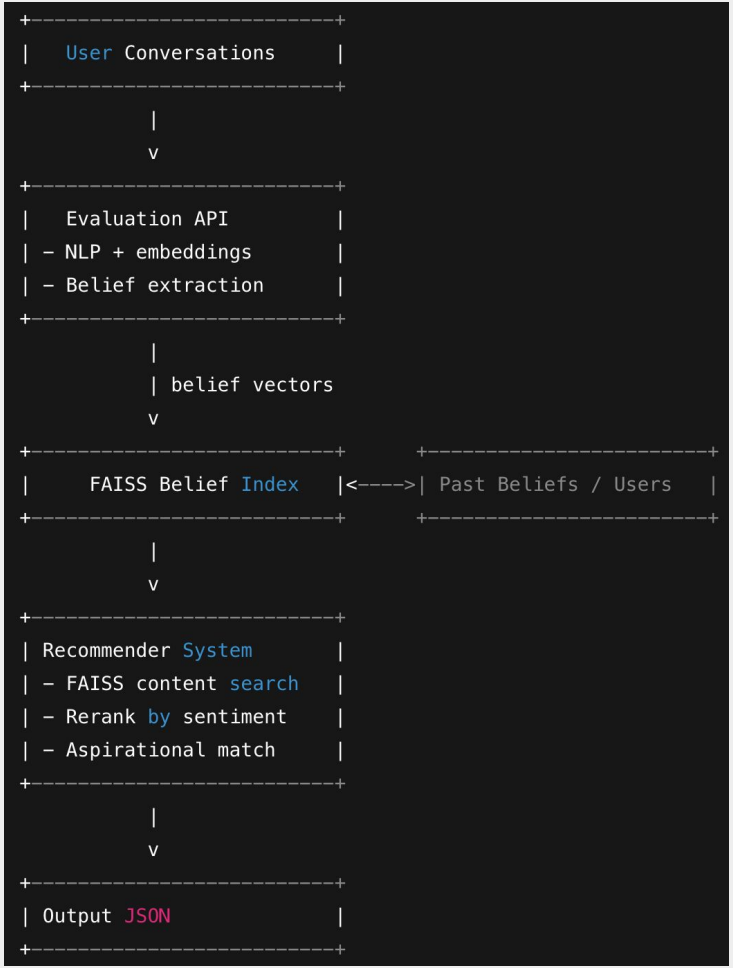
Conversational Evaluation + Recommendations

Nudging Toward Better Health Outcomes with StoryBot

Faran Sikandar
June 2025

Overall System Architecture



Metrics to Support Growth Across Multiple User + Biz Domains

- Ⓐ Assigning Value to Conversations
- Ⓑ Developing StoryBot
- Ⓒ Recommending Community Content
- Ⓓ Monitoring Anomalies in Conversation
- Ⓔ User’s Emotional Stability, Growth, Reflection
- Ⓕ Evaluating Outcomes + Results

Ⓐ Assigning Value to Conversations

Metric	Description	Complexity
Message count / round count	Number of messages or back-and-forth turns	Low
Total conversation length	Time between first and last message	Low
Average message length	Mean/median character or token count per message	Low
Emoji count	Total emojis used, or per-message rate	Low
Unique word count / richness	Lexical diversity or type-token ratio	Medium
Sentiment trajectory	Change in user sentiment from start to end	Medium
Belief shift (Δ identity vector)	Cosine distance between initial and final belief embeddings	High
Mood volatility	Standard deviation of sentiment across turns	Medium
Engagement index	Weighted sum: (length × sentiment × depth)	High (custom)
StoryBot impact on affect	Change in user sentiment before vs. after StoryBot replies	High

Ⓑ Developing StoryBot

Metric	Description	Complexity
StoryBot response latency	Time between user message and StoryBot reply	Low
Sentiment of StoryBot responses	Sentiment polarity/emotion classification	Medium
Relevance / on-topic response	Cosine similarity between user query and reply	Medium
User response rate after reply	Whether user replied (or how soon) after reply	Medium
Belief change attributable to bot	Δ belief between messages, conditioned on bot	High
Behavior-changing response score	Heuristic for motivational/reflective effect	High
Prompt injection defense coverage	Detected/blocked injections (pattern or LLM)	High

Ⓒ Recommending Community Content

Metric	Description	Complexity
Belief vector from conversation	Embedding of extracted beliefs or themes	Medium
Post similarity score (FAISS)	Cosine similarity between user and post	Medium
Topical coverage	Distribution of matched post topics	Medium
Identity alignment score	Alignment of content with beliefs	High
Reranking based on sentiment	Reordering by emotion/tone/value	High
Diversity & serendipity metrics	Novelty or variety vs. past content	High

Ⓓ Monitoring Anomalies in Conversation

Metric	Description	Complexity
Emoji usage rate shift	Change in emoji frequency	Low
Language style shift	Token length, syntax, punctuation	Medium
Mood volatility or spike	Jump/drop in sentiment/emotion	Medium
Topic drift or spike	Rare/unrelated topics appear	High
Prompt injection detection	Classifier for prompt hacks	High
Novel phrase frequency	Out-of-vocabulary or rare phrases	High

Ⓔ User’s Emotional Stability, Growth, Reflection

Metric	Use Case	Justification
Confusion / uncertainty signal	StoryBot / anomaly	"I don’t know", "I’m confused" → degraded clarity
Emotional coherence	Conversation value	Consistent tone = strong narrative
Reflective depth score	StoryBot	Pronouns × abstract language = introspection

Ⓕ Evaluating Outcomes / Results

Metric	Description	Complexity
Goal alignment score	Degree to which conversation aligns with user's goals	High
Change in emotional state	Difference in sentiment/emotion from start to end	Medium
Intent fulfillment rate	% of conversations where user intent was fulfilled	Medium
Behavior signal follow-up	Whether user took implied follow-up action	High
Longitudinal sentiment trend	Sentiment/emotion change across sessions	High
Return conversation rate	% of users returning for another conversation	Medium
Net affect improvement	Aggregate positive sentiment shift across users	High