

Assisting Breastfeeding and Maternity Experts in Responding to User Queries with an AI-in-the-loop Approach

Nadjet Bouayad-Agha, Ignasi Gomez Sebastià,
Alba Padró, Enric Pallarès,
David Pelayo and Rocío Tovar
firstname@lactapp.es
LactApp Women Health
Barcelona (Spain)

Abstract

Breastfeeding and Maternity experts are a scarce resource and engaging in a conversation with mothers on such a sensitive topic is a time-consuming effort. We present our journey and rationale in assisting experts to answer queries about Breastfeeding and Maternity topics from users, mainly mothers. We started by developing a RAG approach to response generation where the generated response is made available to the expert who has the option to draft an answer using the generated text or to answer from scratch. This was the start of an ongoing effort to develop a pipeline of AI/NLP-based functionalities to help experts understand user queries and craft their responses.

1 Context

Breastfeeding and Maternity experts are a scarce resource and engaging in a conversation with mothers on such a sensitive topic is a time-consuming effort. We present our effort in assisting experts to answer queries about Breastfeeding and Maternity topics from users, mainly mothers, using NLP and generative AI technology. From the user perspective, the interaction happens in a chat which is a premium service in an m-health application. Table 1 gives some statistics about the expert-user conversations for the first six months of 2024. The numbers show there is an average of 11 messages per conversation.¹

In order to answer a query, the experts follow a custom tailored protocol, which has been carefully drafted and validated by seasoned experts in the team and is recorded in a 10-page document. The document includes the chat service general philosophy, a step-by-step procedure on how to respond to a query as well as the do's and don't's for content

¹A conversation is roughly estimated as a sequence of messages typically happening on a single day, where the sequence starts with a user message and there is at least one user-expert interaction.

#messages	Experts	26479	47%
	Users	29293	53%
	Total	55772	100%
#conversations	Total	6165	100%
#users	Spanish	2561	94.3%
	Catalan	103	3.8%
	English	28	1%
	Others	23	0.9%
	Total	2714	100%

Table 1: Basic statistics about user-expert conversations in the first six months of 2024

as well as style. Example such guidelines include asking the user for minimal necessary information ("anamnesis"); avoiding "expert syndrome" that is offering a diagnosis before having sufficient information; and deriving the mother if the query is too complex, requires too much time to be attended online, or requires medical attention.

In addition, informal discussions and observation of some experts at work revealed that they spend a lot of time going through the lengthy user messages and previous conversations in order to gather key information.

We also learnt that experts informally keep a list of frequently used sentences, answers to common questions and list of blog post URL links at hand to help them craft their response.

In order to focus our efforts, we analyzed the types of messages that the experts receive from the users and identified five main types of user messages the experts might need help with, shown in Table 2, each of which requires a different pipeline of tasks to either help the expert interpret the query and/or generate a response.

We initially implemented a RAG response generation approach that applies to any incoming message above a minimum length. The generated response is made available to the expert who has the choice

query type	example incoming message
simple	Message about compatibility of some medication with breastfeeding
complex	Unhappy baby and mother, with issues of pain, poor weight gain and latching difficulties
multimodal	Weight-related question with weight tracker screenshot or tabular data
"here and now"	Query about current affairs such as IT problem, visits, workshops or follow-up to previous physical interaction
incomplete	User follow-up or response to current conversation

Table 2: Five main types of user interactions

to draft an answer using the generated text or to answer from scratch.

We quickly realized the limitations of our approach such as the poor outcome for complex queries or 'here and now' queries that require extremely fresh documents/information or the burden of having an answer generated for every incoming message regardless whether it contains a query. Furthermore, the approach did not address the need of the experts to parse the current conversation — skimming through a long single paragraph message that seems to have been written in one spurt — in order to sift the wheat from the chaff.

To address these issues, we then developed a pipeline of NLP/AI tasks to help the experts process the user queries and craft their responses.

In both cases, our philosophy is to provide support to our human experts with an 'AI-in-the-loop' (or AI co-piloting) approach. The conversation is still conducted by the experts who make or validate all the decisions so that trust and security — which are a fundamental requirement for user attention in this domain — are not compromised.

In the rest of the paper, for lack of space, we present the design and implementation of the multi-task pipeline (section 2), including the answer generation task initially used. We then briefly explain our initial deployment of a simple response generation in the expert environment and the deployment of the more advanced pipeline in a development environment (section 3). This is followed by the

evaluation of the answer generation task and a preliminary informal evaluation of the other tasks of the pipeline (section 4) before discussing related work (section 5) and drawing some conclusions regarding efforts so far and future work (section 6).

2 Design and implementation

For an agile development, we currently implemented each task as a rest API function using prompt-based generative AI in a zero-shot or in-context learning setting with some of the most performant commercial models on the market (i.e., Claude_x² and GPT_x³). The tasks are implemented using the Haystack Python API to define pipelines and components, to call models and query data stores.⁴ For document chunk retrieval used in answer generation and blog recommendation, we use the AWS OpenSearch data store.⁵

The tasks are pipelined together using an orchestrator, which performs the necessary checks and calls the different rest functions in order, concatenating the results of the different steps — including meta information such as cost, model and number of input and output tokens — in a final json structure.⁶

Figure 1 in the appendix illustrates the tasks flowchart and presents their basic implementation. More specifically, in order to address the first four user interactions in Table 2, we have implemented the following set of tasks:

Conversation detection. This task is about detecting whether the incoming user message starts a new dialogue or is the continuation of the previous conversation (answering a question, reacting to a previous message, elaborating on a previous point, etc). Currently user messages that do not start a new conversation are not processed (thus the last user interaction in Table 2 is ignored).

Text summarization. This task applies to 'complex' messages and involve detecting parts of the query that are about historical context (e.g., mother's previous pregnancies and health issues), current situation (e.g., what is the situation now),

²<https://aws.amazon.com/es/bedrock/claude/>

³<https://platform.openai.com/>

⁴<https://docs.haystack.deepset.ai/>

⁵<https://aws.amazon.com/opensearch-service/>

⁶<https://docs.aws.amazon.com/lambda/latest/operatorguide/orchestrator.html>

questions and emotions.⁷ Summarization is both abstractive and extractive and an example output is shown in table 10 (first two rows) in the appendix. Extracted highlights can be shown in the user message for the expert to verify their contextual relevance in-situ.

Intent detection. This task is about characterizing the query, that is what it is about, who is talking and who is it for. Thus we determine for each message 1) whether it contains a query; 2) whether it is a follow up to a visit, a workshop or some other event; 3) whether it is about a topic that is not breastfeeding (such as maternity in general or child rearing) and 4) who is the author of the message (health professional, friend, family, mother, etc.). We also identify main and secondary intent(s) in the message. The idea is that secondary intents are sub-issues that are related or collateral to the main issue, whilst main issues are independent of one another and must be addressed separately.

Table 11 in the appendix shows an example of input and output of intent detection. In the example, two main intents are identified: 1) baby's latching difficulties causing ongoing pain to mother (secondary intent) and identified as caused by frenulum (secondary intent) by pediatrician; and 2) blood spotting like a period by the mother.

There are currently around 60 possible intents with their optional definition/explanation, such as:

Compatibility: compatibility of products that the mother takes whilst breastfeeding medicines, vitamins, infusions, food supplements, aesthetic treatments, foods, drinks).

Shape of the breasts: such as hypoplasia; tubular breasts; smaller or larger breasts; asymmetrical breasts, soft breasts, one breast producing more than the other; flat, inverted or pierced nipples; breast augmentation, etc".

Information extraction. This is divided in two tasks: a general IE task to extract general information such as baby's age, mother's pregnancy, baby's prematurity; and an intent-specific IE task.

Currently we have two intent-specific IE tasks: one for compatibility and one for 'pain in the breast'. Compatibility IE task is about identifying the object of compatibility (medication, food, drink, product, body or health treatment and activity). Pain IE task is about identifying diagnosis (e.g., mastitis), treatments (e.g., antibiotics, apply-

ing cold, cabbage leaves), symptoms (e.g., fever, pink area, red line) and body zone (e.g., left breast, areola).

Table 11 in the appendix shows a visual representation of general (i.e., baby's age in mention "A month ago I had my baby") and pain-related information extraction. Whilst the model is able to identify different mentions of pain (e.g., "sensation of throbbing in the chest", "pain in the back", etc), more work needs to be done to refine the situations related to pain in the breast such as the identification of the cause of the pain (e.g., bad latching, frenulum, baby's teeth) and types of pain that could help determine an adequate diagnosis.

Machine Translation. This actually consists of two separate tasks: detection of source language and translation to Spanish, and translation of the generated text back to the source language. Although the current volume of queries that are not in Spanish is small, we envisage that this architecture will allow us to offer the chat functionality to other language communities.

Blog URL recommendation. This task uses a data store of 40k blog post chunks and retrieves the list of 100 chunks that are most similar to the query. The unique URLs of those chunks are identified and an average score is computed for each of them and up to 3 of the ones with the highest score are returned as recommendations.

Image-to-text extraction. We used the multi-modal capabilities of Claude 3 and gpt4-vision to extract baby growth data from a table image. This task is triggered if one of the query's main intent is about baby weight and the user adjoined an image. It is currently implemented in a two-step approach: 1) extraction and optional translation of table header, 2) extraction of baby growth data according to the table header.

Retrieval augmented generation. We use a RAG approach by retrieving the top 3 Q&A pairs whose questions are most similar to the incoming query. We currently have in store over 20k standalone Q&A pairs that were obtained by applying a conversation classification model trained on an automatically constructed dataset to detect conversation boundaries and selecting 'simple' conversations as our standalone Q&A dataset, that is, conversations that consist of only one sequence of consecutive user messages followed by only one sequence of consecutive expert messages.

An example input message with RAG, GENERATED and EXPERT responses is shown in Table 12

⁷The complexity of a message is currently determined as a threshold on the number of characters.

in the appendix. This input matches the one in Table 11 with two main intents. The example illustrates how the RAG answer is able to better address the user intents compared to the vanilla generation.

3 Deployment

We currently have a version that is deployed in the experts production environment which only consists in applying Retrieval Augmented Generation. The expert has the option to modify the generated message (i.e., it is presented in an editable text box). She also has the option to ignore the generated text altogether and draft her answer from scratch.

We had to address a number of impromptu issues including overload of model service or occasional noise in the generated json output string that required some preprocessing.

We also implemented the pipeline version described in section 2 and illustrated in Figure 1 in the appendix in a development environment. The orchestrator in this pipeline applies in order optional translation to Spanish, new conversation detection, summarization, intent detection either on extractive summary (if message is complex) or original query, general and intent-based information and image-to-text extraction, answer generation and optional translation of answer to source language, in addition to retrieval-based blog URL recommendation. The tasks are parameterized so as to enable the personalization of some results, such as the preferred languages of the expert (she may be able to attend queries in different languages), or the user source language (so as to translate the RAG generated response back into the user source language).

4 Initial evaluation and monitoring

In this section we first report on the evaluation and monitoring of our initial implementation, i.e., the answer generation task, before discussing preliminary evaluation of the other tasks.

4.1 Answer generation

In order to best calibrate the generation configuration, we evaluated the generation on a dataset of 100 randomly picked Q&As, the details of which are given in appendix A.⁸

Table 3 shows the semantic similarity of generated responses with expert responses for those 77 out of 100 responses that were

⁸The 100 Q&As of the dataset were obviously excluded from the Q&As RAG datastore.

rag?	semantic similarity				message avg length	
	mpnet2		ol3		exp	gen
	avg	med	avg	med		
yes	71.2	72.2	62.5	64.6	357	465
no	70.3	71.9	60.9	63.1		403

Table 3: Answer generation preliminary evaluation

rag-all = metrics on all 100 messages with attempted RAG,
rag-only = metrics on the 77 messages with retrieved documents,
ol3 = text-embedding-3-large,
mpnet2 = paraphrase-multilingual-mpnet-base-v2

generated with document retrieval augmentation. We computed the similarity using both the multilingual sentence transformer paraphrase-multilingual-mpnet-base-v2 (mpnet2) model (Reimers and Gurevych, 2019)⁹ and OpenAI Embeddings Large v3 (ol3) model¹⁰. The semantic similarity of responses is higher with RAG than without RAG. However, the automated response is longer than the expert response, especially for RAG.

For example, the RAG response in Table 12 in the appendix is more relevant to the user query than the vanilla response and this is reflected in the higher semantic similarity, which is 82.5% (RAG) vs 72.4% (vanilla) with mpnet2 and 69.1% (RAG) vs 66.6% (vanilla) with ol3. In addition, the number of characters in the original language is 705, 338 and 455 in ground truth, RAG and vanilla responses respectively.

4.2 Expert response monitoring

We perform continuous monitoring of the generation pipeline using two metrics. First we assess the expert messages conversion, that is the percentage of expert messages that were drafted from AI-generated responses. We also measure, for those messages that are converted, the mpnet2 semantic similarity between the generated and the final answer that is sent to the user.

Table 3 shows the message conversion and semantic similarity from 12th of February to 31st of May 2024 for response generation using only the Retrieval Augmented Generation (as mentioned in section 3) on all messages above a minimal size threshold.

The gap in similarity between February and March has to do with adjustments in the gener-

⁹https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

¹⁰<https://platform.openai.com/docs/models/embeddings>

month	#exp	#gen	%conv	sem sim
February	3523	270	8%	72.3%
March	3083	541	18%	78.3%
April	3887	762	20%	80.6%
May	4523	948	21%	80.2%

Table 4: Conversion rate and semantic similarity (02/12/2024 to 05/31/2024)

ative prompt. For example we instructed the model to generate a smaller answer and included word limits on the different parts of the message (such as validation), as the model tended to be too wordy. The increase in conversion has to do with the onboarding of the different experts.

We monitor these metrics on a daily basis and results are displayed in a visualization dashboard ¹¹. The idea is that we can see how our work impact the proportion of messages sent using the AI pipeline.

Thus, once our new pipeline is in production, the percentage of queries that can be converted will be based on 'answerable' queries only. However this is not the whole picture as ultimately we will want to know how the other tasks impact the experts in their work. This could be done using a time to response metric and/or by performing some live reviews/interviews with the experts.

4.3 Preliminary evaluation of other tasks

For intent detection, we manually annotated the 100 user messages in the dataset presented in appendix A and compared them with the predictions.

Tables 5 and 6 show some statistics and evaluation results for intent detection. Table 6 shows that although only 37% of text instances have their intent prediction fully matching the ground truth, this goes up to 91% instances having some match (i.e., partial+total match). Table 5 also reveals that the automatic detection tends to over-classify, e.g., 265 intents predicted vs 175 intents in ground truth overall.

Table 13 in the appendix presents the evaluation of the most predicted intents. Precision oscillates between 29% (for "extraction, conservation and preparation of maternal milk") and 86% (for "Baby rejects breast"). More work is needed with respect to the evaluation of the results (e.g., the distinction between main and secondary intents, intent coverage) and the refinement of intents specifications in order to improve precision.

¹¹<https://lookerstudio.google.com/>

		prediction	ground truth
intents	#	265	175
	#unique	47	41
per instance	%none	3	5
	max	8	5
	avg	3	2
	median	2	2

Table 5: Some statistics about intent evaluation data

w.avg precision	0.69
w.avg recall	0.89
%total match	37
%partial match	54
%no match	9

Table 6: Evaluation of intent detection

Given the complexity and subjectivity of summarization evaluation (Akkasi et al., 2023), we opted for a goal-oriented automated evaluation. We picked the 25 "complex" user queries, that is, queries over 500 characters from the 100 query dataset (appendix A) and generated their summaries. We used these summaries as input to generate an answer to the query (gen-sum answers). Given answers from full inputs (gen-full answers) and expert answers, we computed $\text{similarity}(\text{gen-sum}, \text{expert})$ using mpnet2 metric and compared it with $\text{similarity}(\text{gen-full}, \text{expert})$ computed using the same approach. Both gen-full and gen-sum were generated without retrieval augmentation, using the same generation parameters (see table 14 in the appendix) but the prompt for gen-sum was slightly modified to describe the input format, an example of which is shown in the third row of table 10 in the appendix.

The results of the summarization evaluation are shown in table 7 and the size of responses and inputs are shown in table 8. The results show that the semantic similarity of gen-sum with expert answers approaches that of the semantic similarity of gen-full whilst the average size of gen-sum answer is closer to that of expert answer compared to gen-full.

We evaluated information extraction on messages from our 100 query dataset (appendix A) with specific ground truth intents, namely 12 messages about "pain in the breast", 9 messages about "compatibility" and 3 messages with both "pain in

	min	median	avg	max
gen-full vs expert	0.34	0.68	0.64	0.85
gen-sum vs expert	0.27	0.67	0.64	0.89
sum vs full input	0.60	0.77	0.77	0.89

Table 7: Semantic similarity between expert answers and answers generated from full input (gen-full) or from summary (gen-sum); and between summary and full input (computed over 25 instances)

	min	median	avg	max
gen-sum	379	469	479	602
gen-full	565	701	812	1738
expert	100	406	448	1183
full input	503	805	841	1647
summary	398	618	668	958

Table 8: Size (in # characters) of answers generated from summary (gen-sum) and from full input (gen-full); and size of expert response, full input message and summary input

the breast" and "compatibility". The evaluation was done by taking into account partial matches, that is overlapping mentions. For example, mention "21 weeks pregnant" may be identified as "pregnancy" entity whilst in the ground truth, the entity just spans the smaller mention "pregnant".

The results of information extraction evaluation are presented in table 9. Half of the false positives have to do with body parts detected in segments of texts that were not about pain, so detecting body parts in this way is probably too simplistic. Also, we found that the concept of pain as a symptom comes in all sorts of variations or circumstances: pain when sleeping face down, pain when pressing hard, pain when breastfeeding. This is important for the expert for determining the issue.

Regarding image-to-text extraction of baby growth information, we found it only works well

tp	73
fp	20
fn	3
precision	0.78
recall	0.96

Table 9: Evaluation of Information Extraction

for good quality snapshots of digital tables (tables from online trackers for example) but gives poor results when snapshot is taken with poor lighting and angle, and the table contains manuscript data. Thus a more robust approach is needed such as training our own image-to-text extraction model. The experts also explained that they sometimes ask the user to send growth data which the user obliges but as text in tabular format, so detecting this information in-situ in the text and rendering it in a table and eventually a graph is also another requirement.

For the evaluation of conversation detection and URL recommendation, we looked at 70 users and their 538 messages during a given period and evaluated first conversation detection and then blog recommendation on the first messages of each of the 104 true conversations. For conversation detection, we got a precision of 81% and a recall of 95%. For blog URL recommendation, we performed a strict evaluation where every recommended URL is evaluated and a loose evaluation where a true positive is when at least one of the recommended URLs is correct. With strict evaluation, we get a precision of 27% and a recall of 94%. With loose evaluation, we get a precision of 67% and a recall of 100%, so there is room for improvement.

5 Related work

Although task-oriented chatbots and virtual agents have been at the forefront of AI and NLP applications and research for many years, for many domains this implementation remain challenging and costly and its adoption met with dissatisfaction or mistrust (Kraus et al., 2023). The relatively recent advent of Large Language Models (LLMs) and so-called *Generative AI* has brought new promises but also new challenges such as hallucinations and poor relevance.

In the healthcare domain, several approaches have been used to mitigate those issues such as LLM fine tuning to adapt to diagnosis style and prompt engineering to improve consistency (Shi et al., 2024), or applying knowledge- and NLP-intensive approaches such as Xia et al (2022) who combine symptom recognition and disambiguation and knowledge graph reasoning (which they call 'triage') before performing an entity-aware prompt-based generation.

Other approaches aim to assist healthcare professionals instead of replacing them. For example, Madeira et al (2020) provide chat operators of a

mental healthcare service with query classification and a list of suggestions to be discussed. Xie et al (2024) investigate how LLMs can help doctors in daily tasks that are "repetitive [by] nature (e.g., case summarization, preoperative education), relatively low medical risk (e.g., triage), [or require]... extensive information requirements (e.g., medication inquiry)."

6 Conclusions and Future Work

We have presented our ongoing journey into developing AI-driven functionality to assist experts in addressing user queries about maternity and breastfeeding. Starting with a RAG approach, we gave our experts the option to draft their response from a generated text. The monitoring and feedback received allowed us to quickly realize that not every user message could be treated equally, so we followed up with a more complex pipeline for the conditional generation of answers, where we could guarantee a higher relevance, coverage and faithfulness (Es et al., 2024). We also realized the need to help the experts not only draft their response but also understand the current and past conversation and so we expanded the pipeline with understanding tasks such as information extraction .

Our approach is to incrementally put in place and test a set of functionalities that can work for our experts. In doing so we must take into account the following criteria:

Cost. Proprietary LLMs are costly. Whilst those out-of-the-box models allow us to quickly get a grasp of the workability of our pipeline, we consider implementing some of our own models in the future for certain tasks, such as intent detection (because it has such as large input prompt).

Trust. One of the most-valued features of the app and associated chat is the trust it generates and builds amongst our users and this is something that cannot be compromised. This is why we favor extractive understanding through summarization and information extraction, so that the expert can always see the information in context and hence trust its veracity. Trust is something we always need to keep in mind when developing our system.

Accuracy. A large proportion of user queries is complex because they involve a personal history with all its contingencies and sometimes stem from the mother's need to express herself (and often her desperation) and feel understood and validated. For those queries, retrieval may be poor and so the

initial solution is to help experts understand the query, though eventually, it could be processed and become more manageable.

We currently have several fronts to pursue the integration of the AI-pipeline. Firstly, though an initial version of the UI has been developed that integrates the AI functionalities, it needs more work to get usable and work for the experts. This interface should include a feedback system, in which the expert can signal, at least minimally, any issues with the information she is given. Some of the tasks, such as generating a growth table and graph from tabular data or a screenshot, should be performed on demand whilst others should be triggered as the messages arrive as there is some latency involved.

Secondly we need to improve the accuracy and coverage of some of our tasks. For example, we could include more templates for extracting information about other intents. We also need to test and refine them with the help of our experts. For answer generation, we have several pending tasks such as: hybrid retrieval and reranking, fine-tuning an LLM to adjust better to the experts verbal diagnosing style (Shi et al., 2024), and incorporating meta-data for document filtering (Gao et al., 2023). We also need to address fact-checking in order to minimize expert's edition of the answer, such as discrepancies between the response and the user message or an incorrect diagnosis or suggestion (Vishwanath et al., 2024).

Thirdly we can incrementally add new tasks to address expert needs as they emerge. For example we are currently working on summarizing the user conversation history which is something that experts spend a lot of time doing for recurring users. We are also working on message concatenation, because sometimes user input arrives in several installments.

Finally, we are considering a rule-based approach to generating minimal follow-up questions on some intents to gather missing information: for example asking weight or age of the baby if needed to answer a query about, say compatibility of breastfeeding with medication.

References

Abbas Akkasi, Kathleen Fraser, and Majid Komeili. 2023. [Reference-free summarization evaluation with large language models](#). In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP*

- Systems*, pages 193–201, Bali, Indonesia. Association for Computational Linguistics.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158, St. Julians, Malta. Association for Computational Linguistics.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *Published in ArXiv*.
- Sarit Kraus, Yaniv Oshrat, Yonatan Aumann, Tal Hollander, Oleg Maksimov, Anita Ostroumov, and Natali Shechtman. 2023. **Customer service combining human operators and virtual agents: a call for multi-disciplinary ai research**. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’23/IAAI’23/EAAI’23. AAAI Press.
- Thiago Madeira, Heder Bernardino, Jairo Francisco De Souza, Henrique Gomide, Nathália Munck Machado, Bruno Marcos Pinheiro da Silva, and Alexandre Vieira Pereira Pacelli. 2020. **A framework to assist chat operators of mental healthcare services**. In *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–7, Online. Association for Computational Linguistics.
- Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. 2019. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. **Medical dialogue system: A survey of categories, methods, evaluation and challenges**. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2840–2861, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.
- Prathiksha Rumale Vishwanath, Simran Tiwari, Tejas Ganesh Naik, Sahil Gupta, Dung Ngoc Thai, Wenlong Zhao, SUNJAE KWON, Victor Ardulov, Karim Tarabishy, Andrew McCallum, et al. 2024. Faithfulness hallucination detection in healthcare ai. In *KDD Workshop on Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare*, Barcelona, Spain.
- Fei Xia, Bin Li, Yixuan Weng, Shizhu He, Kang Liu, Bin Sun, Shutao Li, and Jun Zhao. 2022. **Med-ConQA: Medical conversational question answering system based on knowledge graphs**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 148–158, Abu Dhabi, UAE. Association for Computational Linguistics.
- Wenya Xie, Qingying Xiao, Yu Zheng, Xidong Wang, Junying Chen, Ke Ji, Anningzhe Gao, Xiang Wan, Feng Jiang, and Benyou Wang. 2024. LLMs for doctors: Leveraging medical llms to assist doctors, not replace them. *arXiv preprint arXiv:2406.18034*.

A Dataset used for evaluation

For the evaluation of the different tasks we use a dataset of 100 user queries that were randomly picked from a larger dataset of ‘hard queries’. These hard queries originated from a legacy dataset of a large number of queries classified by one of our most seasoned experts according to 30 possible topics. A hard query was either a query that the expert could not classify (for example because the topic did not match any of the available options) or a ‘noisy query’ as determined by the Cleanlab package (Northcutt et al., 2019)¹². Examples of noisy queries included queries with more than one topic, or complex queries with overlapping messy topics (like the primary and secondary intents we try to distinguish). The small evaluation dataset was made up of an equal number of noisy and non classifiable datasets.

B Examples of input-output

Table 10 shows a user message in the first row with highlights as identified by the summarization task, the output of which is shown in the second row.¹³ The third row is the input to the generation experiment for summarization evaluation presented in section 4 where contents of each summary section are aggregated so that an answer can be generated given the summary.¹⁴

Table 11 shows an example of information extraction output (in annotation tool) and intent detection (as a json).

Table 12 shows an example of a response to the query in 11 generated with and without retrieval augmentation, as well as the expert response. In the

¹²<https://github.com/cleanlab/cleanlab>

¹³The annotation tool used is Label Studio.

¹⁴All input texts thereafter are pseudonymized, translated into English and some are truncated ([..]). The json outputs are given without the metadata.

RAG case, 3 Q&A pairs are retrieved, two regarding the first main intent (latching and pain) and one regarding the second intent (menstruation).

C Sample per intent detection evaluation

Table 13 shows a sample of intent detection evaluation for the top 5 most assigned labels.

D LLM model parameters for the different tasks

Table 14 shows the LLM model parameters for the different tasks implemented in the pipeline. The choice is determined by availability, capabilities, cost and performance at the time of implementation. For example, GPT4O is cheaper than GPT4 but was showing less accuracy so we use it for less inference heavy tasks such as conversation detection or translation.

E Message processing pipeline

Figure 1 shows the decision flowchart of the pipeline, from the user input flagged as beginning a new conversation to the eventual generation of a response and intermediate steps such as image-to-text, summarization, intent detection, RAG or question generation. All the tasks in purple are currently implemented in a zero-shot or in-context learning setting with custom-made prompts and off-the-shelf models detailed in table 14.

<p>background 1 sentiments 2 current situation 3 questions 4</p> <p>1 Hello everyone, I really feel misunderstood. My baby turns 2 in April. She is so demanding with the breast that we have trouble sleeping lately. She breastfeeds every 1.30 hours, it is exhausting. At first I put up with it but I really need to sleep. And as happens in some places, I have a husband. The baby's father. But he blames me for getting her used to it. That it is a habit because the baby is with me all day and at night I co-sleep with her. I am seeing what I am doing wrong that everyone, including my husband, criticizes me. I stopped working to be with her.</p> <p>2 With the oldest, weaning happened overnight. At 18 months. She was mixed feeding. And I started giving her a bottle at night before going to sleep and she would fall asleep until the next day.</p> <p>3 With the youngest it is very difficult, she takes a bottle but she is very much under my wing. My husband says he can't sleep because he works... I can't sleep either, but according to him I'm a housewife and I'm with the girls all day and I can go to bed for a nap in the afternoon. The reality is that there are times when I can and there are times when I can't because I clean, tidy, cook, educate and take care of my daughters, etc. He thinks that my job is easy... He always belittles me.</p> <p>4 I think it's the beginning of weaning her, but she's in that crisis where she undresses me anywhere and hits me, I don't know what to do. Really. I think it's my fault for continuing to breastfeed on demand. She can be asking me every 10 minutes. And if I don't give it to her she gets angry... I'm trying to make weaning work so I can rest a little better. Even when she had pneumonia I gave her the breast. But I feel like my body is tired.</p> <p>5 By the way, I want you to know that there is no pregnant friend to whom I would not give your book about breastfeeding</p>	<pre>{'questions': { 'summary': ['How can I start weaning my daughter who is very demanding ?', 'How can I handle the situation with my husband who blames me for our daughter's dependence?'], 'highlight': ['I think it's the beginning of weaning her, but she's in that crisis where she undresses me anywhere and hits me, I don't know what to do.', 'I'm seeing what I'm doing wrong that everyone, including my husband, criticizes me.']], 'background': { 'summary': ['The mother stopped working to care for her daughters, the youngest of whom is very dependent on breastfeeding.', 'The girl's father blames the mother for the girl's dependence.'], 'highlight': ['I stopped working to be with her.', 'But he blames me for getting her used to it.']], 'current situation' { 'summary': ['Mother is exhausted due to her daughter's constant breastfeeding demands and lack of sleep.', 'She is trying to initiate weaning but is facing resistance from her daughter.'], 'highlight': ['She breastfeeds every 1.30 hrs, it's exhausting.', 'I'm trying to make weaning work so I can rest a little better.']], 'sentiments': { 'summary': ['frustration', 'tiredness', 'guilt', 'incomprehension'], 'highlight': ['I really feel misunderstood.', 'I think it's my fault for continuing to breastfeed on demand.', 'He thinks that my job is easy... He always belittles me.', 'But I feel like my body is tired .']]}}</pre>
<ul style="list-style-type: none"> - Background: The mother stopped working to care for her daughters, the youngest of whom is very dependent on breastfeeding, The girls father blames the mother for the girls dependence - Current situation: Mother is exhausted due to her daughter's constant breastfeeding demands and lack of sleep; She is trying to initiate weaning but is facing resistance from her daughter. - Questions: How can I start weaning my daughter who is very demanding?; How can I handle the situation with my husband who blames mefor our daughter's dependence? - Sentiments: frustration, tiredness, guilt, incomprehension 	

Table 10: Input and outputs of summarization

pain:diagnosis 1	pain:method_diagnosis 2	pain:body_zone 3	pain:treatment 4	pain:symptom 5	general:baby_age 6
<p>1 A month ago I had my baby by cesarean section, since she was born she has been breastfeeding, and from the beginning she hurt my nipple a lot and she was not sucking well because my breasts were lumpy, so the pediatrician opted for the option of cutting the frenulum. It is true that it has improved somewhat but the pain and the sensation of throbbing in the chest and pain in the back is still there. On top of that, I am worried because since yesterday I started spotting like a period, not in abundance but an intense red color, can it be menstruation even though I am breastfeeding? This whole situation has me desperate and worried.</p>					
<pre> {"has_query": true, "is_follow_up": false, "author": "mother", "other_theme": false, "intents": [{"intent_name": "Latching difficulties, postures, suction", "cluster_id": "1", "main": true} {"intent_name": "Pain in the breast", "cluster_id": "1", "main": false}, {"intent_name": "Frenulum", "cluster_id": "1", "main": false}, {"intent_name": "Menstruation y fertility", "cluster_id": "2", "main": true }]] </pre>					

Table 11: Input and outputs of information extraction and intent detection

Input	A month ago I had my baby by cesarean section, since she was born she has been breastfeeding, and from the beginning she hurt my nipple a lot and she was not sucking well because my breasts were lumpy, so the pediatrician opted for the option of cutting the frenulum. It is true that it has improved somewhat but the pain and the sensation of throbbing in the chest and pain in the back is still there. On top of that, I am worried because since yesterday I started spotting like a period, not in abundance but an intense red color, can it be menstruation even though I am breastfeeding? This whole situation has me desperate and worried.
Gen	Hello! Congratulations on your baby. I understand that you are going through a difficult time. It is normal for your period to return even if you are breastfeeding, especially after the first month. Sore nipples and back pain can be due to poor breastfeeding posture or a breast infection. I would recommend that you see a lactation specialist or your doctor for an examination and the best advice.
RAG	Hello! Congratulations on your baby. I understand that you feel desperate and worried, it is normal. Pain and stinging can be normal after cutting the frenulum, but if they persist, I would recommend that you see a lactation consultant. As for bleeding, up to 52 days postpartum it is normal to have bleeding that looks like menstruation, but it is the end of the blood residue that was in the uterus. If it is very abundant or has a bad smell, make an appointment with a gynecologist.
Expert	It's not easy to breastfeed with pain... When you can, could you explain to me in a little more detail what the pain is like and when you feel it? How is it going with the weight of the baby, is it gaining weight correctly? If you want, we could try to check the latch with a video from here, I don't know if that would be okay with you. As for the bleeding you mentioned, it can be normal. Within 52 days after giving birth, heavy bleeding similar to menstruation can often appear, which is known as "partillo". If you are breastfeeding on demand and exclusively and the bleeding does not have a bad smell, this is most likely what it is. It is not a period, because it is not accompanied by ovulation.

Table 12: Expert, RAG and Generated Response for Input Query

	Label	#	TP	FP	TN	FN	precision
1	Pain in the breast	22	16	6	78	0	0.73
4	Extraction, conservation and preparation of maternal milk	17	5	12	83	0	0.29
3	Compatibility	16	11	5	84	0	0.69
6	Latching, posture and suction difficulties	15	8	7	84	1	0.53
2	Baby rejects breast	14	12	2	86	0	0.86

Table 13: Sample per intent evaluation

task	model and params
conversation detection	gpt-4o-2024-05-13
generation	gpt-4-0613
image extraction	anthropic.claude-3-sonnet-20240229-v1:0 max_tokens=5000
information extraction	gpt-4o-2024-05-13 max_tokens=1000
intent detection	gpt-4-0613
retrieval (generation)	cohere.embed-multilingual-v3 retrieval threshold = 0.65 retriever_top_k=3
retrieval (blog recommendation)	text-embedding-3-large retrieval threshold = 0.5 retriever_top_k=100
summarization	gpt-4-0613
translate	gpt-4o-2024-05-13

Table 14: LLM Generation Tasks Model Parameters
(unless otherwise indicated, temperature for generative models is 0.1 and maximum token length is 500)

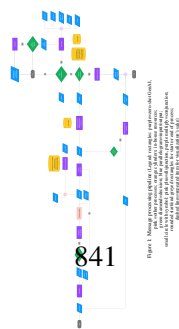


Figure 1. Diagram of a neural network architecture for a classification task.