# LLM-Supported Parenting Chatbot

## 1. Executive summary

This project introduces an advanced, language learning model (LLM)-supported chatbot designed to provide accurate, clear, and accessible parenting advice from newborn care up to the child's first year. The chatbot utilizes reliable information from sources like the CDC and Just in Time Parenting, aiming to be a dependable guide for parents.

The main goal is to offer a user-friendly tool that provides age-specific solutions to parenting questions, ensuring content is understandable regardless of the user's background or reading proficiency. The chatbot was evaluated using a set of 10 FAQs from Northwest Family Clinics, achieving a 70% success rate in accurately answering queries, with further assessments on readability confirming its outputs are suitable for a 7th-grade reading level.

This executive summary underscores the development and testing of a digital parenting aid, highlighting its potential to transform access to parenting advice, with ongoing improvements recommended to expand its knowledge base and enhance user interaction.

## 2. Goals of the project, data, and evaluation methods

The vast amount of parenting information available today is often disorganized, scattered, and comes from uncertain sources. This can pose significant challenges for parents, especially those new to parenting, who are seeking reliable and understandable advice tailored to their specific needs. This project addresses these challenges by developing a language learning model (LLM)-supported chatbot designed to provide accurate, comprehensible answers to parenting questions, starting from newborn care up to the child’s first year.

The primary goal is to develop a chatbot that leverages information from credible sources to provide supportive, age-specific solutions that are relevant and practical for parents at various stages of their child’s early development. To narrow down the scope, the chatbot focuses on parenting information for parents of children aged 0-1. To ensure the chatbot is accessible to all users, responses are crafted to be easily understandable.

To provide parents with credible information, data for this project were carefully selected from reputable organizations, including the Centers for Disease Control and Prevention (CDC) and respected online parenting resources like Just in Time Parenting and HealthyParents.org. These sources provide a solid foundation of verified and reliable information that underpins the chatbot’s knowledge base.

To evaluate the chatbot's fidelity, a rigorous assessment was conducted focusing on its accuracy in addressing 10 frequently asked questions sourced from Northwest Family Clinics. This clinic is renowned for its comprehensive parenting resources, and its FAQs are prominently ranked third in organic search results for "parenting FAQ for baby," indicating their widespread relevance and applicability. This selection of FAQs provides a realistic benchmark for testing the chatbot’s capabilities, aiming to refine the tool to better meet the real-world needs of parents seeking reliable support.

The evaluation process included both manual reviews and automated comparisons using GPT-4. For baseline comparisons, the same set of FAQs was input into GPT-4o-mini, a freely available model, to generate answers without specialized prompt engineering. The responses from GPT-4o-mini were then juxtaposed with those from the chatbot and the original FAQ answers to provide a comprehensive view of the chatbot’s performance relative to both a standard AI model and expert-generated content.

In addition to accuracy, readability was another crucial evaluation metric. The New Dale-Chall Readability Formula was used to assess the comprehensibility of the chatbot's responses. The New Dale-Chall formula is particularly suited for this purpose as it incorporates a list of 3,000 words that are commonly understood by 4th-grade students, providing a reliable measure of whether the text is suitable for the general public. This aspect of the evaluation helps confirm that the chatbot's outputs are easily understood by parents, regardless of their reading proficiency, thereby enhancing the usability and effectiveness of the digital parenting tool.

## 3. GenAI application design

A screenshot of a chat

Description automatically generated*Image 1: system design of the chatbot in Dify*

The information resource was segmented into four distinct databases, each aligned with specific content categories relevant to child development, child care, health and well-being, and food and nutrition, as classified by UNICEF. It is possible for some documents to be cross listed across more than one database due to overlapping content. The system is designed with four specialized agents, each dedicated to one of these databases, ensuring targeted and relevant responses.

Upon receiving a user query, the system first employs a question classifier to determine the most appropriate agent to handle the question based on its content. Once classified, the query is forwarded to the selected agent. The response process begins with the agent utilizing a language learning model (LLM) node, which employs prompt engineering techniques to generate an initial answer based on the information retrieved from the corresponding database.

Following the initial response generation, the answer undergoes further refinement by another LLM node. This node is specifically tasked with adjusting the response to ensure it maintains a positive tone and is easily understandable at an elementary reading level, all while preserving the original meaning of the content. The final output from this LLM node is then presented to the user as plain text. For example, in Image 2, the chatbot answers the user input query "Do I Need to Wash Baby’s Clothes Before Their First Use?”.

Throughout the system, all LLM nodes are powered by the Gemini 1.5 Flash 001 model. This setup was chosen to evaluate whether freely available models could meet the required standards for accuracy, readability, and tone, thus making the application more accessible to a wider audience.

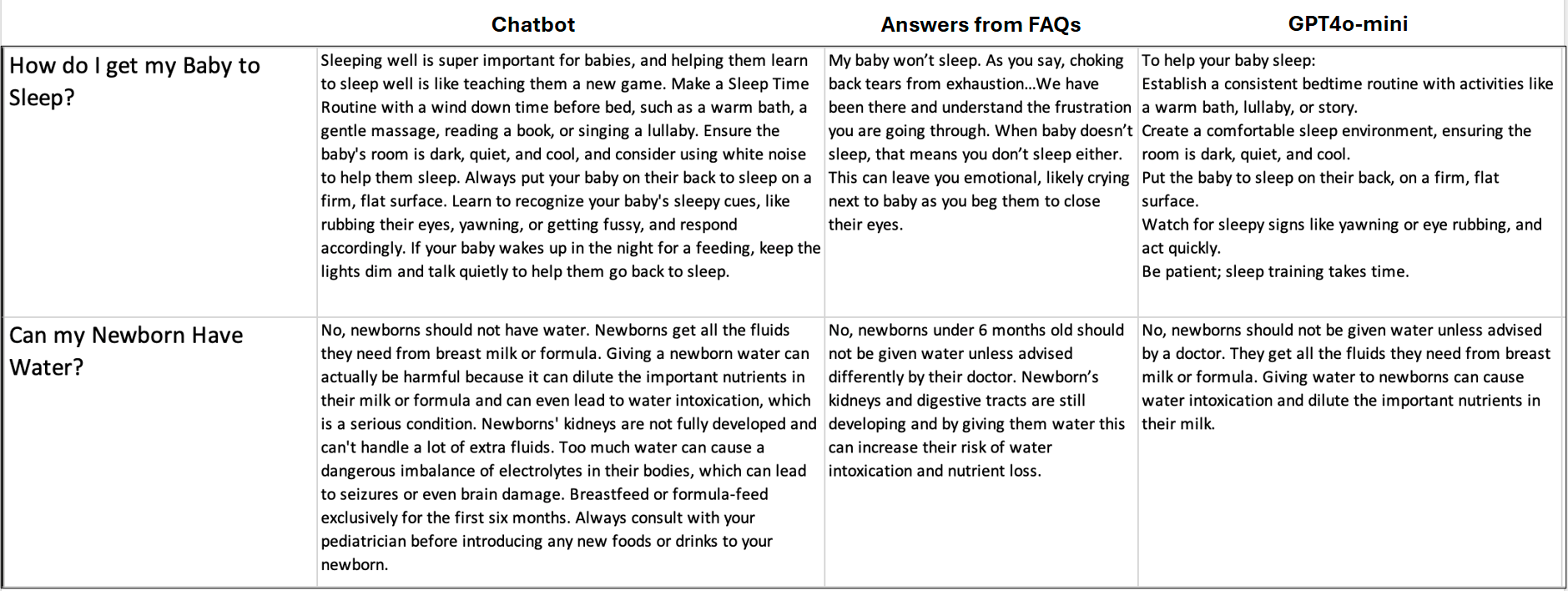
A white text on a white background

Description automatically generated*Image 2: Example Answer to "Do I Need to Wash Baby’s Clothes Before Their First Use?*

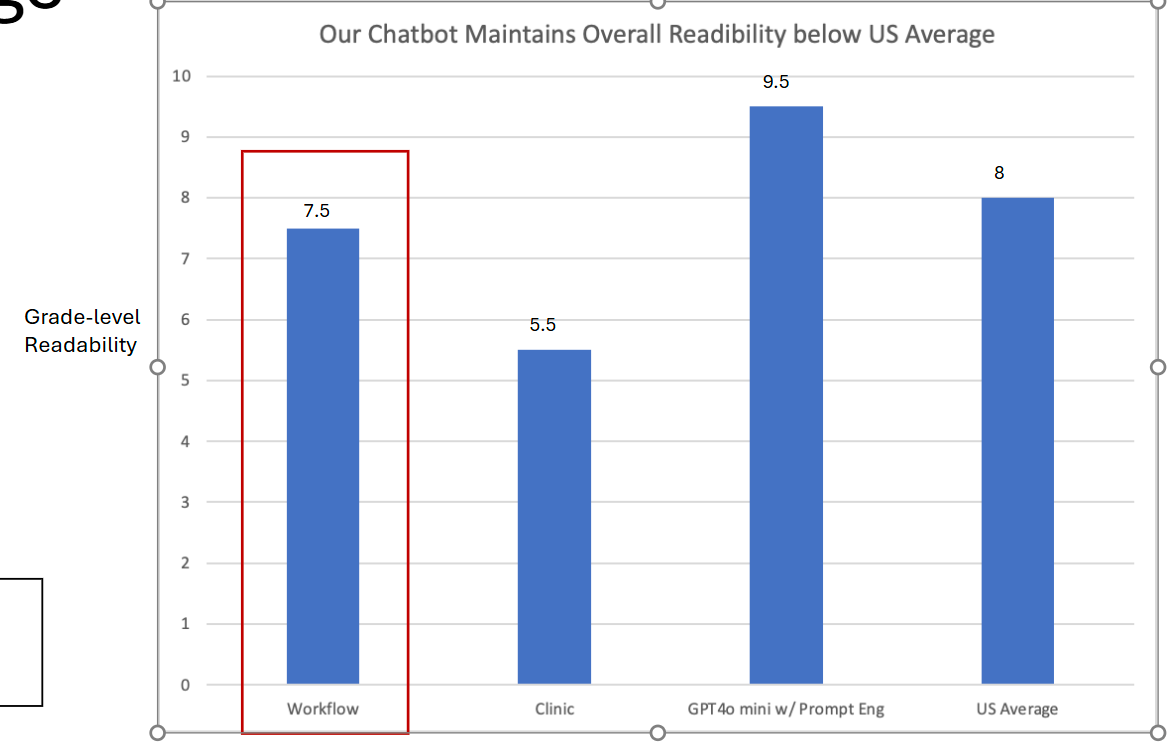
## 4. Evaluation using data and main findings

In the evaluation of the chatbot's accuracy, a set of 10 parenting FAQs was used to test the system. The responses generated by the chatbot were compared with the expert answers provided in the FAQs. The chatbot successfully answered 7 out of the 10 questions; the remaining three questions were not answered correctly as they fell outside the scope of the chatbot's knowledge base, indicating areas where the chatbot lacked information. Answers from GPT4o-mini without prompt engineering were also included as the baseline.

For the questions that were answered successfully, a comprehensive assessment was carried out involving both manual review and automated comparison using GPT-4. This dual-method evaluation confirmed that the chatbot's responses were consistently accurate, achieving a 100% match rate with the answers in the FAQs and those generated by GPT-4o-mini. The precision of the chatbot's answers underscores its effectiveness in handling typical parenting queries. Details and examples of these comparisons are illustrated in Image 3, showcasing the chatbot’s performance.

*Image 3: Example answer comparison among the Chatbot, the FAQs, and GPT4o-mini*

Another key evaluation metric was readability. Utilizing the New Dale-Chall Readability Formula, grade-level readability scores were generated for each of the chatbot's responses. The average readability score across the seven accurately answered questions was approximately 7.5, indicating that the content is suitable for 7th-grade students. This level of readability is more accessible than the outputs from GPT-4o-mini without prompt engineering, and also lower than the US national average readability score of 8. These findings, detailed in Image 4, demonstrate the chatbot’s ability to provide answers that are easily understandable, enhancing its usability for a broader audience.



*Image 4: Readability Scores among answers by Chatbot, FAQs, and GPT4o-mini*

## 5. Managerial insights

The development and evaluation of the LLM-supported chatbot for answering parenting questions provide valuable insights into the potential for digital parenting solutions. The chatbot's ability to accurately answer questions underscores the viability of using free tier models like Gemini 1.5 Flash 001, complemented by a sophisticatedly designed knowledge base.

A key observation from the evaluation is the importance of the scope and depth of the knowledge base. The chatbot's failure to answer three out of ten questions highlights the necessity for a more comprehensive database that covers a wider range of parenting topics, including less common queries. This indicates that expanding the chatbot's data sources or integrating a more diverse dataset could significantly enhance its utility. Additionally, designing a mechanism to inform users when their queries fall outside the chatbot’s current knowledge base could improve user experience and manage expectations.

Furthermore, the readability results, with the chatbot achieving an average score suitable for 7th-grade students, demonstrate the feasibility of crafting content that is easily understandable. To further enhance usability, future iterations could implement a feedback loop that continuously tests and refines the output to meet specific readability standards, ensuring accessibility for a broader audience.

A noticeable limitation in the current design is the lack of interaction among the agents. Enabling communication and data sharing between the agents could lead to more comprehensive responses and a more integrated user experience.

Additionally, the evaluation process could benefit from more sophisticated methodologies, such as retrieval-augmented generation (RAG) evaluations. This would provide a deeper understanding of the chatbot's performance and help pinpoint areas for improvement in both the technology and content delivery, ultimately leading to a more effective and reliable digital parenting chatbot.

## 6. Sources for more information, including a bibliography on the subject

Readability score calculation:

<https://readabilityformulas.com/readability-scoring-system.php>

10 FAQs from Northwest Family Clinics:

<https://www.northwestfamilyclinics.com/blog/10-questions-every-new-parent-has-about-caring-their-newborn>

Healthychildren.org:

<https://www.healthychildren.org/English/Pages/default.aspx>

Just in Time Parenting:

<https://jitp.info/>

UNICF Parenting:

<https://www.unicef.org/parenting/>

CDC Parenting:

<https://www.cdc.gov/child-development/positive-parenting-tips/index.html>