disabling diplopia, and in severe cases, vision loss.[1]

Currently there are limited nonINVASIVE treatment options that improve proptosis and diplopia. The most recent European Group on Graves' Orbitopathy (EUGOGO) guidelines recommend a cumulative dosage of 4.5 to 5.0 g of intravenous methylprednisolone (IVMP) over 12 weeks for most patients with moderate to severe active TED.[4] Although data demonstrate that IVMP is associated with reduced inflammation, the dose, timing of administration, and duration of therapy vary in the literature, making it challenging to compare the clinical results, particularly on the progressive outcomes of proptosis and diplopia. A 2-mm reduction in proptosis and a 1-grade improvement in diplopia have been considered clinically meaningful in prior TED clinical trials.

On January 21, 2020, teprotumumab became the first US Food and Drug Administration-approved treatment for TED.[5,6] 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-controlled, double-masked, randomized clinical trials (RCTs) of patients with moderate to severe TED demonstrated that teprotumumab was associated with clinically significant reductions in inflammation, proptosis, and diplopia over 24 weeks.[7,8]

To our knowledge, there are currently no studies directly comparing the efficacy of the most recommended dose of IVMP with teprotumumab or placebo; as such, matching-adjusted indirect comparisons (MAICs) simulating direct comparisons between treatments can be used to estimate 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 reported in the literature and (2) to compare these results with teprotumumab and placebo in patients with moderate to severe active TED using MAICs.

Key Points

Question Is teprotumumab more efficacious than intravenous methylprednisolone (IVMP) for proptosis and diplopia?

Findings This meta-analysis and matching-adjusted indirect comparison showed an association with small improvements in proptosis from baseline for IVMP vs placebo (−0.16 mm); associated proptosis improvements were statistically significantly greater with teprotumumab vs IVMP (treatment difference, −2.31 mm). For diplopia response, IVMP was not favored over placebo while teprotumumab was favored over IVMP.

Meaning Improvements in proptosis and diplopia with IVMP vs placebo may be small/not clinically relevant; in this meta-analysis, teprotumumab was associated with greater improvements in proptosis and diplopia vs IVMP, but clinical trials are needed to confirm the clinical relevance of this finding.

Methods

Patients Receiving Teprotumumab and Placebo

Data sources included deidentified patient-level data for teprotumumab or placebo from the phase 2 (NCT01868997) and 3 (NCT03298867) trials and published aggregate-level data for IVMP (4.5-5 g over 12 weeks). Data for patients receiving teprotumumab or placebo were obtained from 2 published trials: a phase 2 trial that included 43 patients and 45 patients in the teprotumumab and placebo groups, respectively, and a phase 3 trial that included 41 patients and 42 patients in the teprotumumab and placebo groups.[7,8] Given the similar inclusion and exclusion criteria, data were pooled to obtain treatment arms with 84 randomized patients and 87 randomized patients for teprotumumab and placebo, respectively.

Literature Review for IVMP

A literature review was conducted to identify existing published literature assessing the most commonly recommended dose of IVMP among patients with moderate to severe active TED.[9] PubMed and Embase were searched for relevant RCTs and observational studies from database inception to date of search (October 5, 2020) using a search strategy that included key terms and controlled vocabulary (eg, "intravenous steroid," "Graves' orbitopathy," "thyroid eye disease," "Graves' ophthalmopathy") (search strategy presented in eAppendix 1 in the Supplement). Results were filtered to include only studies conducted in humans. Regular alerts were established to capture any recent studies until April 1, 2021.

Screening and Selection Criteria

Study inclusion was based on PICOS (population, intervention, comparator, outcomes, and study design) criteria established a priori. Briefly, only studies including patients with moderate to severe active TED receiving treatment with IVMP at a dosage of 4.5 g to 5 g over 12 weeks and reporting at least 1 of the 2 outcomes of interest (ie, change from baseline in proptosis in millimeters and/or Bahn-Gorman diplopia score) were included.[10] Two reviewers (R.A.Q. and R.B.) independently reviewed each title and abstract to identify eligible studies. Full texts of eligible studies were also examined for inclusion criteria and then reviewed to catalog the results.

Data Extraction

Data were extracted by a single reviewer (R.A.Q.) and verified for accuracy by a second reviewer (R.B.). Data extraction was completed using a standardized form and included study characteristics (eg, authors, study design), eligibility criteria (ie, inclusion and exclusion criteria), patient baseline characteristics (eg, sample sizes, sex, age, smoking status), and trial outcomes (eg, change from baseline in proptosis).

jamaophthalmology.com — JAMA Ophthalmology April 2022 Volume 140, Number 4 329

Downloaded From: https://jamanetwork.com/ on 05/31/2022

Limited Context Window

8 Relevant Chunks

OpenAI gpt-3.5-Turbo

System Message
(Controls Model Behavior)

Instructions for Question Answering
(Structure, source citations, comprehensiveness, etc..)

User Question

Context
(Only 6 Relevant Chunks)

Answer with Source Citations

Context Window 16K

Multiple tasks to perform in the backend to generate response for any query

Massive Multi-task Language Understanding - MMLU

- MMLU benchmark evaluates AI models' multitasking accuracy in various tasks. It helps assess AI performance in tasks from simple math to complex legal reasoning.
- Higher scores indicate better multitasking performance

Model	GPT–3.5-Turbo	GPT-4-Turbo
MMLU Score	70.0 (5-shot)	80.4 to 86.4 (-)

Multiple tasks and Instruction for Question Answering

Understand the user's question	Check if the context has the needed information	If information is missing, Immediately reply with No context found	Extract relevant info and its source	Construct a focused answer with citations	Use headings, sub-headings, and bullets	Apply HTML tags for organization
Cite sources using [${{number}}]	Ensure every sentence has a citation	Cite only the most pertinent results	Provide distinct answers for different entities	Format multiple citations as [${{number1}}] [${{number2}}]	Don't answer if context lacks info	Don’t provide answers without context.

Agenda

User feedback Consolidation

Causal Analysis on constructive feedback

Proposed solution

Refinement Plan

Proposed Solution

LLM Model Comparison and Model Upgrade plan

- Comprehensive Model Comparison: A detailed comparison of gpt-3.5-turbo, gpt-4-turbo, gpt-4, and gpt-4-32k models.
- Latency and Cost Optimization Plan: Strategies for optimizing latency and cost during the model upgrade process.

Chat suggestions for Iterative Process Improvement

- Follow-up Question Suggestions: Companion suggests follow-up questions to enhance user experience and foster an educational, iterative process.

Incorporating Images in Responses

- Image Parsing and Tagging: Parsing images from PDFs and tagging them to the appropriate context/chunk.
- Displaying Relevant Images: Displaying images when the context is referenced in the response.

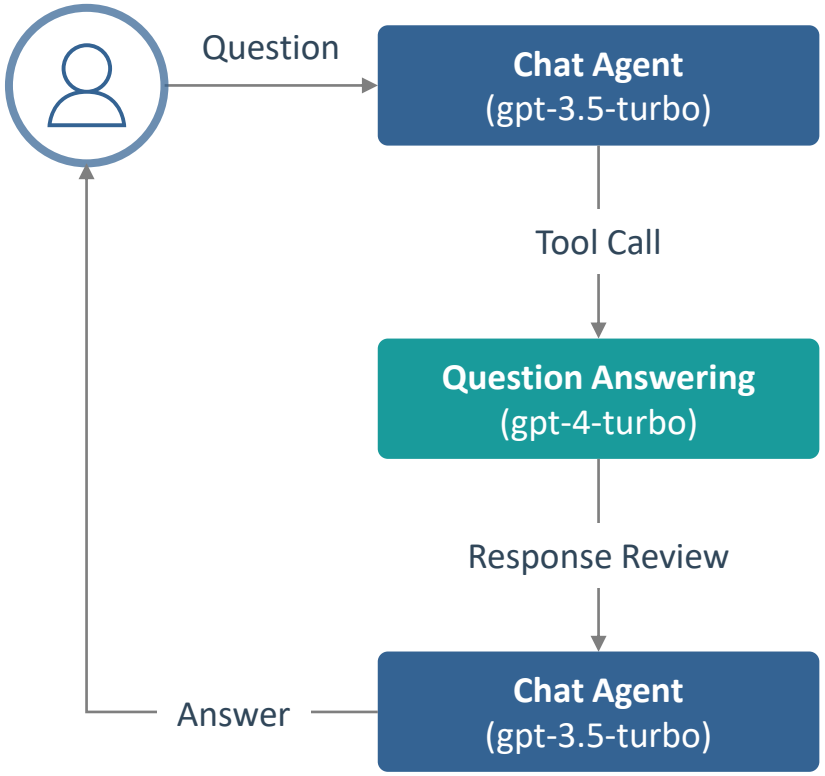
LLM Model Comparison and Model Upgrade plan

Proposed Option

	GPT-3.5-Turbo	GPT-4-Turbo	GPT-4	GPT-4-32k
Input Context Window	16,385	128,000	8,192	32,768
Max Output Tokens	4,096	4,096	4,096	4,096
Pricing – Input Tokens	$0.50 / 1M tokens	$10.0 / 1M tokens	$30.0 / 1M tokens	$60.0 / 1M tokens
Pricing – Output Tokens	$1.50 / 1M tokens	$30.0 / 1M tokens	$60.0 / 1M tokens	$120.0 / 1M tokens
Estimated Latency	15-25 Sec	45-60 Sec	> 1 min	> 1.5 mins
Capability	Moderate	Highest (estimated)	High	High
GPQA – Graduate level Google-Proof Q&A	~ 28	46.5	~ 35	~ 35
HellaSwag – Commonsense reasoning	85.5	96.0	95.3	95.3
MMLU – Multi tasking Language understanding	70.0 (5 shot)	80.4 to 86.4 (-)	86.4 (5 shot)	86.4 (5 shot)

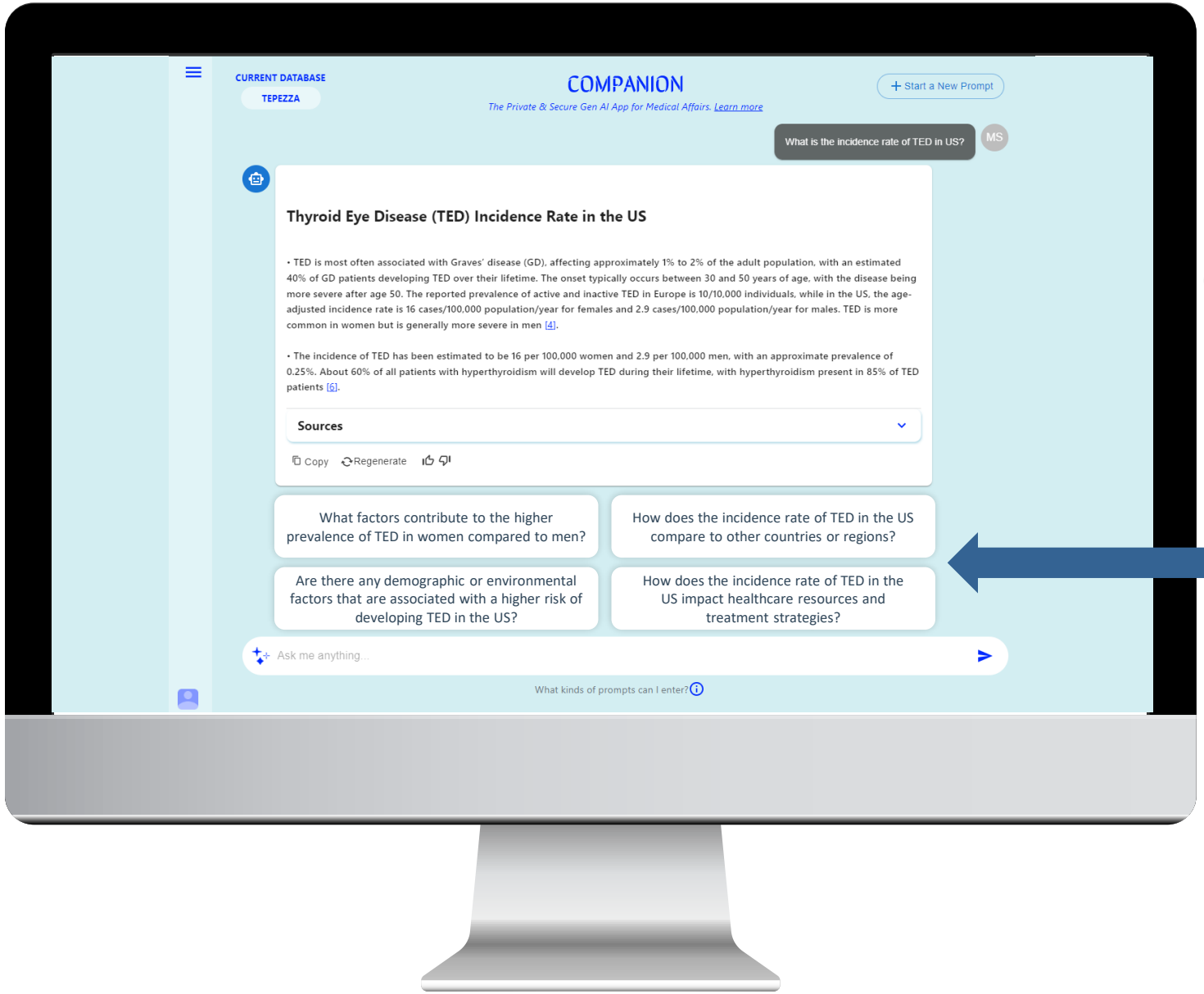
Latency and Cost Optimization Plan

Using advanced models only where necessary



Displaying live status to improve user engagement

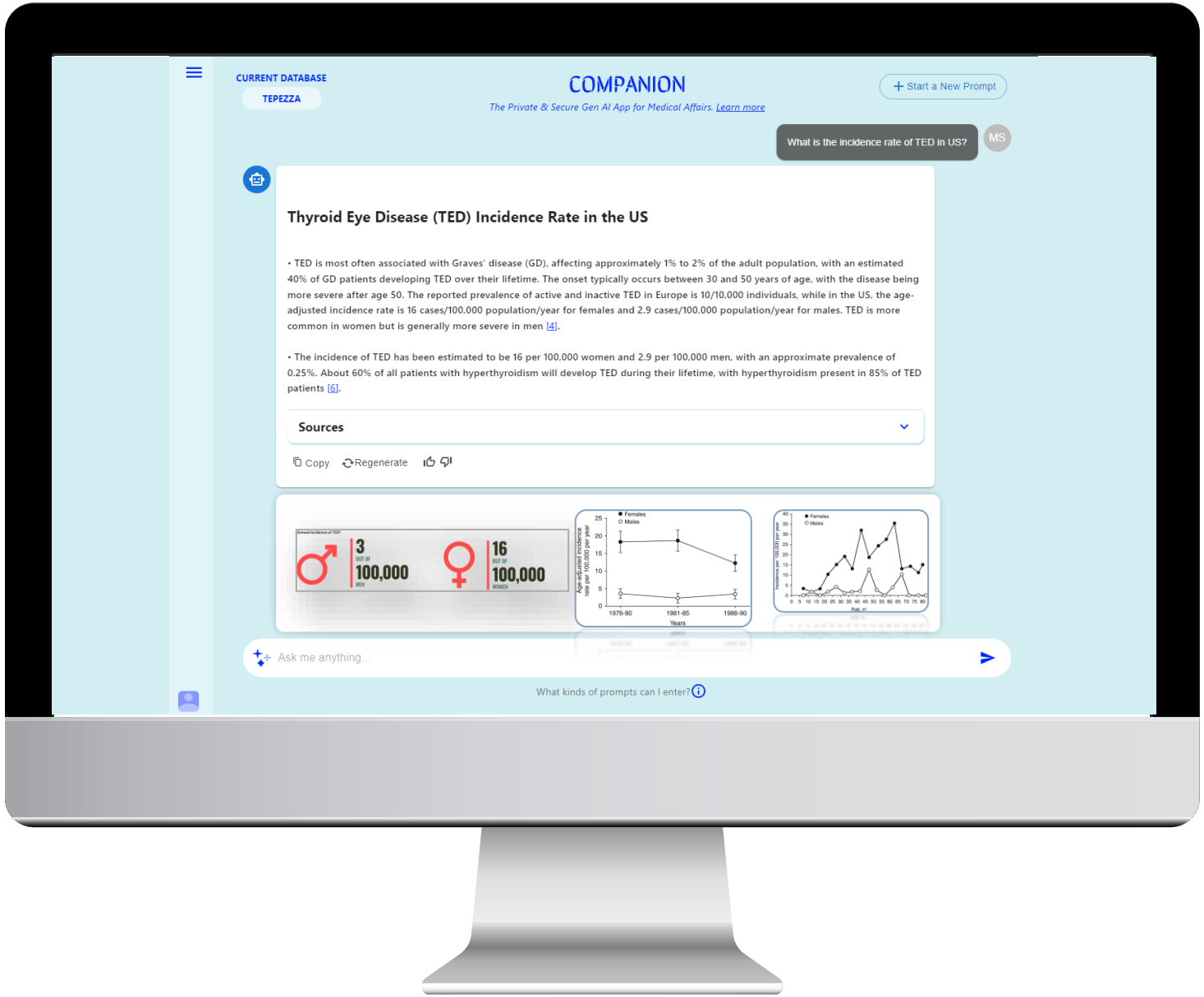
Chat suggestions for enhancing Iterative Process



Adding suggestions to deep dive into topics, ask for more specific information and follow up questions to enable educational and iterative process

Can we add Chat suggestion for repeated process

Incorporating Images in Responses



Can we Incorporate Images and Graphs in the Response