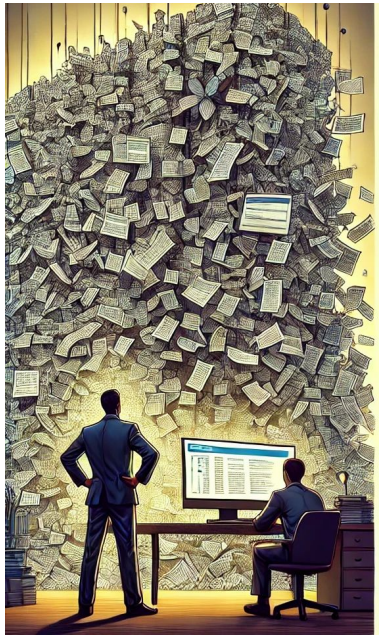
Zeteo Health Assistant: RAG-based Chatbot with Hyper-Personalization

BIS 686 Capstone Project Presentation

Tom Shin

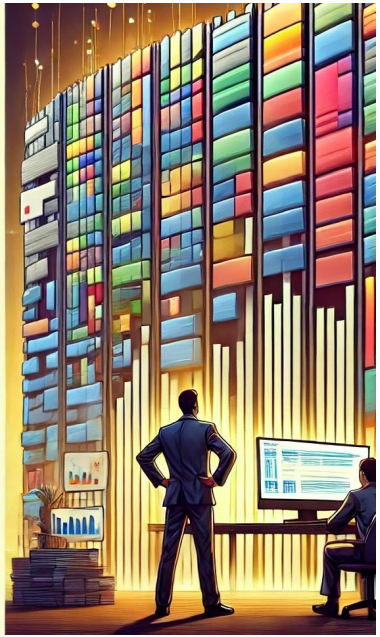
Mentored by Eric Landry

Introduction *What & Why*



Paper-based search result

Internet-based search result



AI-based search result

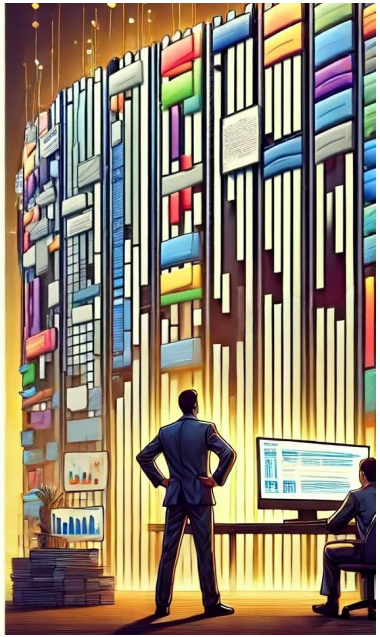
Generated using OpenAI's DALL-E. Comparison illustration depicting stages of text data processing. Created on December 9, 2024.

Introduction *What & Why*

Discreet selection enhanced by Hyper-Personalization

Paper-based search result

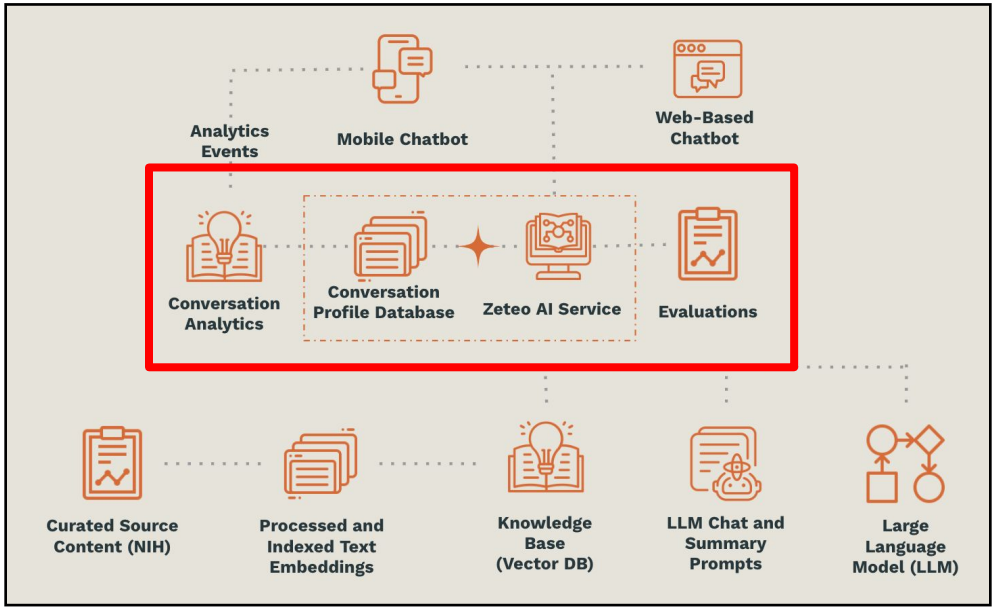
Internet-based search result



AI-based search result

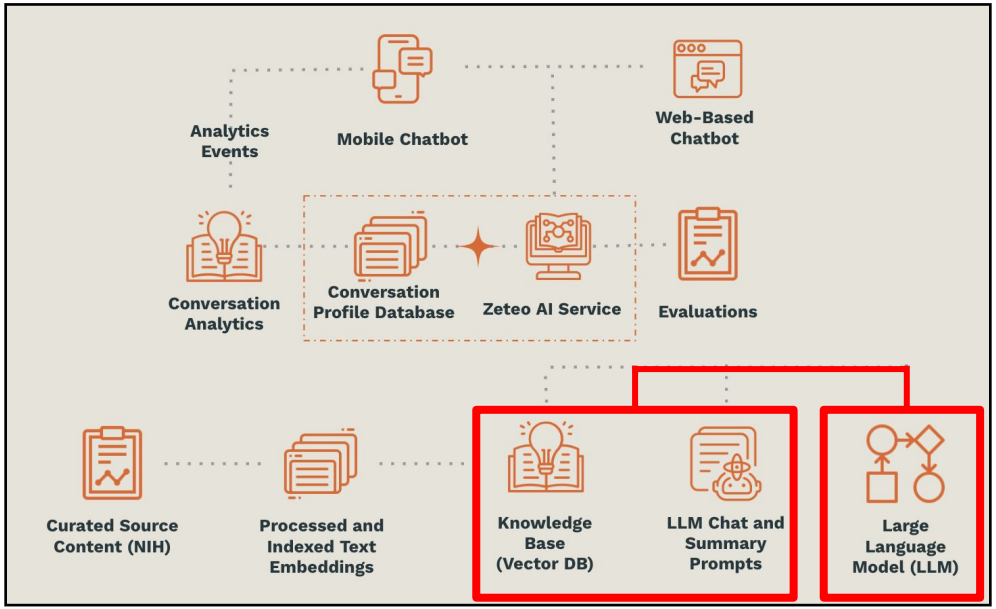
Generated using OpenAI's DALL-E. Comparison illustration depicting stages of text data processing. Created on December 9, 2024.

Introduction *Key Focus*



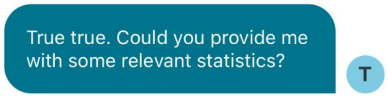
The Zeteo Health conversational AI framework uses a Retrieval-Augmented Generation (RAG) Large Language Model (LLM) architecture, powered by a professionally curated knowledge base to deliver context-aware, personalized responses

Introduction *RAG-LLM?*



Retrieval-Augmented Generation (RAG) is the process of optimizing the output of a LLM, so it references an authoritative knowledge base outside of its training data sources before generating a response[1]

Scoping *Last semester*

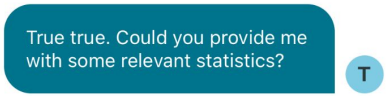


How should information be delivered? Could formats like bullet points improve clarity? Intent & Preference

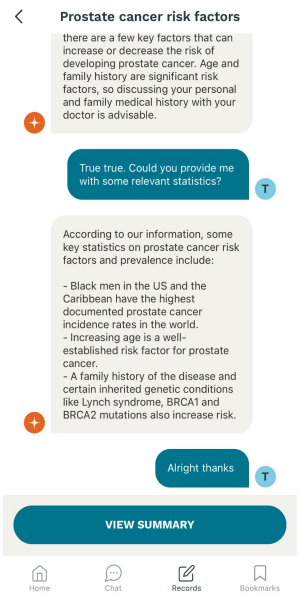
How easy or difficult should information be? Would the user's comprehension level enhance response quality? Reading Comprehension

Last semester's LLM misread comprehension level prompts, limiting output length not comprehension, based on flawed assumptions tying user input to understanding

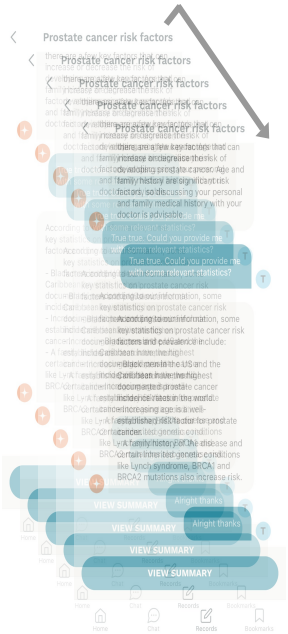
Scoping *Defining Features by Level of Interaction*



Single utterance



Multi-turn conversation

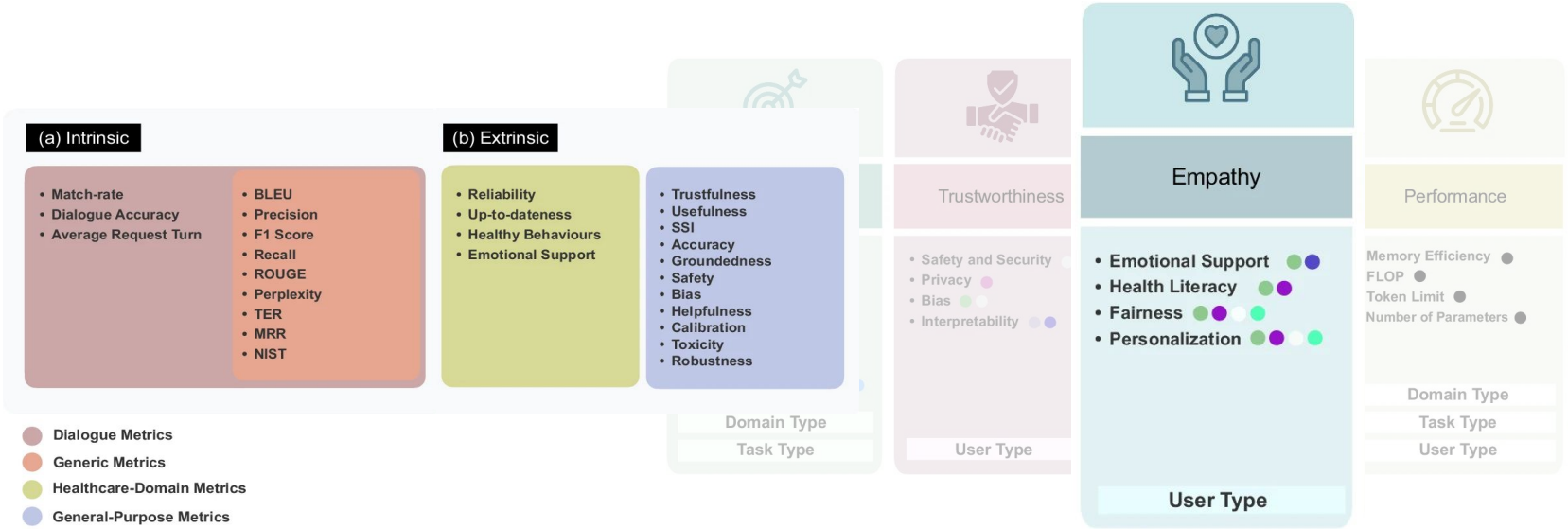


History

Shifted focus from evaluating single utterance quality to multi-conversation analysis

→ Redefine measurement criteria to capture meaningful features in dynamic, multi-turn healthcare assistant interactions

Scoping *Existing Metrics*



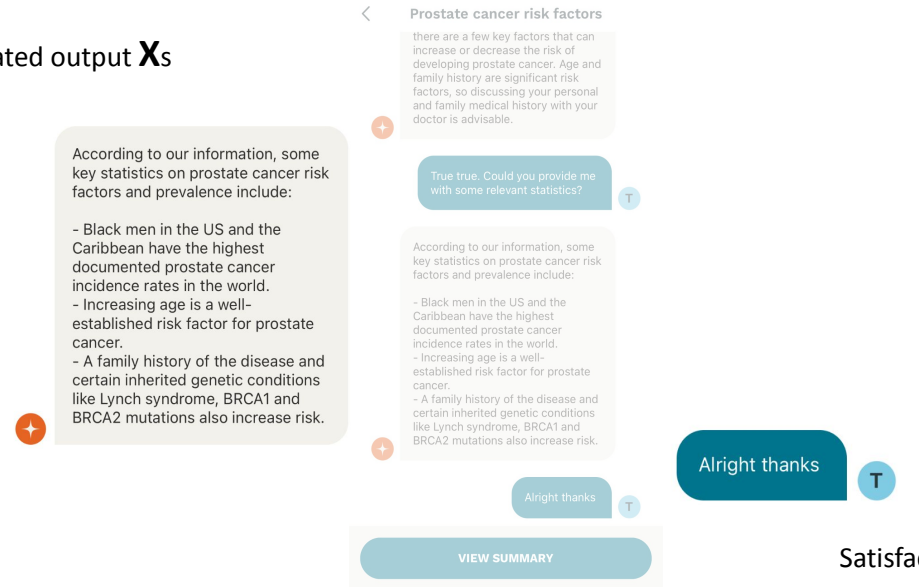
- **Intrinsic** metrics (*e.g., BLEU, ROUGE*) focus on language, not medical concepts or patient well-being
- **Extrinsic** methods (*e.g., human judgment*) are narrow, missing holistic healthcare needs

Overlook empathy, trust-building, personalization, and emotional support[2]

Scoping *Multi-turn Conversation*

Assistant

Features from generated output **X**s

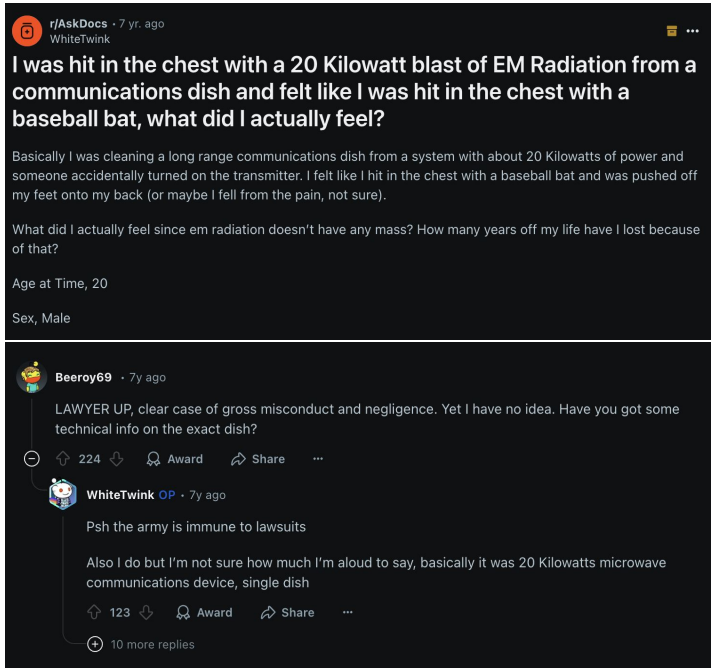


User

Satisfaction as destination **Y**

Investigated 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"*



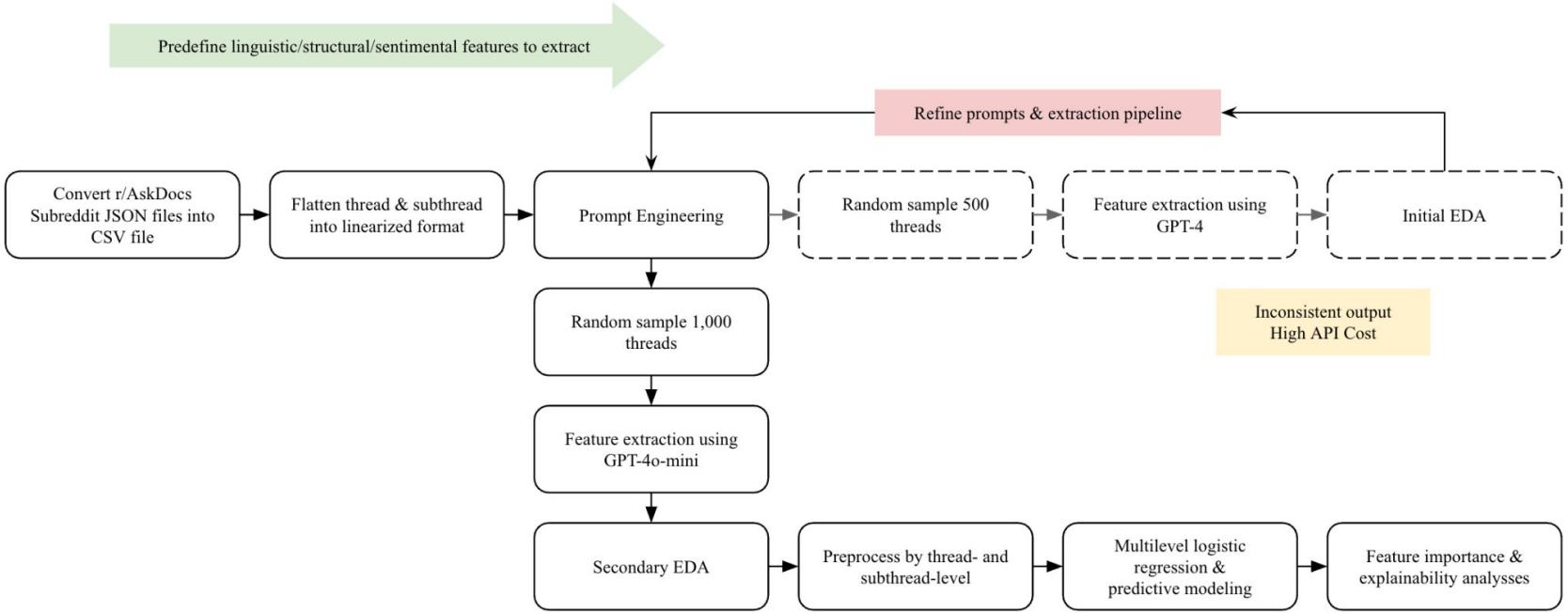
Prior studies utilized movie ratings to assess subjective conversational quality[3]

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

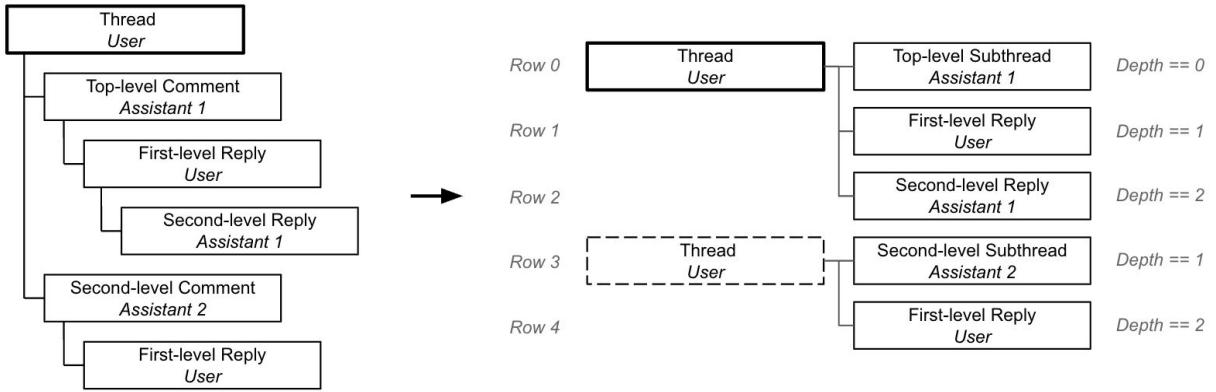
Methods *Project Design*

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

Methods *Workflow Diagram*

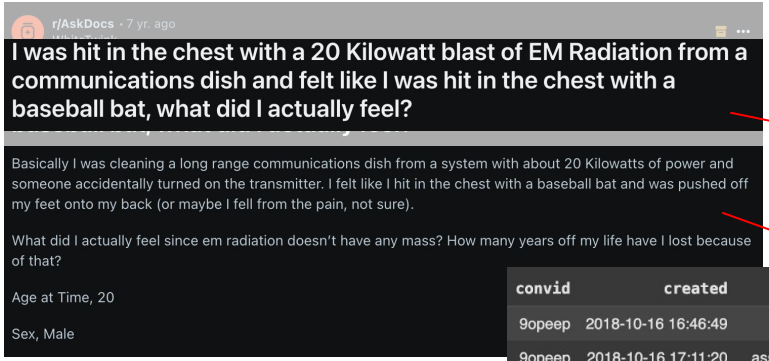


Data *Conversion*

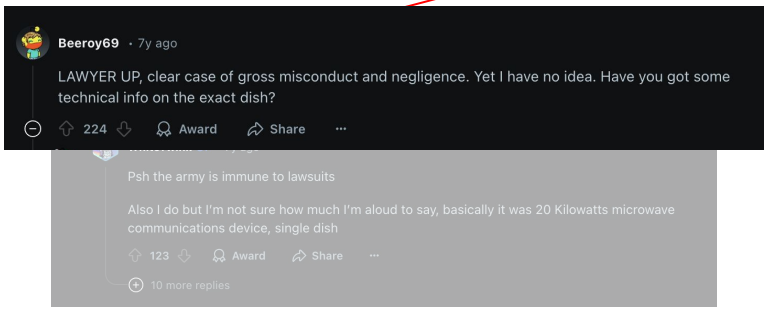


Transformed Reddit's thread-like conversation data into a linear table/dataframe, structuring user threads with root posts branching into subthreads, each representing individual user-assistant conversations

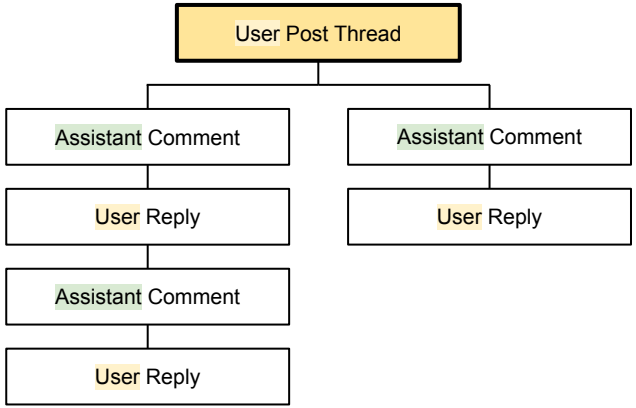
Data *Conversion*



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...



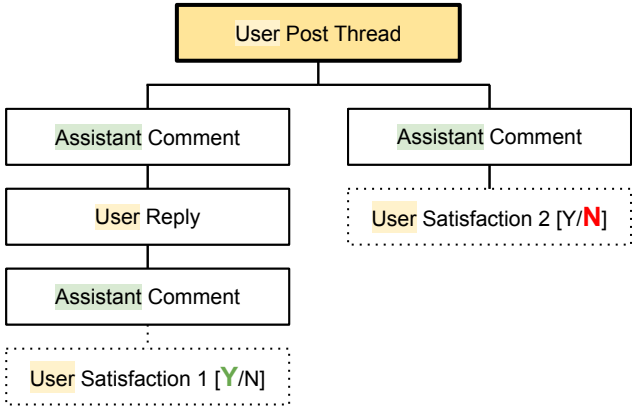
Data *Conversion*



conversation_id	id	parent_id	subthread_id	speaker	role
22woxm	22woxm			Too-Cool	user
22woxm	cgr68sb	22woxm	22woxm_1	Carboxia	assistant1
22woxm	cgr80w3	22woxm	22woxm_2	doublecross	assistant2
22woxm	cgrh1eq	cgr80w3	22woxm_2	Too-Cool	user
22woxm	cgrh23p	cgr68sb	22woxm_1	Too-Cool	user
22woxm	cgrlram	cgrh1eq	22woxm_2	doublecross	assistant2
22woxm	cqsqdc7	cgrlram	22woxm_2	Too-Cool	user
23136h	23136h			jysalia	user
23136h	cguavmv	23136h	23136h_2	allhailbrightfutur	assistant1
23136h	cguavwm	cguavmv	23136h_1	autowikibot	assistant2
23136h	cgufd9b	cguavmv	23136h_2	jysalia	user
23136h	cgugbvb	cgufd9b	23136h_2	allhailbrightfutur	assistant1
23136h	cguh50k	cgugbvb	23136h_2	jysalia	user
23136h	cguhhut	cguh50k	23136h_2	allhailbrightfutur	assistant1

Flattened thread format into linear conversation format

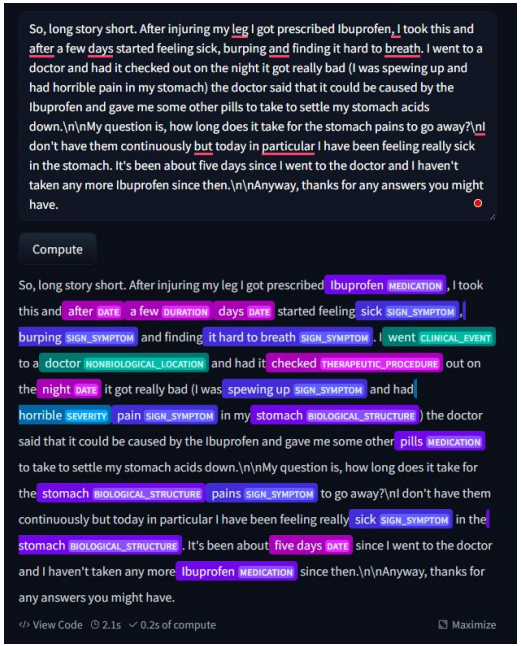
Data *Feature Extraction*



conversation_id	id	parent_id	subthread_id	speaker	role	Resolved?
22woxm	22woxm			Too-Cool	user	TRUE
22woxm	cgr68sb	22woxm	22woxm_1	Carboxia	assistant1	
22woxm	cgr80w3	22woxm	22woxm_2	doublecross	assistant2	
22woxm	cgrh1eq	cgr80w3	22woxm_2	Too-Cool	user	
22woxm	cgrh23p	cgr68sb	22woxm_1	Too-Cool	user	
22woxm	cgrlram	cgrh1eq	22woxm_2	doublecross	assistant2	
22woxm	cqsqdc7	cgrlram	22woxm_2	Too-Cool	user	
23136h	23136h			jysalia	user	FALSE
23136h	cguavmv	23136h	23136h_2	allhailbrightfutur	assistant1	
23136h	cguavwm	cguavmv	23136h_1	autowikibot	assistant2	
23136h	cgufd9b	cguavmv	23136h_2	jysalia	user	
23136h	cgugbvb	cgufd9b	23136h_2	allhailbrightfutur	assistant1	
23136h	cguh50k	cgugbvb	23136h_2	jysalia	user	
23136h	cguhhut	cguh50k	23136h_2	allhailbrightfutur	assistant1	

Manual review of LLM-annotated **resolution**, supplemented with user satisfaction and sentiment score trajectory

Data *Feature Extraction*




```
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 = 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[3]

→ Used LLM *GPT-4o-mini & few-shot prompting* instead of Named Entity Recognition 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}"""
```

```
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

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.

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 left-side resolution 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

Revamped prompt and extraction pipeline to cut computation costs while boosting coherence and precision with more context

Data *Old vs. Refined Overview*

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

Data *Features Extracted [ALL]*

Feature	Description	Value	Extraction Level	Example	LLM Used
Metadata					
Raw-Reddit					
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,,	✗
timestamp	Post timestamp	Integer	Metadata	1392350217	✗
score	Post rating	Integer	Row	1	✗
Processed					
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
Extracted					
Structure					
length_of_text	Word count	Integer	Row	5	✗
readability	Text readability	Float	Row	5.0	✗
Conversation					
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	✓
Component					
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	✓
Sentiment					
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	✓
Destination					
resolution	Concern resolved	Boolean	Thread	False	✓
thankfulness	User gratitude	0-2/None	Row	1	✓

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

Data *Categorical Feature Labels*

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!=0*: 8: Question 9: Answer 10: Feedback 11: Follow up
body_area	0: N/A 1: Head 2: Chest 3: Stomach 4: Right arm 5: Left arm 6: Right leg 7: Left leg 8: Back 9: Neck
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

Feature	Schema
urgency	0: Non-urgent 1: Moderate 2: Urgent 3: Emergent
question_type	*User*: 1: More information 2: Easier interpretation 3: Other *Assistant*: 4: More information 5: Easier interpretation 6: Other
vibe	1: Positive 0: Neutral -1: Negative
thankfulness	0: None 1: Somewhat 2: Very None

Table 2: Specific Labeling Schema for Cardinal/Ordinal Features

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	✗
readability_bin	Binned readability score	1/2/3/4	Row	Easy	✗
Numerical Features					
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	✓

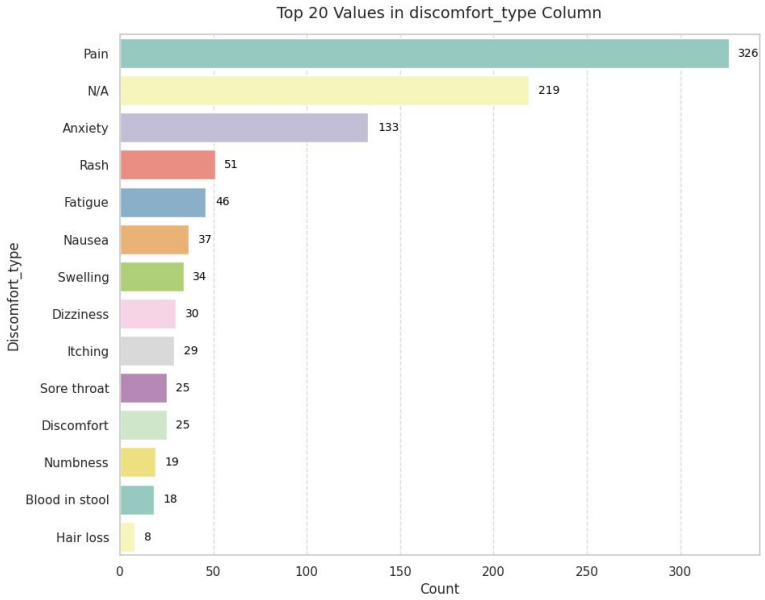
Table 5: Final Features Used in the Full Machine Learning Model (14 Features)

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

Result *EDA: Discomfort Type (Root Thread / User)*

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.

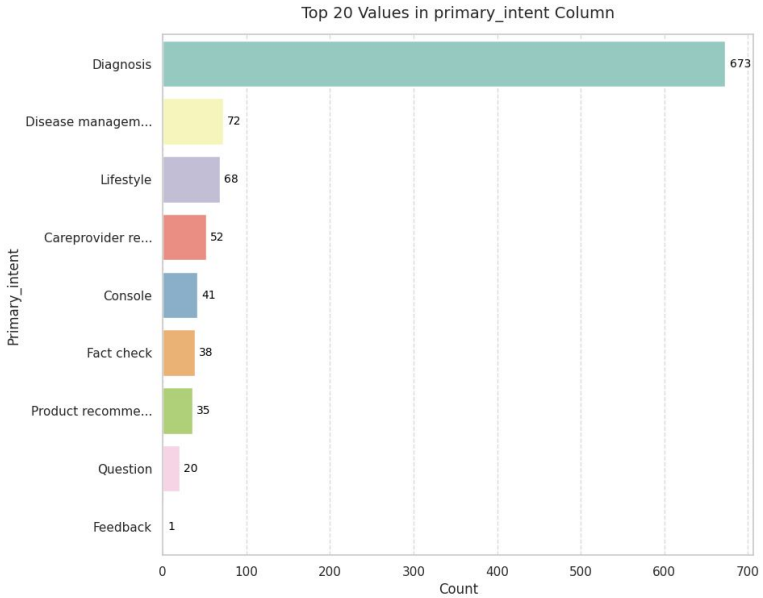
N/A Utterance Sample

N/A utterance sample = simple health-related question

Result *EDA: Primary Intent (Root Thread / User)*

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!=0*: 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%



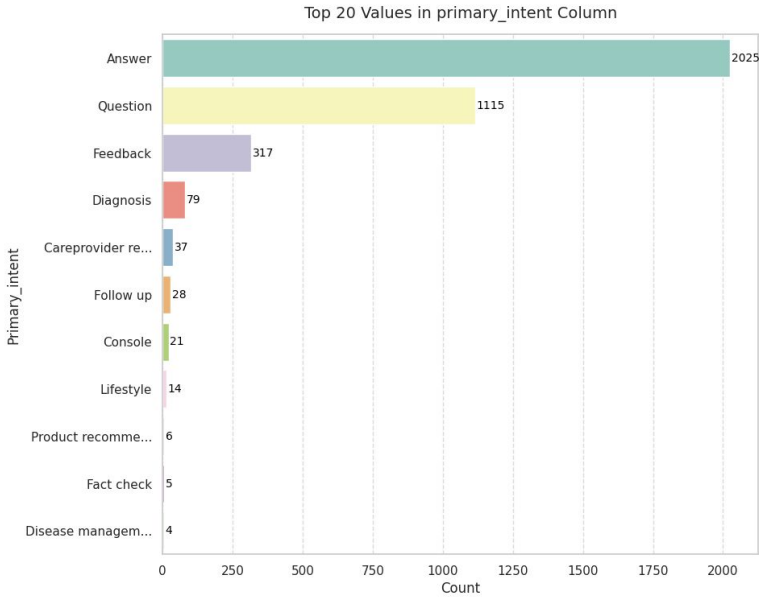
Diagnosis	I am 17, female and 109lb. Also 5ft 4 if that helps. For the last 6 months-a year, I've been feeling sick bloated after every time I eat something, even if it's something small. It's not that I'm eating too much because when I've not eaten all day, even if I eat a small meal, I feel very sick.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

Result *EDA: Primary Intent (Subthread / Assistant(s))*

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!=0*: 8: Question 9: Answer 10: Feedback 11: Follow up

	Primary_intent	Count	Percentage
0	Answer	2025	55.46%
1	Question	1115	30.54%
2	Feedback	317	8.68%
3	Diagnosis	79	2.16%
4	Careprovider recommendation	37	1.01%
5	Follow up	28	0.77%
6	Console	21	0.58%
7	Lifestyle	14	0.38%
8	Product recommendation	6	0.16%
9	Fact check	5	0.14%
10	Disease management	4	0.11%



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.

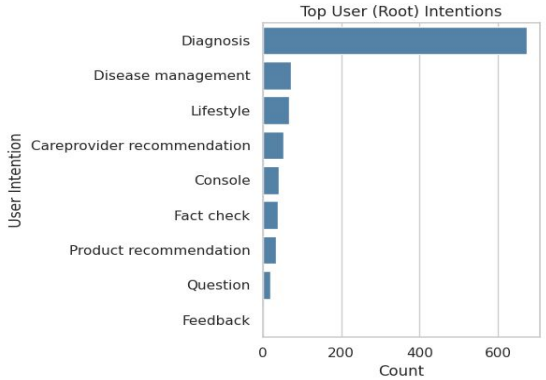
Answer

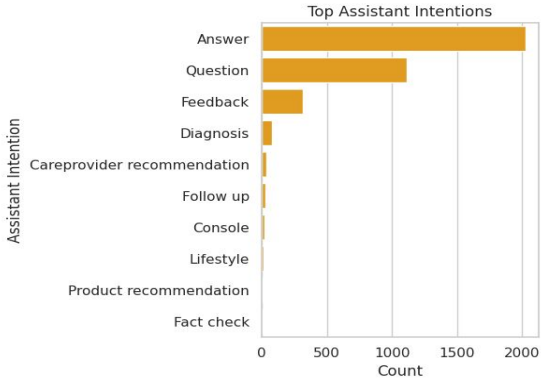
Wait, you take oral penicillin daily because of your splenectomy? Does a physician prescribe that? My mind is boggled...

Question

Though refined, some gaps exist where more controlled/detailed (fine-tuned) or more rule-based categorization to utterances and the speaker, mixed with LLM automation

Result *EDA: Room for Improvement*





Answer	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?

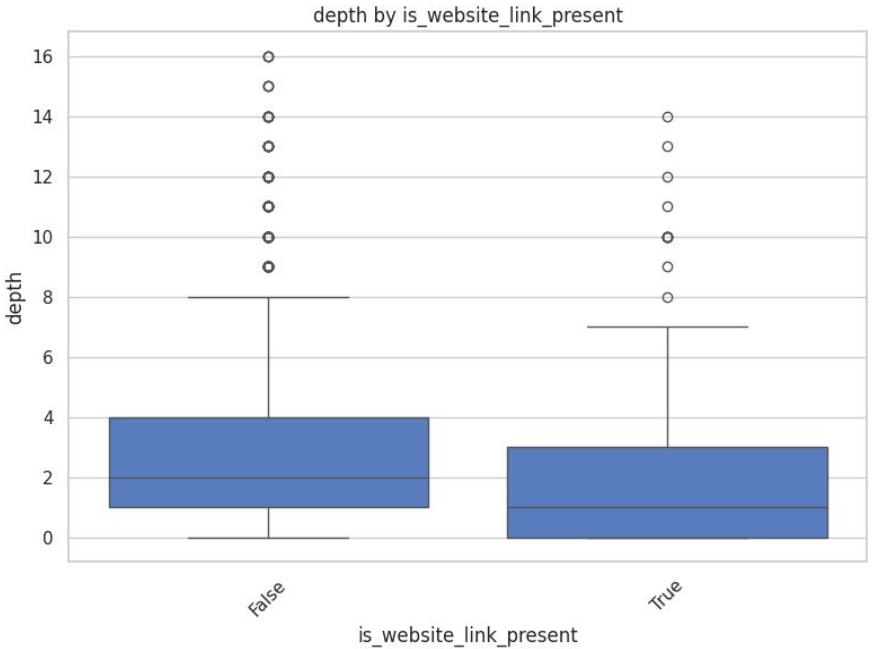
	Has_question	Sample 1
0	True	NSAIDs (non steroidal anti inflammatories) are known to play havoc on your stomach lining. Make sure you have plenty to eat when taking them and always stick to the required dosage. The pain should ease after 3-5 days once your stomach lining has regenerated. What tablets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

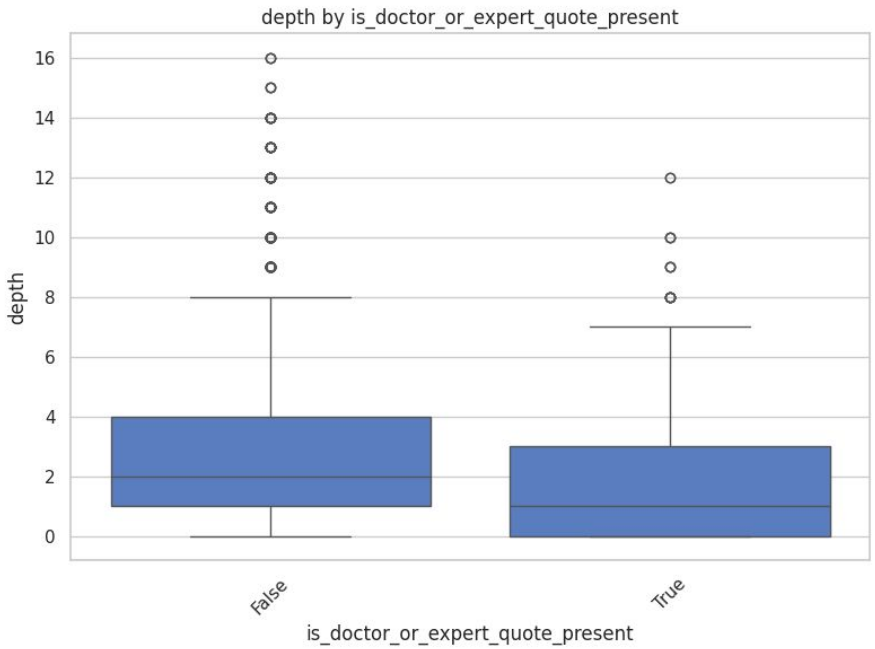
More information	NSAIDs (non steroidal anti inflammatories) are known to play havoc on your stomach lining. Make sure you have plenty to eat when taking them and always stick to the required dosage. The pain should ease after 3-5 days once your stomach lining has regenerated. What tablets are you on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

Assistant's `Is Answer` → `Has Question` → `More Information` == Assistant's answer with follow-up question

Dashboard is the way to go

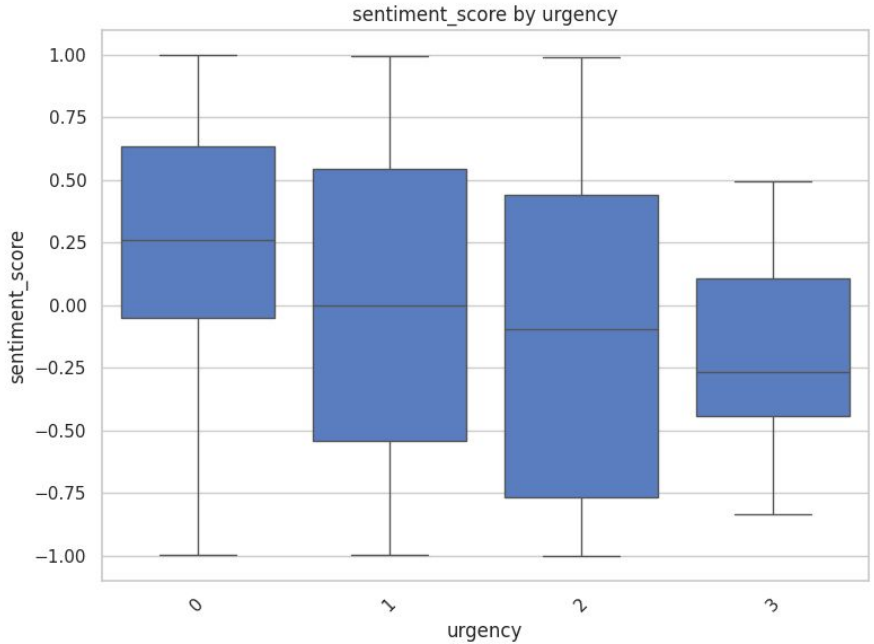
Result *Association - Boxplot: Depth*

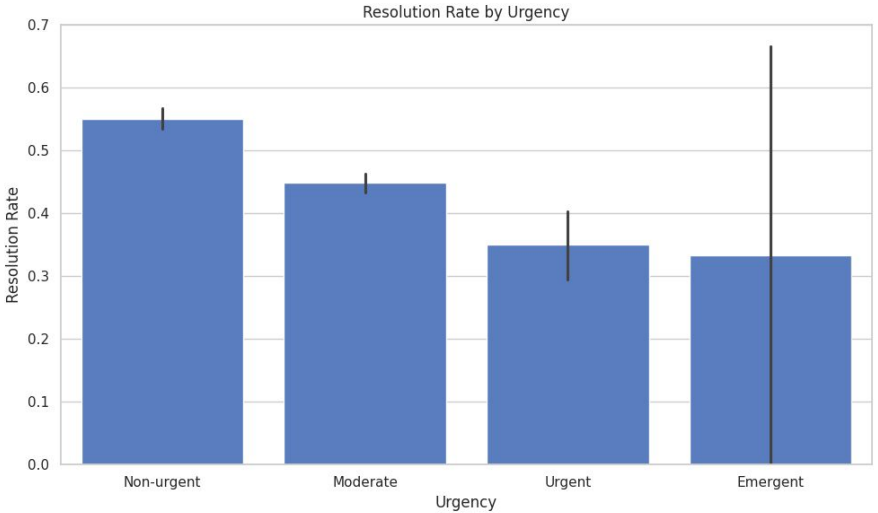




Conversations without website links or doctor quotes have greater depth, suggesting links and quotes may shorten discussions, with outliers reaching high depths

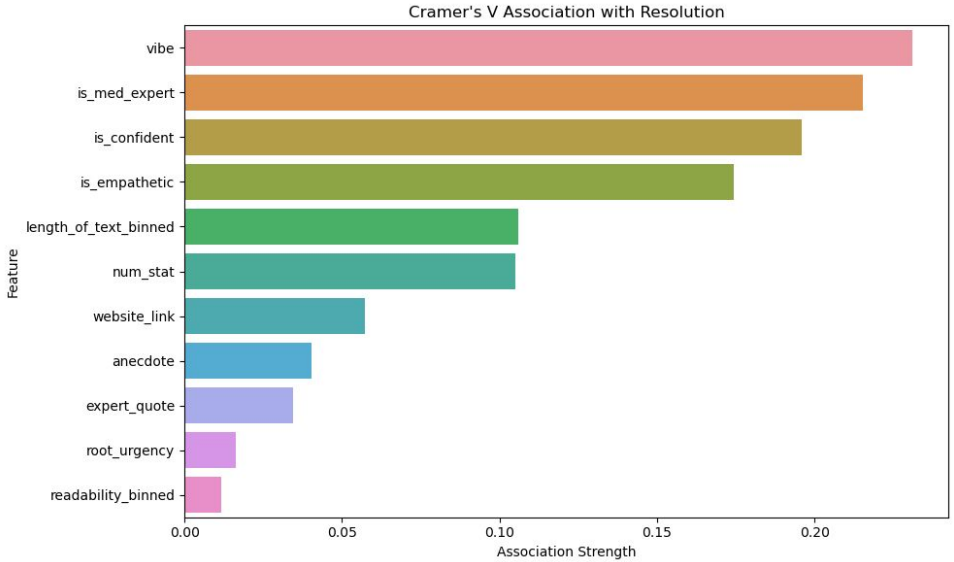
Result *Association - Boxplot: Urgency*

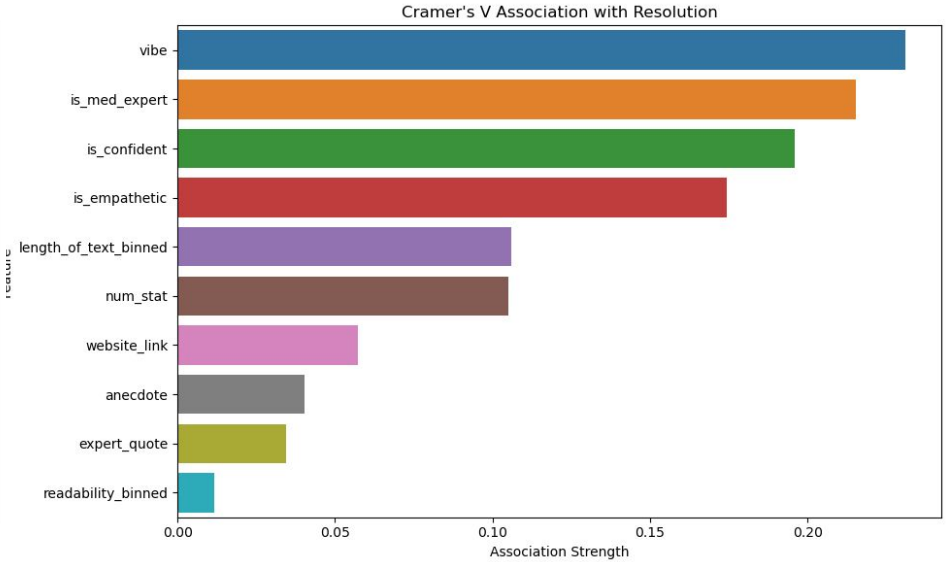




Non-urgent queries have more positive sentiment and higher resolution rates than urgent ones, which lean negative and are less resolved due to complexity

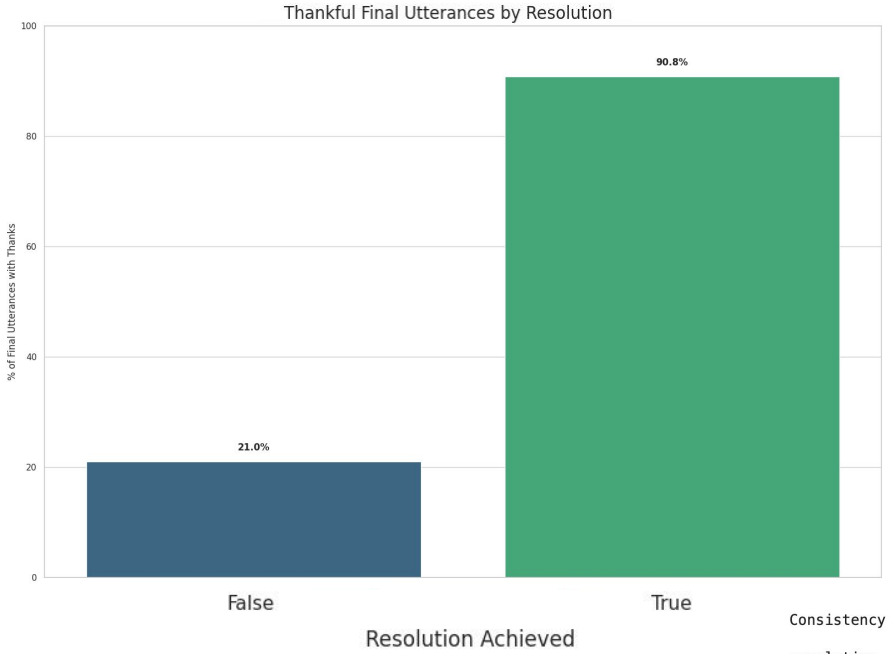
Result *Association - Cramer's V Plot*

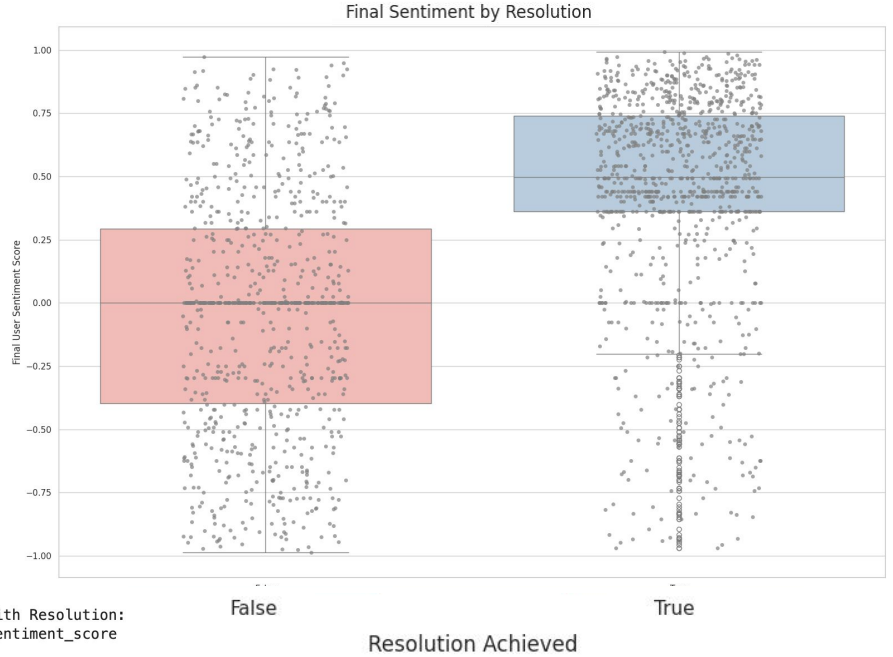




Cramer's V plot shows vibe strongly correlates with resolution, with minimal impact on other feature associations after dropping `root_urgency`, `score_sum`, and `medical_term` in the reduced model

Result *Association - Resolution Validation*



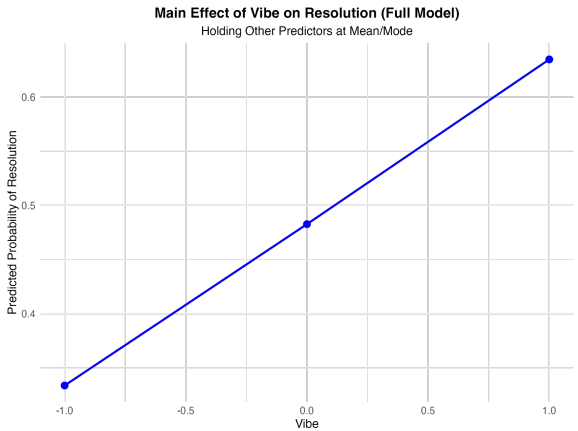


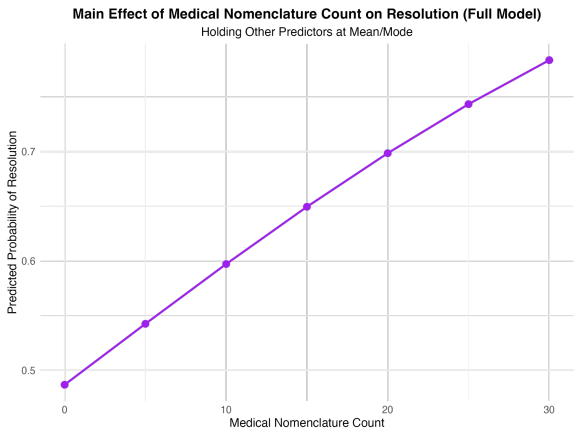
Consistency with Resolution:

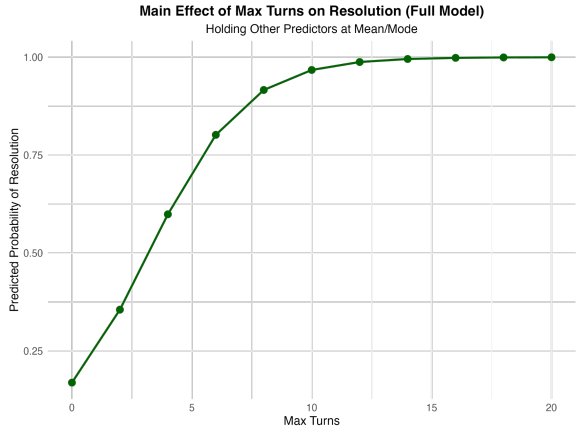
resolution	sentiment_score
False	−0.039923
True	0.441578

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*

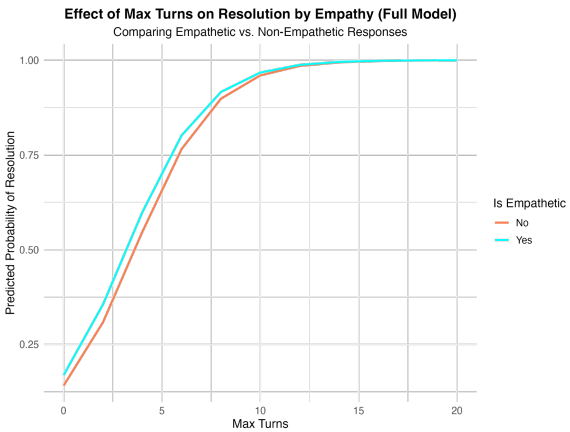


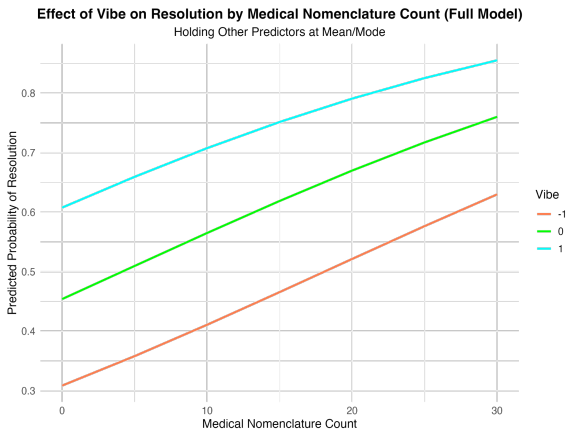


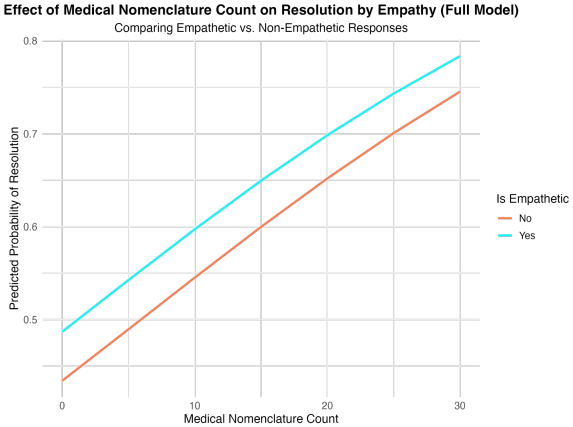


Used multilevel logistic regression (via R Lme4 library) to account for the nested structure (subthreads within conversations)

Result *Interaction Plots*







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

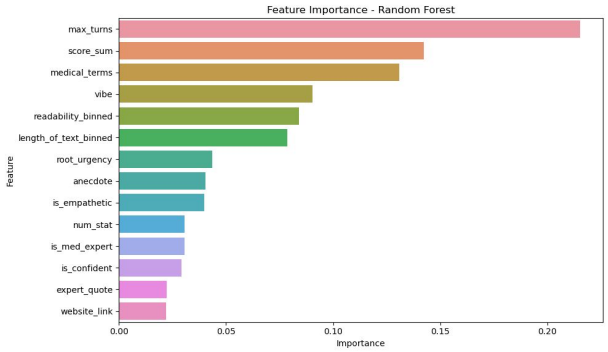
Result *Predictive Model Performance*

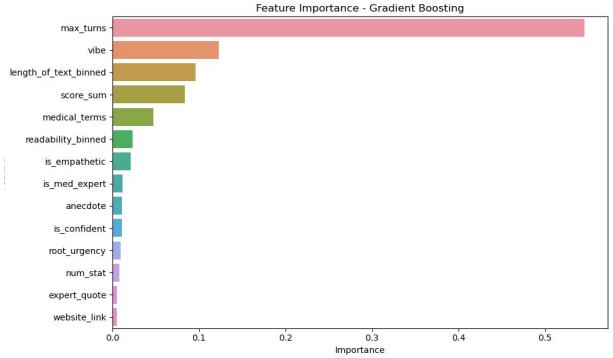
Model (Scenario/Run)	ROC Metric	F1 Metric
Before Refinement (Older Run with Erroneous Data)		
Random Forest (Scenario 1)	0.563 (ROC AUC)	0.814 (CV F1 Mean)
HistGradientBoosting (Scenario 2)	0.587 (ROC AUC)	0.362 (CV F1 Mean)
Random Forest (Scenario 3)	0.613 (ROC AUC)	0.542 (CV F1 Mean)
After Refinement (Run with Further Dropped Features, 11 Features)		
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)

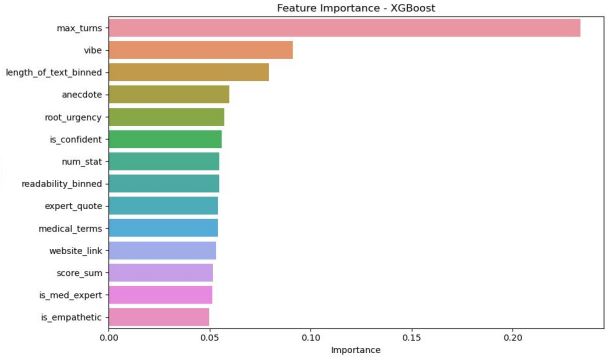
Table 7: Combined Model Performance Before and After Refinement

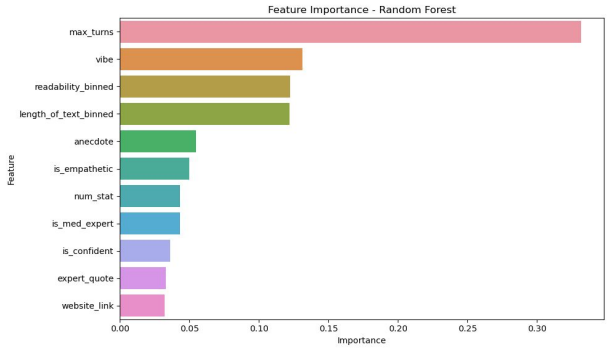
Refining the dataset and dropping low-impact features significantly improves model performance

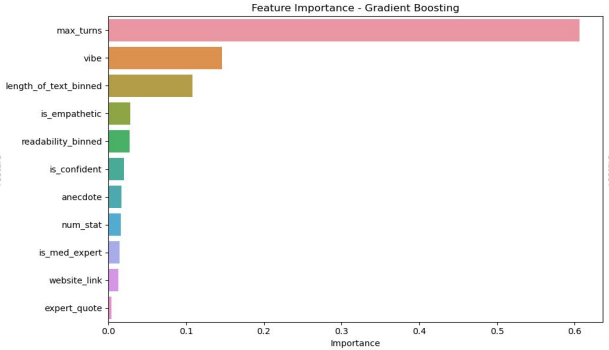
Result *Feature Importance*

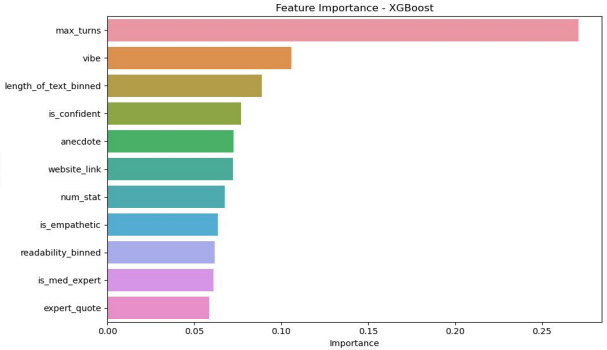






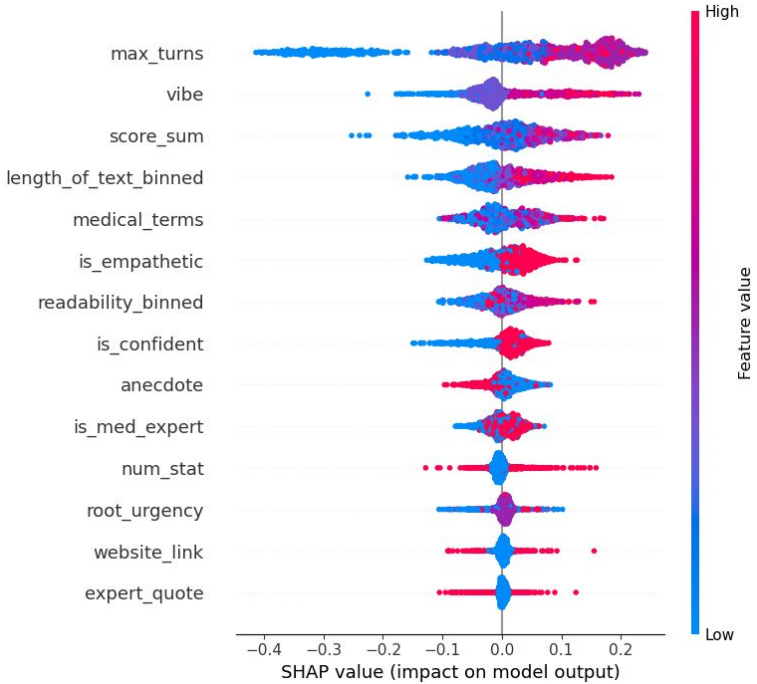


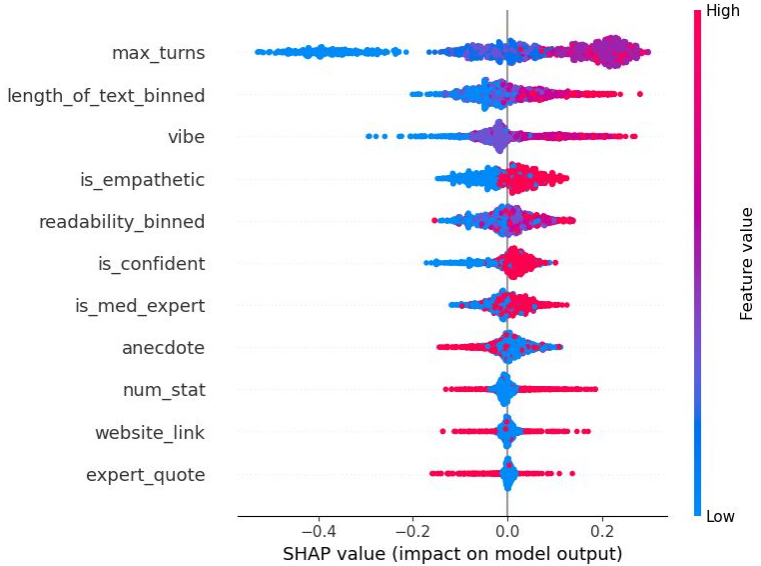




`Max_turns`, `vibe`, and `length_of_text_binned` are top predictors of resolution across Random Forest, Gradient Boosting, and XGBoost, with the reduced model (below) maintaining predictive power after dropping low-importance features

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*

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