

# **Zeteo Health Assistant: RAG-based Chatbot with Hyper-Personalization**

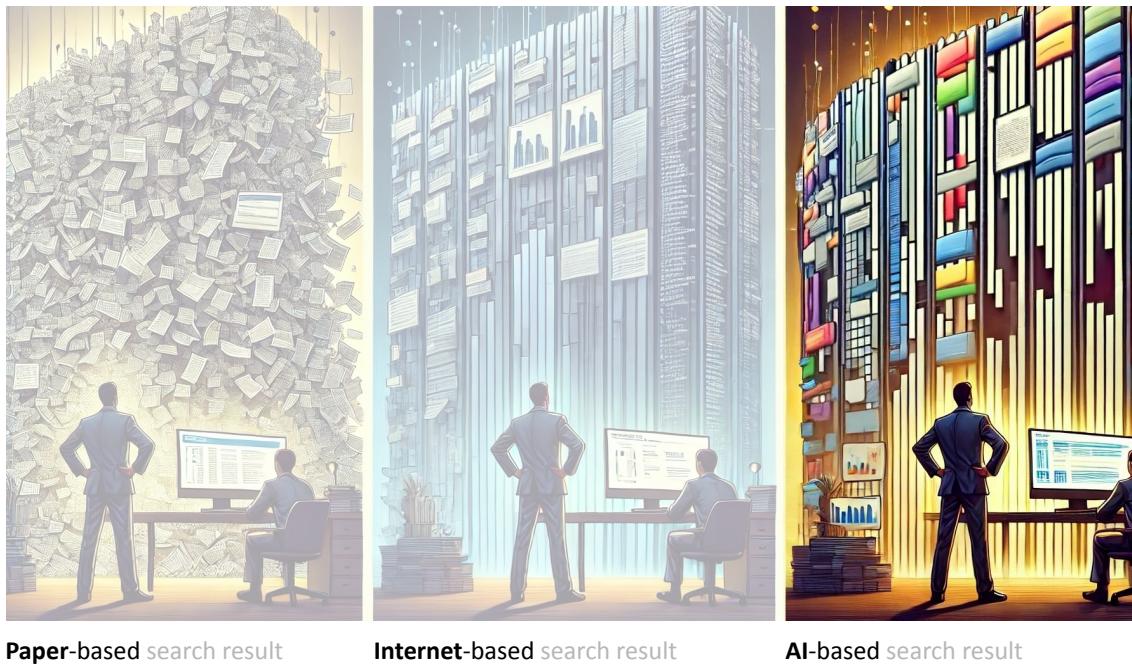
BIS 686 Capstone Project Presentation

Tom Shin

Mentored by Eric Landry

# Introduction *What & Why*

Discreet selection enhanced by Hyper-Personalization



**Paper-based** search result

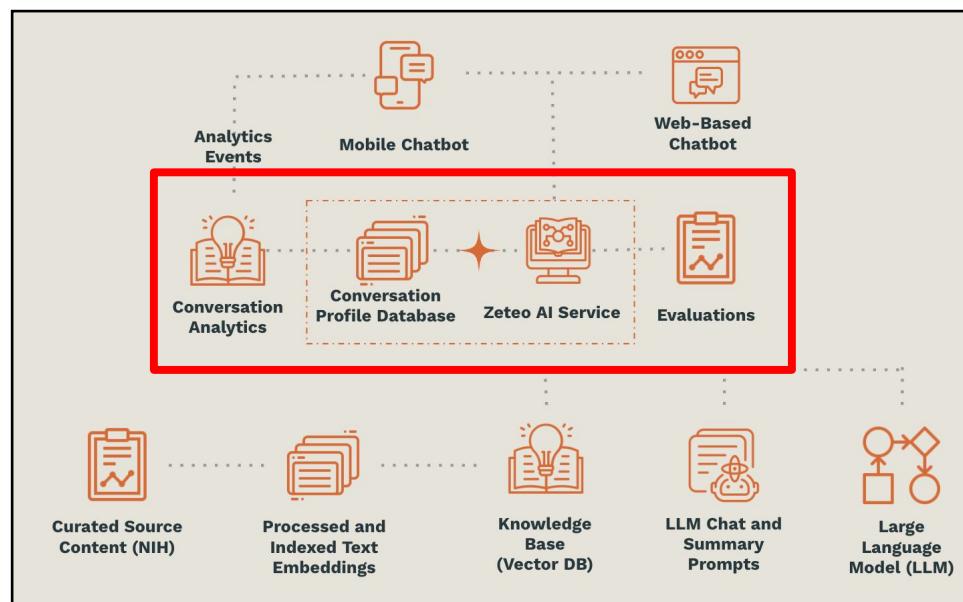
**Internet-based** search result

**AI-based** search result

Geoffrey Hinton, a deep learning pioneer, noted that LLMs, like human memory, "fabricate details"

Need to tailor AI-driven healthcare information to users' specific contexts

# Introduction *Key Focus*



The Zeteo Health conversational AI framework uses a Retrieval-Augmented Generation (RAG) Large Language Model (LLM) architecture, powered by a professionally curated knowledge base to deliver context-aware, personalized responses

# Scoping *Multi-turn Conversation*

The screenshot shows a conversational interface between an **Assistant** and a **User**.

**Assistant:**

- According to our information, some key statistics on prostate cancer risk factors and prevalence include:
  - Black men in the US and the Caribbean have the highest documented prostate cancer incidence rates in the world.
  - Increasing age is a well-established risk factor for prostate cancer.
  - A family history of the disease and certain inherited genetic conditions like Lynch syndrome, BRCA1 and BRCA2 mutations also increase risk.

**User:**

- True true. Could you provide me with some relevant statistics?

**Assistant:**

- According to our information, some key statistics on prostate cancer risk factors and prevalence include:
  - Black men in the US and the Caribbean have the highest documented prostate cancer incidence rates in the world.
  - Increasing age is a well-established risk factor for prostate cancer.
  - A family history of the disease and certain inherited genetic conditions like Lynch syndrome, BRCA1 and BRCA2 mutations also increase risk.

**User:**

- Alright thanks

**Buttons:**

- VIEW SUMMARY

As LLMs evolve beyond simple Q&A to power ubiquitous conversational platforms,  
the focus shifts to managing continuous, nuanced, and flowing dialogues

# Scoping Defining Features by Level of Interaction

**Single utterance**

True true. Could you provide me with some relevant statistics?

**Multi-turn conversation**

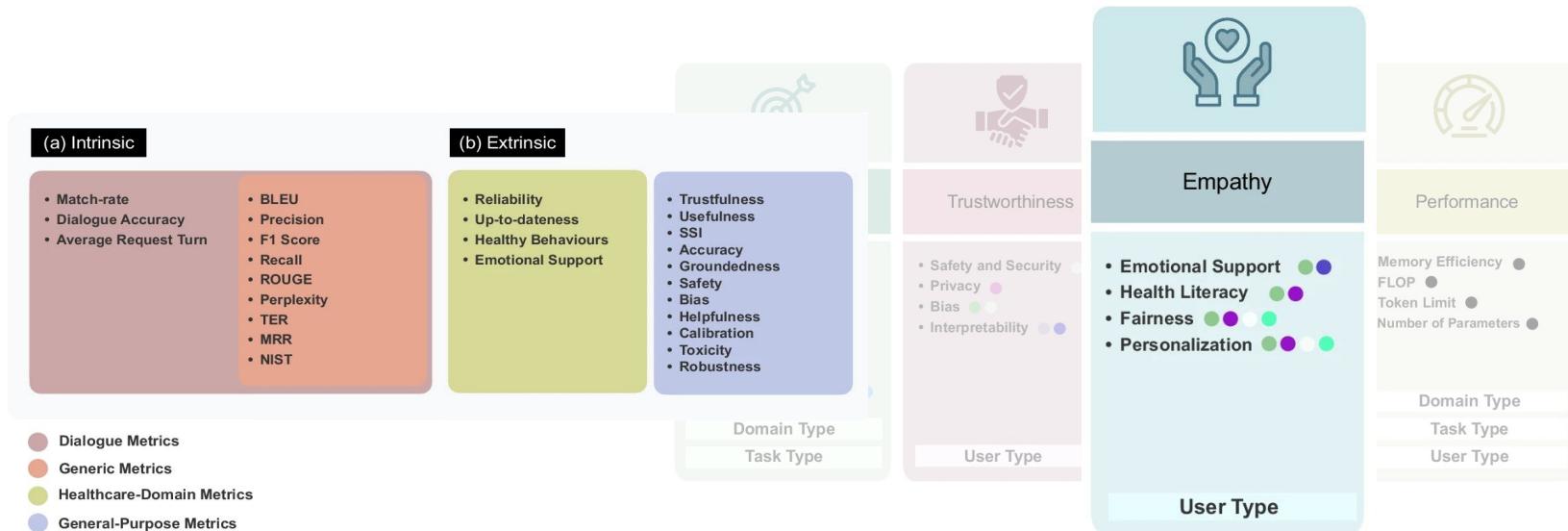
True true. Could you provide me with some relevant statistics?  
According to our information, some key statistics on prostate cancer risk factors and prevalence include:  
- Black men in the US and the Caribbean have the highest documented prostate cancer incidence rates in the world.  
- Increasing age is a well-established risk factor for prostate cancer.  
- A family history of the disease and certain inherited genetic conditions like Lynch syndrome, BRCA1 and BRCA2 mutations also increase risk.

**History**

True true. Could you provide me with some relevant statistics?  
According to our information, some key statistics on prostate cancer risk factors and prevalence include:  
- Black men in the US and the Caribbean have the highest documented prostate cancer incidence rates in the world.  
- Increasing age is a well-established risk factor for prostate cancer.  
- A family history of the disease and certain inherited genetic conditions like Lynch syndrome, BRCA1 and BRCA2 mutations also increase risk.

LLM evaluations should go beyond accuracy and precision, emphasizing how well contextual nuances are integrated across multi-turn conversations for coherent and adaptive dialogue

# Scoping Existing Metrics



- **Intrinsic** metrics (e.g., BLEU, ROUGE) focus on language, not medical concepts or patient well-being
- **Extrinsic** methods (e.g., human judgment) are narrow, missing holistic healthcare needs

Overlook empathy, trust-building, personalization, and emotional support<sup>[2]</sup>

# Scoping Multi-turn Conversation

## Assistant

### Features from generated output Xs

The screenshot shows a conversational interface with two main components: the Assistant and the User.

**Assistant:** The left side of the interface. It features a sidebar with a blue gradient bar and a "SEARCH" button. Below the sidebar, there's a "Recent" section with a "View all" link and a "New messages" section with a "View all" link. A "Logout" button is at the bottom.

**User:** The right side of the interface. It shows a list of messages in a chat-like format. The messages are color-coded: light gray for the AI and teal for the user. The AI's messages include:

- "According to our information, some key statistics on prostate cancer risk factors and prevalence include:
  - Black men in the US and the Caribbean have the highest documented prostate cancer incidence rates in the world.
  - Increasing age is a well-established risk factor for prostate cancer.
  - A family history of the disease and certain inherited genetic conditions like Lynch syndrome, BRCA1 and BRCA2 mutations also increase risk.
- "True true. Could you provide me with some relevant statistics?"
- "According to our information, some key statistics on prostate cancer risk factors and prevalence include:
  - Black men in the US and the Caribbean have the highest documented prostate cancer incidence rates in the world.
  - Increasing age is a well-established risk factor for prostate cancer.
  - A family history of the disease and certain inherited genetic conditions like Lynch syndrome, BRCA1 and BRCA2 mutations also increase risk.

The user's messages are:

- "Alright thanks"

**Satisfaction as destination Y:** A teal arrow pointing to the user's message "Alright thanks".

Investigate healthcare assistant output components to gauge user satisfaction, laying groundwork for an evaluation framework, despite challenges in finding multi-conversation data with clear satisfaction labels

## Data Reddit “/AskDocs”

r/AskDocs • 7 yr. ago  
WhiteTwink

I was hit in the chest with a 20 Kilowatt blast of EM Radiation from a communications dish and felt like I was hit in the chest with a baseball bat, what did I actually feel?

Basically I was cleaning a long range communications dish from a system with about 20 Kilowatts of power and someone accidentally turned on the transmitter. I felt like I hit in the chest with a baseball bat and was pushed off my feet onto my back (or maybe I fell from the pain, not sure).

What did I actually feel since em radiation doesn't have any mass? How many years off my life have I lost because of that?

Age at Time, 20  
Sex, Male

Beeroy99 • 7y ago

LAWYER UP, clear case of gross misconduct and negligence. Yet I have no idea. Have you got some technical info on the exact dish?

WhiteTwink OP • 7y ago

Psh the army is immune to lawsuits

Also I do but I'm not sure how much I'm aloud to say, basically it was 20 Kilowatts microwave communications device, single dish

123 Award Share ...  
10 more replies

Prior studies utilized movie ratings to assess subjective conversational quality<sup>[3]</sup>

→ Evaluate user need fulfillment through feature extraction from iterative medical domain conversations

# Methods *Goals*

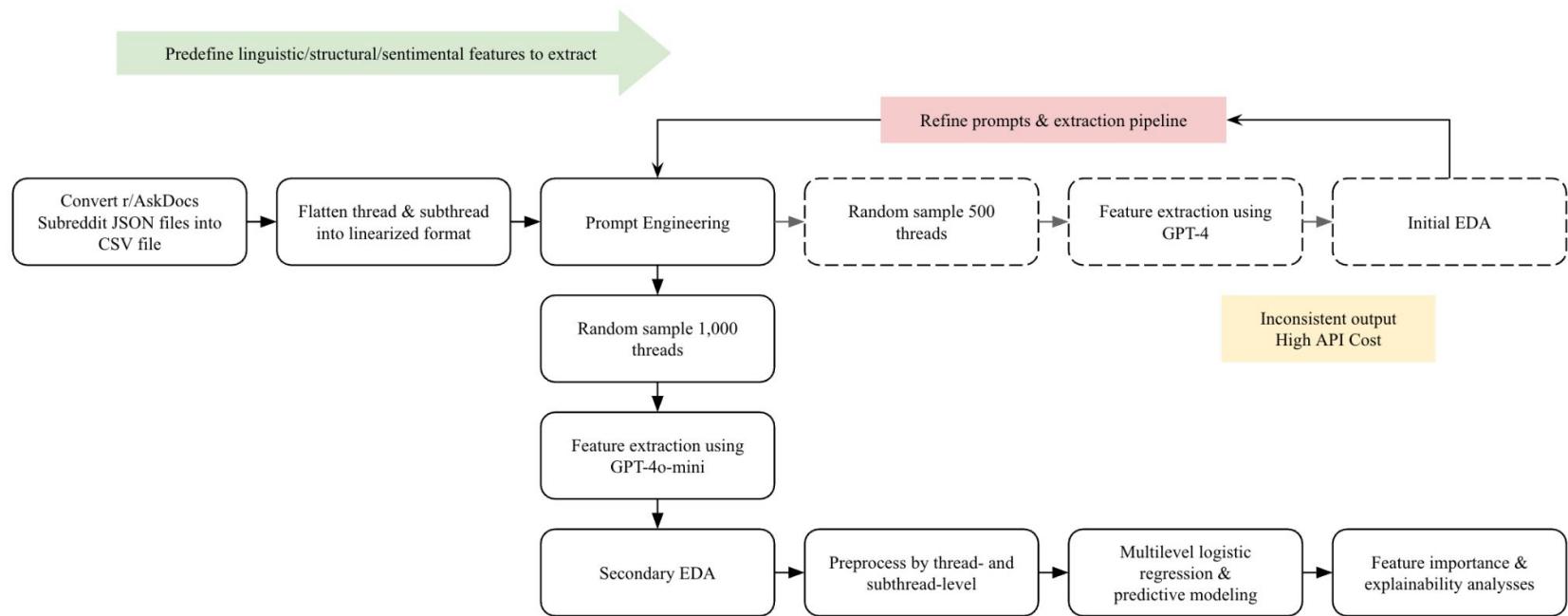
## 1) Benchmarking Response Quality

Analyze successful multi-turn conversations (e.g., Reddit's "resolved, satisfied" threads) to define quantitative (readability, terminology) and qualitative (credible sources, expert quotes) standards for coherent, user-focused AI responses

## 2) Feature Extraction for Adaptive Analysis

Identify recurring patterns (e.g., follow-up questions, user intent) in multi-turn dialogues to guide real-time analytics and prompting strategies for consistent, meaningful, and storable AI outputs

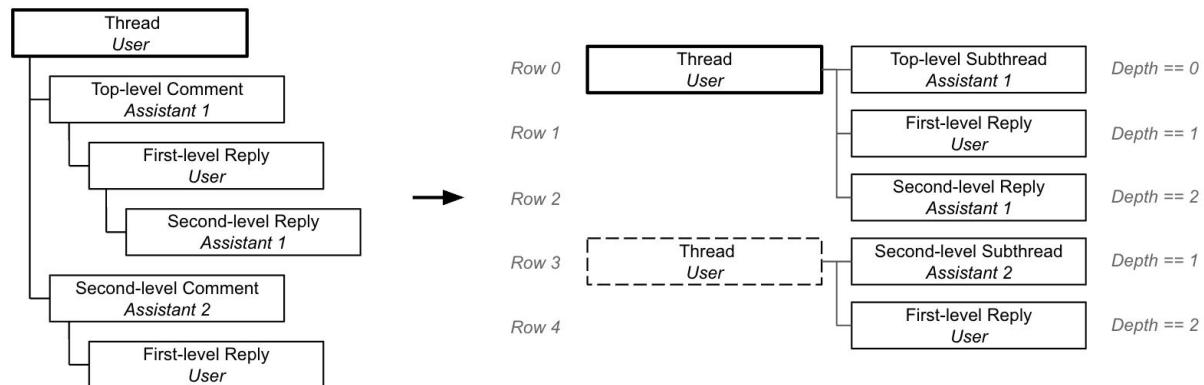
# Methods Workflow Diagram



## Methods *Project Design*

- 1) Collect/transform/preprocess multi-turn conversational data with LLM-labeled outcomes: user satisfaction/inquiry resolution
- 2) Identify and extract features
- 3) Conduct exploratory and association analyses to identify trends in medical concerns and user needs to guide
- 4) Perform multilevel logistic regression to account for conversation-level variation
- 5) Apply predictive modeling with explainability to identify key features for improved input collection

## Data Conversion



Transformed Reddit's thread-like conversation data into a linear table/dataframe, structuring user threads with root posts branching into subthreads, each representing individual user-assistant conversations

## Data Conversion

The screenshot shows a Reddit post from r/AskDocs and its replies, with a data conversion table overlaid.

**Post:**

r/AskDocs · 7 yr. ago  
I was hit in the chest with a 20 Kilowatt blast of EM Radiation from a communications dish and felt like I was hit in the chest with a baseball bat, what did I actually feel?

Basically I was cleaning a long range communications dish from a system with about 20 Kilowatts of power and someone accidentally turned on the transmitter. I felt like I hit in the chest with a baseball bat and was pushed off my feet onto my back (or maybe I fell from the pain, not sure).

What did I actually feel since em radiation doesn't have any mass? How many years off my life have I lost because of that?

Age at Time, 20  
Sex, Male

**Replies:**

Beeroy69 · 7y ago  
LAWYER UP, clear case of gross misconduct and negligence. Yet I have no idea. Have you got some technical info on the exact dish?

Psh the army is immune to lawsuits

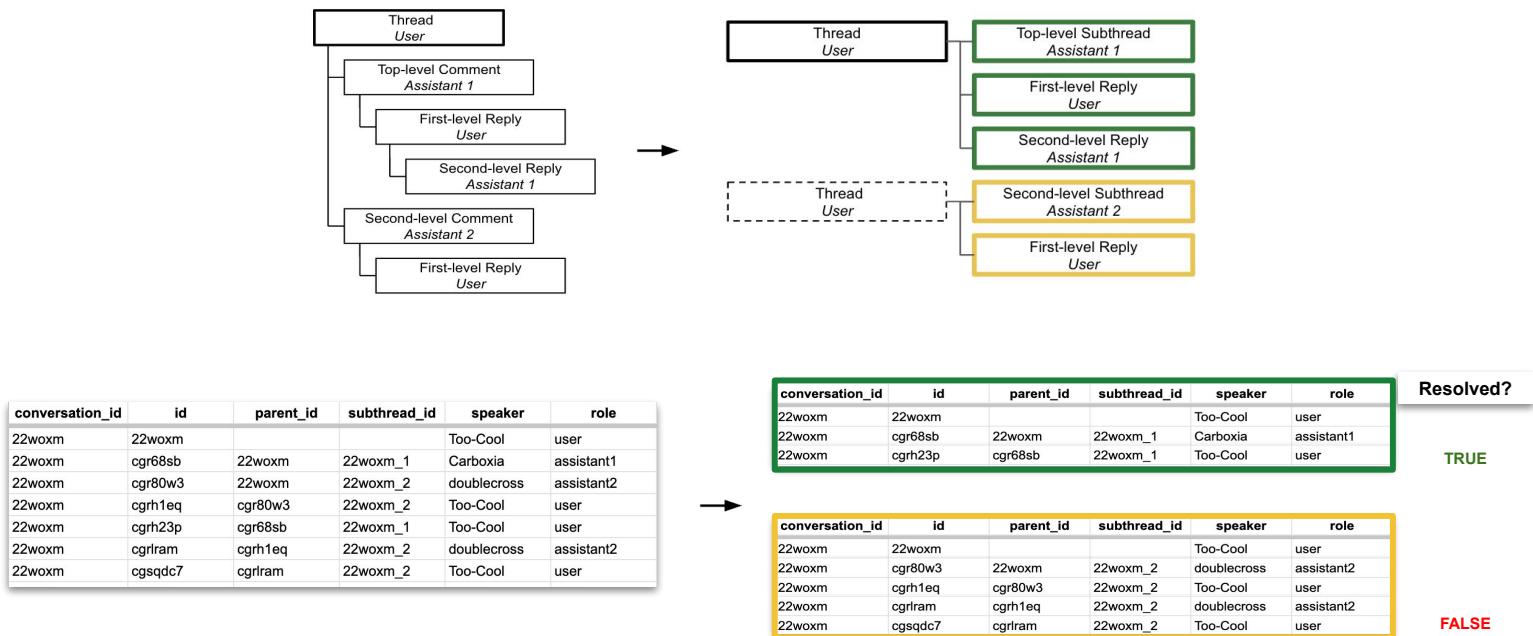
Also I do but I'm not sure how much I'm aloud to say, basically it was 20 Kilowatts microwave communications device, single dish

123 Award Share ...  
10 more replies

**Data Conversion Table:**

convid	created	role	content	id	reply_to	score	title
9opeep	2018-10-16 16:46:49	user	Basically I was cleaning a long range communic...	9opeep	NaN	158.0	I was hit in the chest with a 20 Kilowatt blas...
9opeep	2018-10-16 17:11:20	assistant1	LAWYER UP, clear case of gross misconduct and ...	e7vpd8r	9opeep	217.0	I was hit in the chest with a 20 Kilowatt blas...
9opeep	2018-10-16 17:12:48	user	Psh the army is immune to lawsuits \n\nAlso I ...	e7vphdj	e7vpd8r	113.0	I was hit in the chest with a 20 Kilowatt blas...
9opeep	2018-10-16 17:23:19	assistant2	I'm not a physicist or a doctor of the right t...	e7vqalr	9opeep	28.0	I was hit in the chest with a 20 Kilowatt blas...
9opeep	2018-10-16 17:26:55	assistant3	I'd Google it, or ask some people in the physi...	e7vqkku	e7vphdj	9.0	I was hit in the chest with a 20 Kilowatt blas...

## Data Conversion



Flattened thread format into linear conversation format

## Data Feature Extraction

The screenshot shows a user interface for processing a patient's叙述. On the left, a text input field contains a story about ibuprofen side effects. A 'Compute' button is visible below the text. The right side displays two code snippets for generating prompts for an LLM.

**component\_prompt:**

```
component_prompt = f"""Identify all structural components in this text (multiple can apply):  
- Numeric Fact, Link, Quote, Anecdote, Personal Experience, List, Paragraph.  
Return as a comma-separated list (e.g., 'Numeric Fact, Paragraph').  
Text: {t}"""
```

**resolution\_prompt:**

```
resolution_prompt = f"""Does this thread resolve the initial concern from the depth=0 post?  
- 'True' or 'False'.  
Initial text: {initial_text}  
Full thread: {full_thread}"""
```

Instruction-tuned LLM models *LLaMA & GPT* show strong generalization across tasks, highlighting their potential for clinical NLP<sup>[4]</sup>

→ Motivated from the approach, used LLM *GPT-4o-mini & few-shot prompting* for flexibility and generalizability

## Data First Extraction

```
resolution_prompt = f"""Does this thread resolve the initial concern from the depth=0 post?  
- 'True' or 'False'.  
Initial text: {initial_text}  
Full thread: {full_thread}"""
```

Resolution

**TRUE**

**text\_clean**

I actually don't usually eat foods like this, but I guess I could do with drinking more water and eating more fiber. I'll try this, thank you!

Resolution

**FALSE**

**text\_clean**

See, that's part of what puzzled me. It is the left lung that doesn't get sick, even though that was the injured one. The neurological damage on the left affecting the right side does make sense, but how it does is the question. I won't treat this like an official diagnosis or anything, but it is interesting to think about.

```
component_prompt = f"""Identify all structural components in this text (multiple can apply):  
- Numeric Fact, Link, Quote, Anecdote, Personal Experience, List, Paragraph.  
Return as a comma-separated list (e.g., 'Numeric Fact, Paragraph').  
Text: {t}"""
```

Top 20 Values in component\_structure Column

	Component_structure	Count	Percentage
0	Paragraph	2245	33.06%
1	Personal Experience, Paragraph	1099	16.18%
2	Numeric Fact, Paragraph	721	10.62%
3	Paragraph, Personal Experience	430	6.33%
...			
17	Paragraph, List	38	0.56%
18	Anecdote, Paragraph	27	0.40%
19	Personal Experience, Numeric Fact, Paragraph	24	0.35%

Initial extraction using general prompting without task-specific fine-tuned LLM yielded inconsistent results *right* with acceptable resolution *left* but numerous dataset inconsistencies found during manual review

## Data Refined Extraction

```
resolution_prompt = f"""Analyze a medical conversation subthread to determine if the user's initial concern, expressed in the first post, has been resolved. Resolution requires at least one user message in the subthread to explicitly indicate satisfaction or acknowledgment of an answer (e.g., gratitude like 'Thanks', 'That helps', or agreement like 'Got it', 'This makes sense').
```

Select 'True' if any user text shows resolution, or 'False' if no user message indicates resolution or if the user expresses ongoing confusion or dissatisfaction (e.g., 'I'm still not sure', 'That doesn't help'). Assistant responses alone (e.g., providing advice) are insufficient unless a user confirms resolution.

Return the result in JSON format. Consider the initial post and the full subthread context, which includes all messages labeled by role. Do not provide explanations or additional text.

Choices:

- True: At least one user message explicitly indicates resolution
- False: No user message indicates resolution, or user expresses dissatisfaction/confusion

Initial User Post (depth=0): {initial\_query}

Subthread Context (all messages with roles): {subthread\_context}

Return: {"resolution": <true/false>}"""

role	text	resolution	thankfulness	sentiment_score	vibe
user	...	FALSE	1	-0.8437	0
assistant1	...	TRUE		-0.5106	0
user	...	TRUE	1	0.1979	0
assistant1	...	TRUE		0.1511	0
user	...	TRUE	1	0.8519	1
user	...	FALSE	1	-0.0673	0
assistant1	...	TRUE		0.7096	1
user	...	TRUE	2	0.74	0
assistant1	...	TRUE		0.0258	1
user	...	TRUE	1	0.7054	1

```
combined_prompt = f"""Analyze the following medical conversation utterance and extract the specified binary features in a single JSON object. Each feature should indicate whether the described content is present (true) or not (false). Do not provide explanations or extra text beyond the JSON object.
```

\*\*Text\*\*: {t}

\*\*Features to Extract\*\*:

1. \*\*is\_website\_link\_present\*\*: Contains a website link (e.g., URLs starting with http, https, www).

Return true if present, false if not.

2. \*\*is\_numeric\_statistical\_fact\_present\*\*: Contains a numeric or statistical fact, such as percentages, ratios, or counts (e.g., "50% of patients recover," "3 out of 4 cases").

Return true if present, false if not.

Return: {"is\_website\_link\_present": <true/false>, "is\_numeric\_statistical\_fact\_present": <true/false>,

role	website_link	credible_source	num_stat_fact	anecdote	expert_quote	...	resolution
user	FALSE	FALSE	TRUE	TRUE	FALSE	...	FALSE
assistant1	FALSE	FALSE	FALSE	FALSE	TRUE	...	FALSE
user	FALSE	FALSE	FALSE	TRUE	FALSE	...	FALSE
assistant1	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE
user	FALSE	FALSE	TRUE	TRUE	FALSE	...	FALSE
assistant2	FALSE	FALSE	TRUE	TRUE	FALSE	...	FALSE
user	FALSE	FALSE	FALSE	TRUE	FALSE	...	FALSE

Optimized prompt and extraction pipeline by parallelizing extraction, expanding the prompt's context window, and adopting MCQ format for the LLM, maintaining performance while reducing computation costs and enhancing coherence and precision<sup>[5]</sup>

## Data Refined Extraction

role	text	resolution	thankfulness	sentiment_score	vibe
user	...	FALSE	1	-0.8437	0
assistant1	...	TRUE		-0.5106	0
user	...	TRUE	1	0.1979	0
assistant1	...	TRUE		0.1511	0
user	...	TRUE	1	0.8519	1
user	...	FALSE	1	-0.0673	0
assistant1	...	TRUE		0.7096	1
user	...	TRUE	2	0.74	0
assistant1	...	TRUE		0.0258	1
user	...	TRUE	1	0.7054	1

resolution\_prompt = f"""Analyze a medical conversation subthread to determine if the user's initial concern, expressed in the first post, has been resolved. Resolution requires at least one user message in the subthread to explicitly indicate satisfaction or acknowledgment of an answer (e.g., gratitude like 'Thanks', 'That helps', or agreement like 'Got it', 'This makes sense').

Select 'True' if any user text shows resolution, or 'False' if no user message indicates resolution or if the user expresses ongoing confusion or dissatisfaction (e.g., 'I'm still not sure', 'That doesn't help'). Assistant responses alone (e.g., providing advice) are insufficient unless a user confirms resolution.

Return the result in JSON format. Consider the initial post and the full subthread context, which includes all messages labeled by role. Do not provide explanations or additional text.

Choices:

- True: At least one user message explicitly indicates resolution
- False: No user message indicates resolution, or user expresses dissatisfaction/confusion

Initial User Post (depth=0): {initial\_query}

Subthread Context (all messages with roles): {subthread\_context}

Return: {"resolution": <true/false>}"""

thankfulness\_prompt = f"""Analyze the user reply in a medical conversation to assess the level of thankfulness expressed toward the assistant's responses. Select the single most appropriate thankfulness level from the choices and return its numerical value in JSON format. Use these guidelines:

- 0: No note (no gratitude, neutral or unrelated, e.g., "Not really", "I don't think so")
- 1: Somewhat thankful (mild gratitude, e.g., "Um okay thanks", "Thanks for the info")
- 2: Very thankful (strong gratitude, e.g., "Thank you so much, this really helped!", "I will definitely try this. Thanks!")

Consider the conversation context to interpret the tone and intent. Do not provide explanations or additional text.

Choices:

- 0: No note
- 1: Somewhat thankful
- 2: Very thankful

Conversation Context (recent messages): {history}

User Reply: {t}

Return: {"thankfulness": <number>}"""

Manually reviewed LLM **resolution** with VADER **sentiment** trajectory and three-tier **thankfulness** label, middle label cushions for polite appreciation of assistant effort, not true satisfaction to validate true satisfaction

## Data Features Extracted [ALL]

Feature	Description	Value	Extraction Level	Example	LLM Used
<b>Metadata</b>					
<i>Raw-Reddit</i>					
conversation_id	Unique thread ID	String	Metadata	1xv9vn	✗
id	Post ID	String	Metadata	1xv9vn	✗
reply_to	changed into parent_id	String/None	Metadata	None	✗
speaker	Poster ID	String	Metadata	healthissue	✗
title	Thread title	String	Metadata	HELP NEEDED	✗
text	Raw post text	String	Row	I have pain... I have pain... I have pain...	✗
timestamp	Post timestamp	Integer	Metadata	1392350217	✗
score	Post rating	Integer	Row	1	✗
<i>Processed</i>					
subthread_id	Subthread ID	String/None	Metadata	1xv9vn_1	✗
parent_id	Parent post ID	String/None	Metadata	None	✗
text_clean	Cleaned text	String	Row	I have pain...	✗
role	Speaker role	user/assistantN	Metadata	user	✗
depth	Post level	Integer	Metadata	0	✗
max_depth	Max thread depth	Integer	Row	4	✗
max_count	Thread post count	Integer	Row	5	✗

Feature	Description	Value	Extraction Level	Example	LLM Used
<b>Extracted Structure</b>					
length_of_text	Word count	Integer	Row	5	✗
readability	Text readability	Float	Row	5.0	✗
<b>Conversation</b>					
primary_intent	Post intent	1-11	Row	6	✓
question_type	Question type	1-6	Row	1	✓
specific_disease	Disease name	String/N/A	Thread	N/A	✓
body_area	Affected body area	0-9	Row	3	✓
discomfort_type	Symptom type	0-13	Row	1	✓
<b>Component</b>					
has_question	Contains question	Boolean	Row	True	✓
is_website_link	Has website link	Boolean	Row	False	✗
is_numeric_fact	Has numeric fact	Boolean	Row	False	✗
is_personal_exp	Has personal story	Boolean	Row	True	✗
is_expert_quote	Quotes expert	Boolean	Row	False	✗
is_credible_source	Has credible source	Boolean	Row	False	✗
medical_terms	Medical term count	Integer	Row	1	✓
medical_expert	Shows expertise	1/0	Row	0	✓
<b>Sentiment</b>					
urgency	Urgency level	Non-urgent to Emergent	Row	Moderate	✓
sentiment_score	Text sentiment	Float (-1 to 1)	Row	-0.5	✗
vibe	Post tone	1/0/-1	Row	-1	✓
is_anxious	Shows anxiety	1/0	Row	1	✓
is_empathetic	Expresses empathy	Boolean	Row	False	✓
is_confident	Expresses confidence	Boolean	Row	False	✓
<b>Destination</b>					
resolution	Concern resolved	Boolean	Thread	False	✓
thankfulness	User gratitude	0-2/None	Row	1	✓

Table 1: Grouped Feature Descriptions, Examples, and Analyzed Viewpoint

Lists features from 1,000 randomly sampled conversation IDs in the *AskDocs* subreddit, including Type (Metadata, Row-based, or Thread-level extraction), and LLM Used (indicating whether an LLM was used for extraction)

## Data Categorical Feature Labels

Feature	Schema
primary_intent	<p><i>Depth=0:</i></p> <ul style="list-style-type: none"> <li>1: Product recommendation</li> <li>2: Diagnosis</li> <li>3: Fact check</li> <li>4: Console</li> <li>5: Lifestyle</li> <li>6: Disease management</li> <li>7: Careprovider recommendation</li> </ul> <p><i>Depth!=0:</i></p> <ul style="list-style-type: none"> <li>8: Question</li> <li>9: Answer</li> <li>10: Feedback</li> <li>11: Follow up</li> </ul>
body_area	<ul style="list-style-type: none"> <li>0: N/A</li> <li>1: Head</li> <li>2: Chest</li> <li>3: Stomach</li> <li>4: Right arm</li> <li>5: Left arm</li> <li>6: Right leg</li> <li>7: Left leg</li> <li>8: Back</li> <li>9: Neck</li> </ul>
discomfort.type	<ul style="list-style-type: none"> <li>0: N/A</li> <li>1: Pain</li> <li>2: Rash</li> <li>3: Anxiety</li> <li>4: Fatigue</li> <li>5: Dizziness</li> <li>6: Itching</li> <li>7: Nausea</li> <li>8: Sore throat</li> <li>9: Discomfort</li> <li>10: Swelling</li> <li>11: Blood in stool</li> <li>12: Hair loss</li> <li>13: Numbness</li> </ul>
urgency	<ul style="list-style-type: none"> <li>0: Non-urgent</li> <li>1: Moderate</li> <li>2: Urgent</li> <li>3: Emergent</li> </ul>
question.type	<p><i>User:</i></p> <ul style="list-style-type: none"> <li>1: More information</li> <li>2: Easier interpretation</li> <li>3: Other</li> </ul> <p><i>Assistant:</i></p> <ul style="list-style-type: none"> <li>4: More information</li> <li>5: Easier interpretation</li> <li>6: Other</li> </ul>
vibe	<ul style="list-style-type: none"> <li>1: Positive</li> <li>0: Neutral</li> <li>-1: Negative</li> </ul>
thankfulness	<ul style="list-style-type: none"> <li>0: None</li> <li>1: Somewhat</li> <li>2: Very</li> <li>None</li> </ul>

Table 2: Specific Labeling Schema for Cardinal/Ordinal Features

Schema labels are defined based on the EDA of samples from the first extraction

Yet require a more comprehensive and clinically reviewed list to ensure accuracy and completeness

## Data Old vs. Refined Data Summary

Metric	Value
Unique Thread Posts	500
Total Subthreads	1091
Avg Subthreads per Thread	2.18
Total Messages	3918
User Messages	2079
Assistant Messages	1839
Avg Depth per Subthread	3.56
Conversations Resolved	406
Percentage Resolved (%)	37.21
Resolution: False	488
Resolution: True	406

Table 3: Data Summary Before Refined Prompt

Metric	Value
Unique Thread Posts	1000
Total Subthreads	2179
Avg Subthreads per Thread	2.18
Total Messages	7817
User Messages	4166
Assistant Messages	3651
Avg Depth per Subthread	3.51
Conversations Resolved	993
Percentage Resolved (%)	45.57
Resolution: False	1186
Resolution: True	993

Table 4: Data Summary After Refined Prompt

Expanded sample size from 1,091→**2,179** conversations, drawn from 84,321 unique thread posts (likely more conversations) up to 10/2018

## Data Preprocessing for Modeling

conversation_id	subthread_id	role	website_link	credible_source	num_stat_fact	anecdote	expert_quote	...	resolution
24i1m1		user	FALSE	FALSE	TRUE	TRUE	FALSE	...	FALSE
24i1m1	24i1m1_1	assistant1	FALSE	FALSE	FALSE	FALSE	TRUE	...	FALSE
24i1m1	24i1m1_1	user	FALSE	FALSE	FALSE	TRUE	FALSE	...	FALSE
24i1m1	24i1m1_1	assistant1	FALSE	FALSE	FALSE	FALSE	FALSE	...	FALSE
24i1m1	24i1m1_1	user	FALSE	FALSE	TRUE	TRUE	FALSE	...	FALSE
24i1m1	24i1m1_2	assistant2	FALSE	FALSE	TRUE	TRUE	FALSE	...	FALSE
24i1m1	24i1m1_2	user	FALSE	FALSE	FALSE	TRUE	FALSE	...	FALSE



conversation_id	subthread_id	website_link	credible_source	num_stat_fact	anecdote	expert_quote	...	resolution
24i1m1	24i1m1_1	0	0	0	1	1	...	0
24i1m1	24i1m1_2	0	1	1	0	1	...	0

Feature	Description	Value	Extraction Level	Example	LLM Used
Structure Features					
text_len_bin	Binned word count of post	1/2/3/4	Row	Medium	✗
Numerical Features					
readability_bin	Binned readability score	1/2/3/4	Row	Easy	✗
med_nomen_count	Count of medical terms in post	Integer	Row	2	✓
max_turns	Number of posts in thread	Integer	Subthread	4	✗
assistant_score_sum	Sum of post ratings	Integer	Thread	5	✗
Binary Features					
is.empathetic	Expresses empathy	1/0	Row	1	✓
is.confident	Expresses confidence	1/0	Row	0	✓
website_link	Contains a website link	1/0	Row	0	✓
num_stat_fact	Contains numeric facts	1/0	Row	0	✓
anecdote	Contains personal story	1/0	Row	1	✓
expert_quote	Quotes a medical expert	1/0	Row	0	✓
medical_expert	Indicates medical expertise	1/0	Row	0	✓
Sentiment/Categorical Features					
vibe	Tone of the post	-1/0/1	Row	0	✓
root_urgency	Urgency level of thread	0/1/2/3	Root Post	Moderate	✓
Destination					
resolution	Concern resolved	True/False	Subthread	False	✓

Table 5: Final Features Used in the Full Machine Learning Model (14 Features)

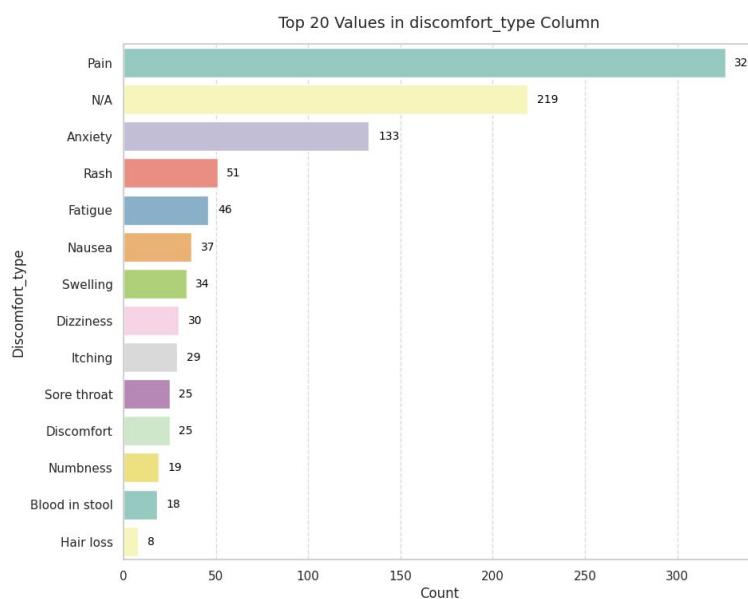
Binary labeling to flatten data, indicating if components are present in assistant responses at the subthread level,

Rather than extracting detailed information given inconsistent extraction & unnecessary

# Result EDA: Discomfort Type (User Initial Thread == Root)

Feature	Schema
discomfort_type	0: N/A 1: Pain 2: Rash 3: Anxiety 4: Fatigue 5: Dizziness 6: Itching 7: Nausea 8: Sore throat 9: Discomfort 10: Swelling 11: Blood in stool 12: Hair loss 13: Numbness

	Discomfort_type	Count	Percentage
0	Pain	326	32.60%
1	N/A	219	21.90%
2	Anxiety	133	13.30%
3	Rash	51	5.10%
4	Fatigue	46	4.60%
5	Nausea	37	3.70%
6	Swelling	34	3.40%
7	Dizziness	30	3.00%



I'm a 20 year old male, 5'10" and about 200 pounds. I went for a walk last night to clear my head. I stayed on the roads around my college campus. I went about 4.3 miles in a little over an hour. I know, it isn't very long, but when I got back I had this pain on the outside edge of my foot and it hurt to walk. Today I can hardly walk without having to lean on something. It doesn't hurt when I'm sitting or when I touch it with my hand, only when I put weight on it. What could this be? Also, I have a few house showings to go to today and tomorrow so what can I do to help the pain?

Pain Utterance Sample

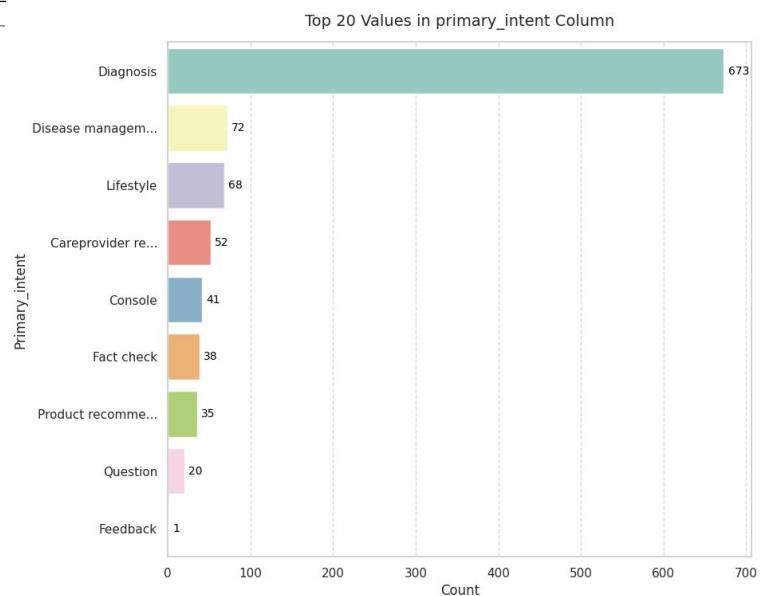
What would the differences be healthwise? I understand that manufactured cigarettes can have 100's of additives. Many of them probably shouldn't belong in my lungs/body.

N/A Utterance Sample

N/A instances are majorly simple health-related inquiries N/A Utterance Sample

## Result *EDA: Primary Intent (User Initial Thread == Root)*

Feature	Schema		
primary_intent	$Depth=0:$ 1: Product recommendation 2: Diagnosis 3: Fact check 4: Console 5: Lifestyle 6: Disease management 7: Careprovider recommendation $Depth=1:$ 8: Question 9: Answer 10: Feedback 11: Follow up		
Primary_intent	Count	Percentage	
0	Diagnosis	673	67.30%
1	Disease management	72	7.20%
2	Lifestyle	68	6.80%
3	Careprovider recommendation	52	5.20%
4	Console	41	4.10%
5	Fact check	38	3.80%
6	Product recommendation	35	3.50%
7	Question	20	2.00%
8	Feedback	1	0.10%



I am 17, female and 109lb. Also 5ft 4 if that helps. For the last 6 months-a year, I've been feeling sick bloated after every time I eat something, even if it's something small. It's not that I'm eating too much because when I've not eaten all day, even if I eat a small meal, I feel very sick. I hope somebody can help, I don't wanna waste time going to the doctors if it's nothing serious. Thanks :)

## Diagnosis

Disease management

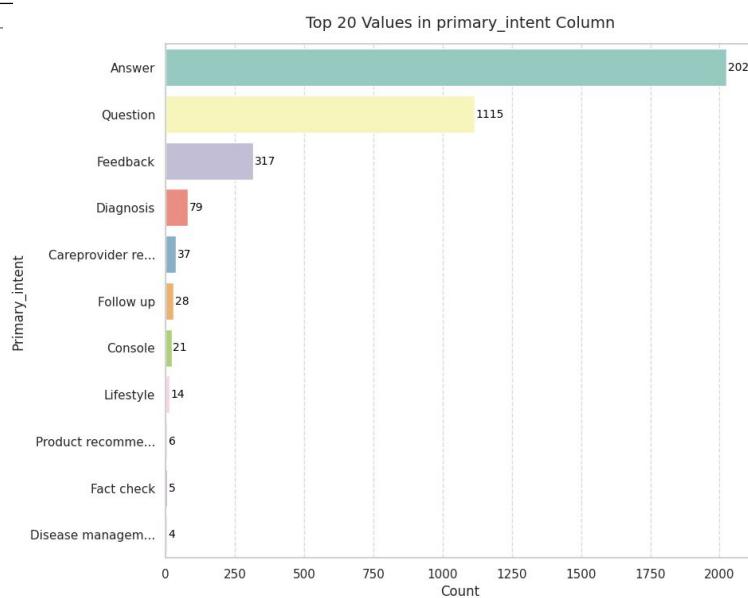
So, long story short. After injuring my leg I got prescribed Ibuprofen, I took this and after a few days started feeling sick, burping and finding it hard to breath. I went to a doctor and had it checked out on the night it got really bad (I was spewing up and had horrible pain in my stomach) the doctor said that it could be caused by the Ibuprofen and gave me some other pills to take to settle my stomach acids down. My question is, how long does it take for the stomach pains to go away? I don't have them continuously but today in particular I have been feeling really sick in the stomach. It's been about five days since I went to the doctor and I haven't taken any more Ibuprofen since then. Anyway, thanks for any answers you might have.

## Top 2 Primary Intent Utterance Samples

Hints that Reddit users' initial thread posts primarily seek diagnoses

# Result EDA: Primary Intent (Assistant(s)' reply)

Feature	Schema
primary_intent	<p><i>Depth=0:</i></p> <ul style="list-style-type: none"> <li>1: Product recommendation</li> <li>2: Diagnosis</li> <li>3: Fact check</li> <li>4: Console</li> <li>5: Lifestyle</li> <li>6: Disease management</li> <li>7: Careprovider recommendation</li> <p><i>Depth!=0:</i></p> <ul style="list-style-type: none"> <li>8: Question</li> <li>9: Answer</li> <li>10: Feedback</li> <li>11: Follow up</li> </ul> </ul>
	<b>Primary_intent</b>
0	Answer
1	Question
2	Feedback
3	Diagnosis
4	Careprovider recommendation
5	Follow up
6	Console
7	Lifestyle
8	Product recommendation
9	Fact check
10	Disease management



I'll get back to you on the name of the tablets when I get home. But just wondering, do you know if drinking milk help with the stomach pains? I think I saw that somewhere but I'm not quite sure if it was true.

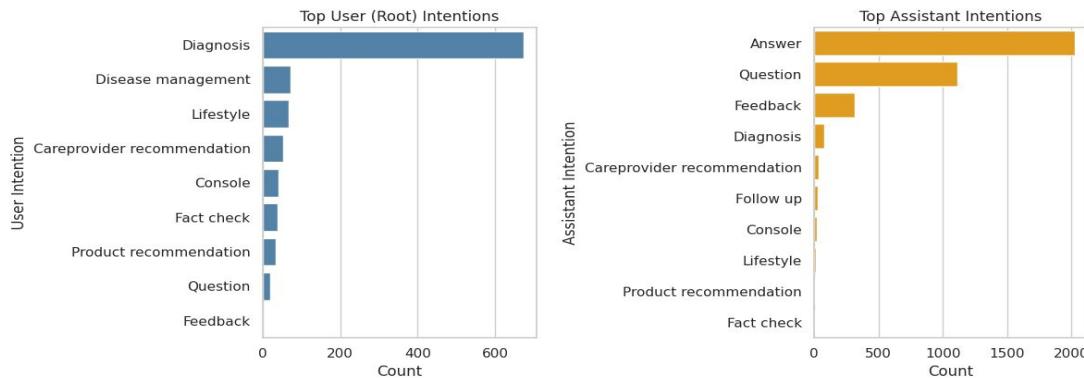
Answer

Wait, you take oral penicillin daily because of your splenectomy? Does a physician prescribe that? My mind is boggled...

Question

Despite refinements, gaps remain, requiring more controlled, fine-tuned, or rule-based categorization of utterances and speakers, integrated with LLM automation

# Result EDA: Room for Improvement



Has_question		Sample 1
Answer	0 True	<p>NSAIDs (non steroidal anti inflammatories) are known to play havoc on your stomach lining. Make sure you have plenty to eat when taking them and always stick to the required dosage. The pain should ease after 3-5 days once your stomach lining has regenerated. What tablets are you on now?</p>

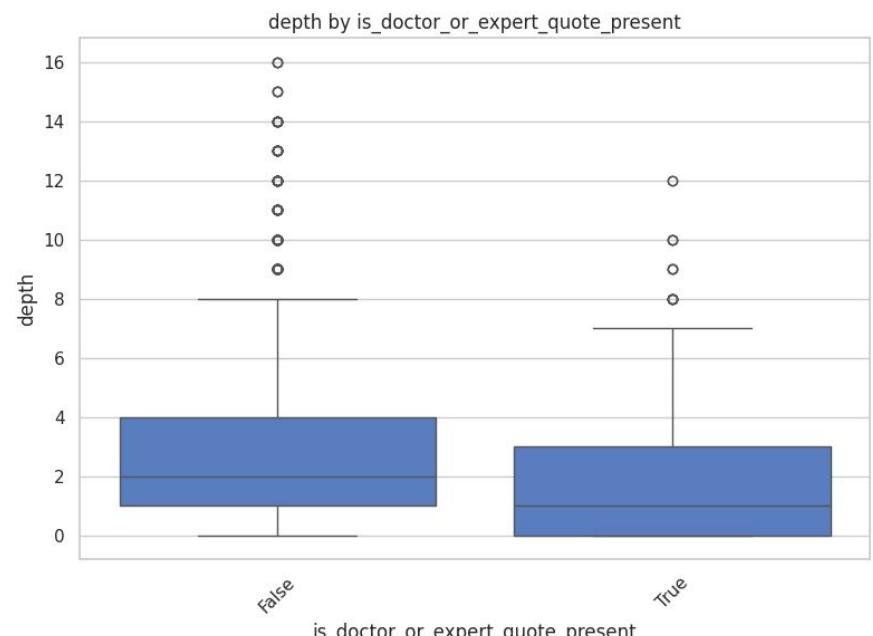
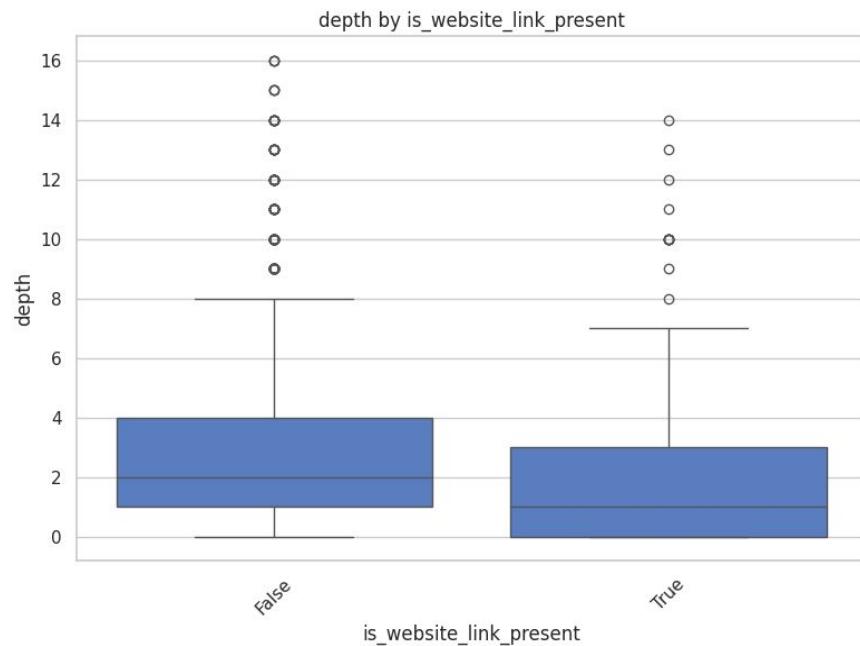
  

More information	NSAIDs (non steroidal anti inflammatories) are known to play havoc on your stomach lining. Make sure you have plenty to eat when taking them and always stick to the required dosage. The pain should ease after 3-5 days once your stomach lining has regenerated. What tablets are you on now?
------------------	--

[Assistant] Is Answer → Has Question → More Information == answer with follow-up question

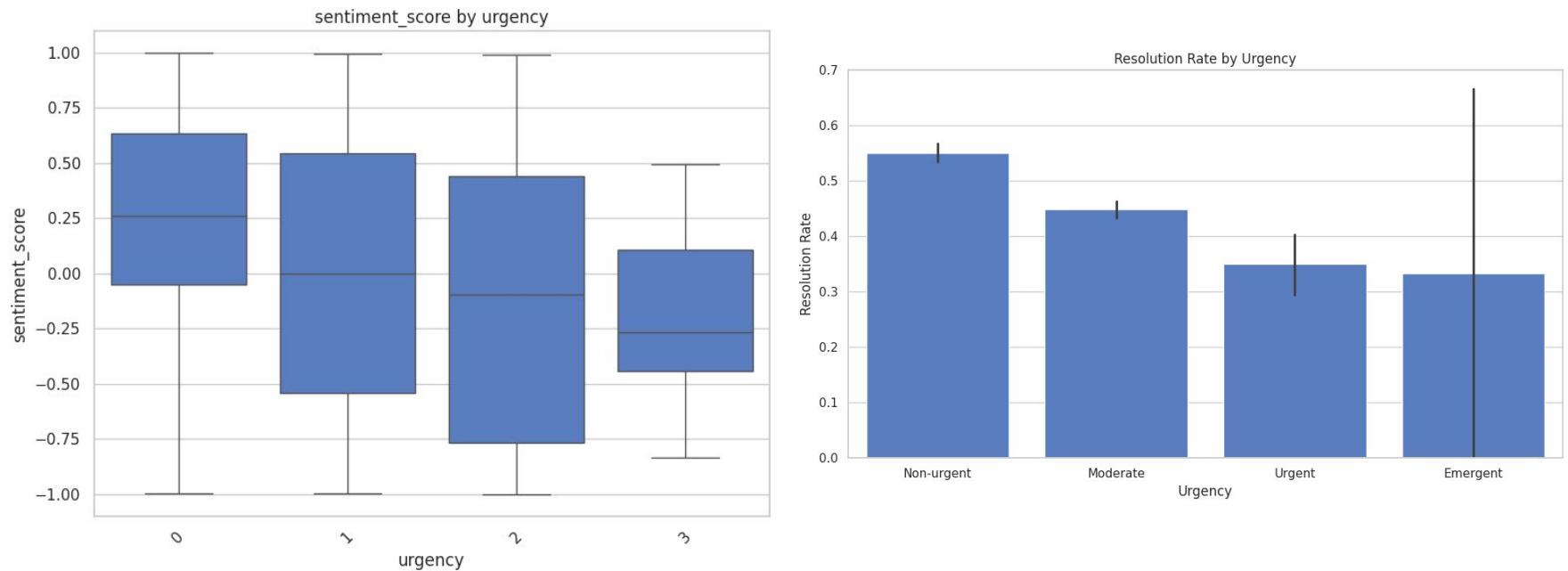
Dashboard is the way to go

## Result Association - Boxplot: Depth



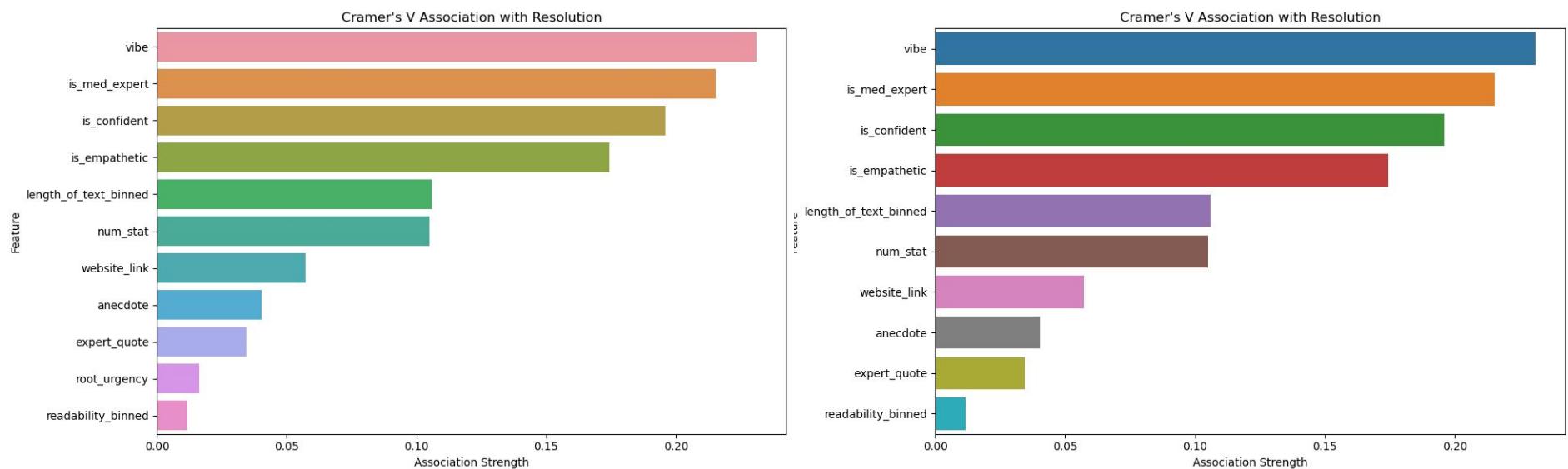
Conversations *without website links or doctor quotes* have greater **depth**,  
suggesting *links and quotes* may shorten discussions, with outliers reaching high **depths**

## Result Association - Boxplot: Urgency



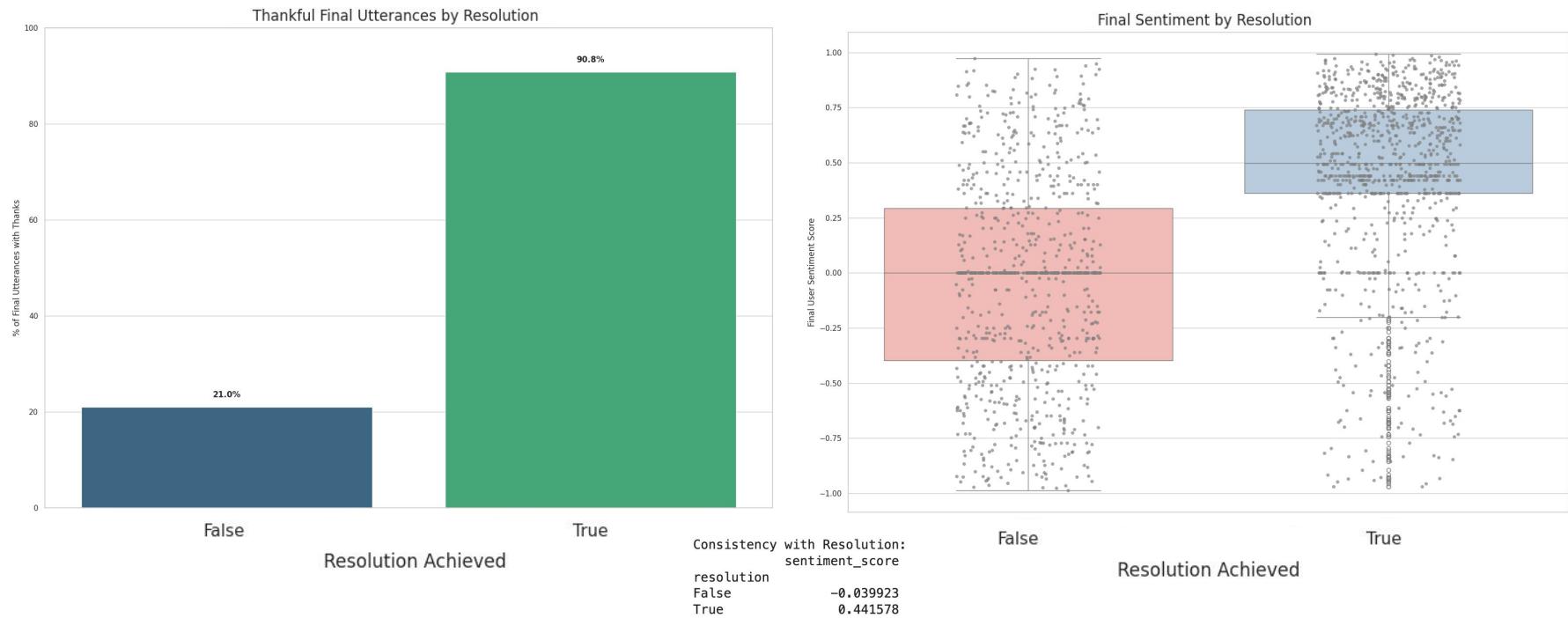
*Non-urgent* queries have more positive **sentiment** and higher **resolution** rates than *urgent* ones,  
which lean negative and are less *resolved* due to complexity

## Result Association - Cramer's V Plot



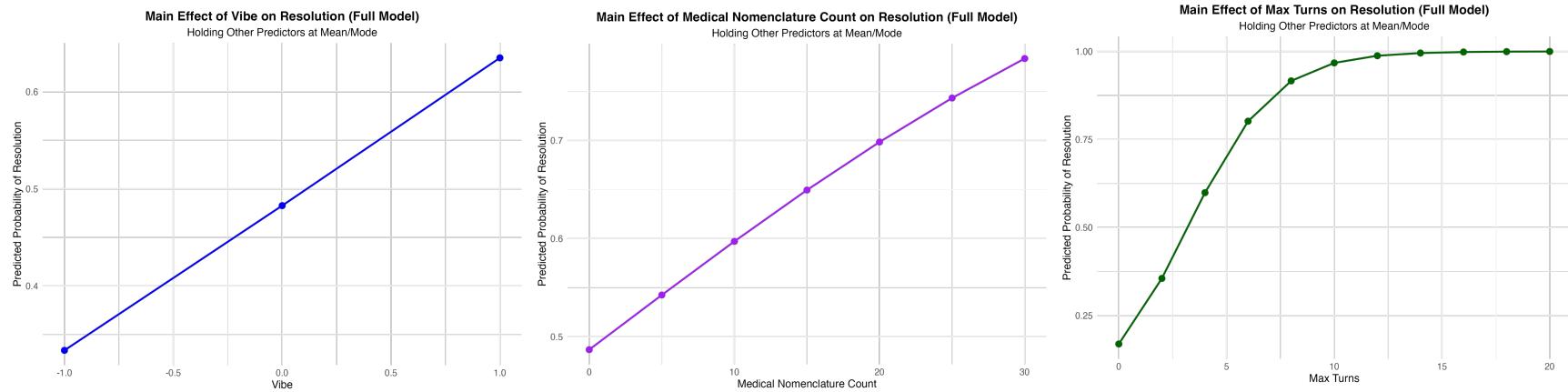
Cramer's V plot reveals a strong correlation between **vibe** (positive/accepting, negative/rejecting, or neutral tone) and **resolution**, with minimal impact on other feature associations in the reduced model after excluding **root\_urgency**, **score\_sum**, and **medical\_term**

## Result Association - Resolution Validation



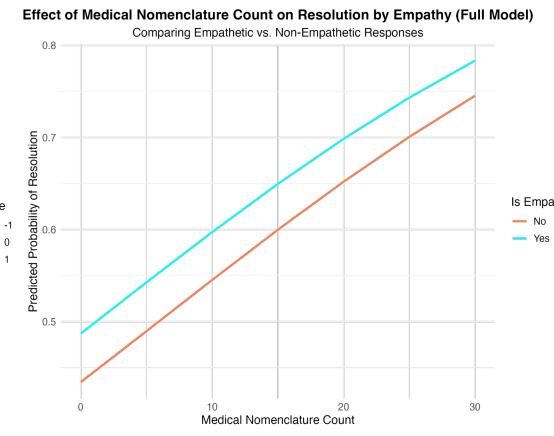
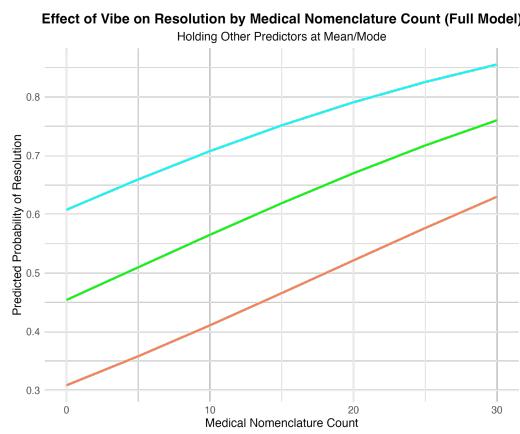
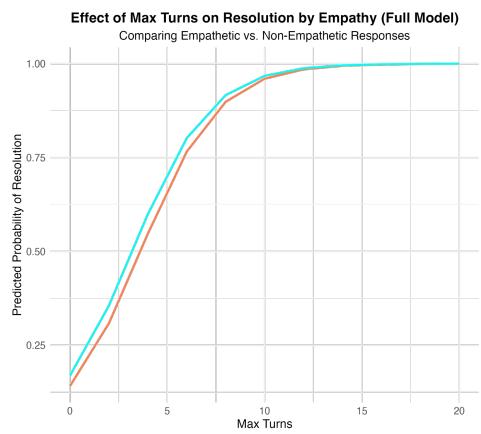
*Resolved* subthreads show 89% **thankfulness** and higher positive **sentiment** (mean 0.44) compared to 21% **thankfulness** and slightly negative **sentiment** (mean -0.04) in *unresolved* subthreads, validating both as strong indicators of inquiry resolution

# Result *Multilevel Logistic Regression*



Used multilevel logistic regression (via R Lme4 library) to account for the nested structure (subthreads within conversations)

# Result *Interaction Plots*



More **turns** strongly increase **resolution** probability,  
*empathetic* responses consistently outperform *non-empathetic* ones, with a small but consistent gap

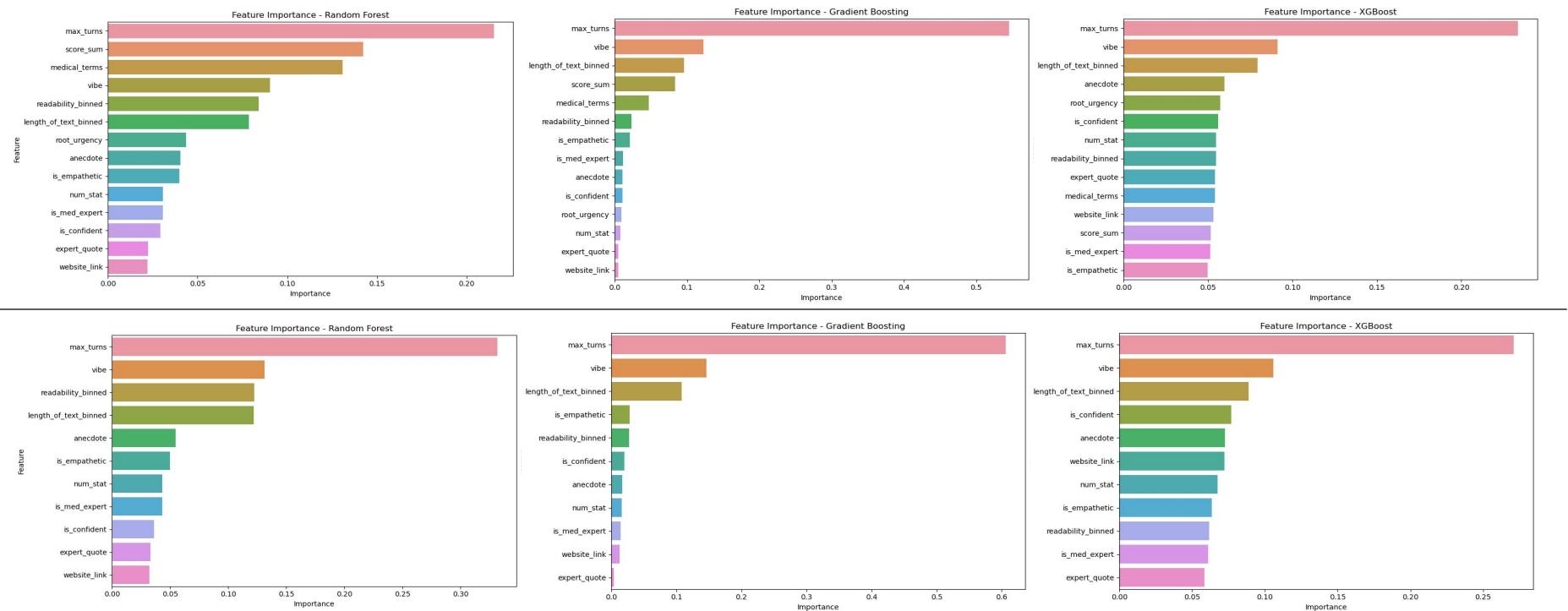
## Result *Predictive Model Performance*

Model (Scenario/Run)	ROC Metric	F1 Metric
<b>Before Refinement (Older Run with Erroneous Data)</b>		
Random Forest (Scenario 1)	0.563 (ROC AUC)	0.814 (CV F1 Mean)
HistGradientBoosting (Scenario 2)	0.587 (ROC AUC)	0.362 (CV F1 Mean)
Random Forest (Scenario 3)	0.613 (ROC AUC)	0.542 (CV F1 Mean)
<b>After Refinement (Run with Further Dropped Features, 11 Features)</b>		
Logistic Regression	0.765 (CV ROC AUC)	0.684 (Macro F1)
Random Forest	0.756 (CV ROC AUC)	0.659 (Macro F1)
Gradient Boosting	0.803 (CV ROC AUC)	0.689 (Macro F1)
XGBoost	0.764 (CV ROC AUC)	0.674 (Macro F1)

Table 7: Combined Model Performance Before and After Refinement

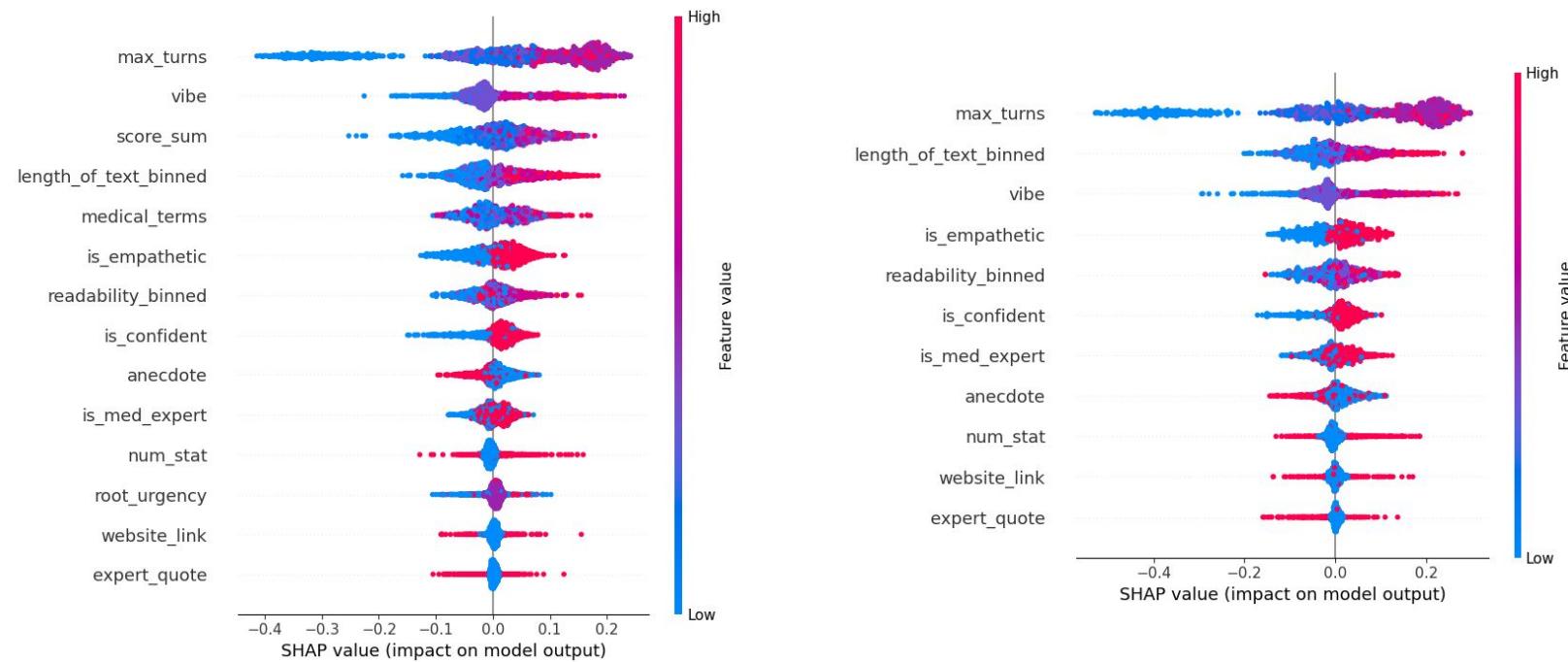
Refining the dataset and dropping low-impact features significantly improves model performance

# Result *Feature Importance*



Max\_turns, vibe, and length\_of\_text\_binned are top predictors of resolution across Random Forest, Gradient Boosting, and XGBoost, with the reduced model *below* maintaining predictive power after dropping low-importance features

## Result SHAP Summary Plot



SHAP summary plot highlights `max_turns` and `vibe` as dominant predictors of `resolution`, with longer conversations and positive tone increasing likelihood (up to +0.3), while dropped features (right) have minimal impact

# Conclusion

- **Encourage Extended Engagement:** Design the chatbot to sustain multi-turn conversations by asking follow-up questions (e.g., about symptom duration or severity), as longer interactions significantly enhance resolution
- **Prioritize Positive and Empathetic Tone:** Train the chatbot to adopt a consistently positive tone (vibe = 1) and empathetic responses, as these increase resolution probability by up to 0.25 and 0.05, respectively, across all contexts
- **Proactive Question Generation:** Implement a question-asking mechanism to extend conversations, focusing on user-relevant prompts (e.g., “Can you describe your symptoms further?”), to maximize engagement and resolution likelihood

## Limitations *General*

<b>Limited Domain Tuning</b>	General models miss clinical nuances without specific data & sample size
<b>Non-Traditional Data Issues</b>	Reddit's own structure & its informal content (images, non-expert replies) hinders consistent extraction
<b>Inconsistent Output</b>	Variable performance in capturing subtle cues across clinical texts (e.g., enhance schema labels)
<b>Prompt Challenges</b>	Optimizing prompts for tasks like urgency classification is difficult
<b>Weak Nuance Detection</b>	Struggles with complex linguistic patterns (e.g., sentiment, intention)
<b>Basic Feature Set</b>	Current features lack depth for row-level interactions
<b>Model Limitations</b>	Advanced models (e.g., deep learning) could improve feature capture but weren't used
<b>LLM Reliability</b>	Competitive but unreliable for critical clinical use without validation

## Limitations *Model Limitation*

- General models lack clinical tuning, reducing accuracy with medical terminology
- Limited control over entities/relations misses nuances like negation or severity
- Inconsistent results across clinical notes without optimized prompts
- Zero-/few-shot settings lag behind fine-tuned models
- Fine-tuning (and possibly newer models) needed for better performance
- Study was experimental

Thank you

# References

- [1] Amazon Web Services. (n.d.). *What is Retrieval-Augmented Generation (RAG)?*
- [2] Abbasian, M., Khatibi, E., Azimi, I. et al. *Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI.* npj Digit. Med. 7, 82 (2024). <https://doi.org/10.1038/s41746-024-01074-z>
- [3] Salemi, A., Mysore, S., Bendersky, M., & Zamani, H. (2024). *LaMP: When large language models meet personalization* (Version 4) [Preprint]. arXiv. <https://doi.org/10.48550/arXiv.2304.11406>
- [4] Hu, Y., Zuo, X., Zhou, Y., Peng, X., Huang, J., Keloth, V. K., Zhang, V. J., Weng, R.-L., Chen, Q., Jiang, X., Roberts, K. E., & Xu, H. (2025). *Information extraction from clinical notes: Are we ready to switch to large language models?* arXiv. <https://arxiv.org/abs/2411.10020>
- [5] Johri, S., Jeong, J., Tran, B.A. et al. An evaluation framework for clinical use of large language models in patient interaction tasks. Nat Med 31, 77–86 (2025). <https://doi.org/10.1038/s41591-024-03328-5>