A Study on Integrating Retrieval-Augmented Generation with Large Language Model for Consulting Support in Development and Mental Health of Children Under 6 Years Old

Thanh Nguyen Van Quoc [1], Hao Nguyen Thi Bich [2], Nhut Nguyen Minh[3], Thuan Nguyen Dinh[4]
Faculty of Information Systems
University of Information Technology - Vietnam National University
Ho Chi Minh City, Vietnam
{21521447@gm.uit.edu.vn, 21522049@gm.uit.edu.vn, nhutnm.17@grad.uit.edu.vn, thuannd@uit.edu.vn}

***Abstract*—Currently, the need for psychological health as well as parents' concerns about the rate of development of children is very high due to the growing number of cases of autism, Autism Spectrum Disorders (ASD), developmental delays have been discovered in recent years. However, most people are not well aware or well-gathered about this issue. Therefore, parents or relatives of the child have not yet given a correct objective assessment of these diseases. The use of the RAG framework, in conjunction with LangChain and using a Large Language Model (LLM) will help people learn and receive better results about mental health-related diseases and developmental milestones that children under 6 years old need to achieve**

I. Introduction

With the continuous development of machine learning and deep learning, AI has been a powerful assistant in supporting people in most areas. The field of child mental health and development assessment is chosen by the team to learn, study the uses and challenges that an RAG, LangChain and LLM architecture can bring and encounter.

Our research objective is to apply the Retrieval-Augmented Generation framework (RAG) in combination with LangChain technology and outstanding LLM models for Vietnamese data. The aim is to assess the reliability and efficiency of these models when applied to the field of pediatric mental health and development. Selected LLM models include Viet-Mistral/Vistral-7B-Chat, SeaLLMs/SeaLLMs-v3-7B-Chat, vilm/vinallama-7b-chat and vietgpt/dama-2-7b-chat. Despite differences in model parameters, this study aims to assess the suitability of each model with the Vietnamese dataset, especially when using a relatively small amount of data to fine-tune. To evaluate model quality, we will use indicators such as METEOR, ROUGE and Cosine Similarity.

II. Related Works

Kailai Yang, Tianlin Zhang, Ziyan Kuang, Qianqian Xie, Jimin Huang, and Sophia Ananiadou [1] presented a paper on the integration of mental health analysis and Llama. The authors argued that Large Language Models (LLMs), particularly Llama, and their fine-tuning could significantly improve precision and clarity in predicting mental health conditions from social media data. Although LLMs such as ChatGPT and GPT-4 show good performance, they still have limitations in providing solutions and empathetic analysis based on user or client input. Therefore, fine-tuning these models with data from social networks has proven their strong generalization ability for a variety of tasks, while maintaining the quality of explanations close to human-level reasoning.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, William El Sayed [2] introduced the Mistral 7B model, an LLM based on the LLaMA architecture, with 7 billion parameters designed to optimize performance and efficiency while still delivering impressive results. According to the authors, Mistral 7B outperforms the open 13B model (Llama 2) and the 34B model (Llama 1) across various benchmarks such as reasoning, mathematics, and code generation. The Grouped-Query Attention (GQA) and Sliding Window Attention (SWA) mechanisms played a crucial role in improving speed and reducing costs during training and fine-tuning. Viet-Mistral/Vistral-7B-Chat, a multi-turn conversational LLM, was created by fine-tuning the Vietnamese dataset with the proposal to extend the tokenizer for better support of the Vietnamese language.

Quan Nguyen, Huy Pham and Dung Dao [3] introduced Vinallama, a large language model of Vietnam based on LLAMA-2, enhanced with more than 800 billion additional training notification codes. The model shows strong fluency in Vietnamese and deep understanding of local culture. Adjusted by 1 million bilingual patterns (English-English), Vinallama-7B-Chat achieved advanced results on important NLP benchmarks of Vietnam such as VLSP, Vmlu and Vicuna benchmarks. This emphasizes its ability for specialized applications in the context of Vietnam, especially in scenarios based on dialogue. At the time the article was studied, Vinallama-7B-Chat performed quite well in the benchmark score compared to

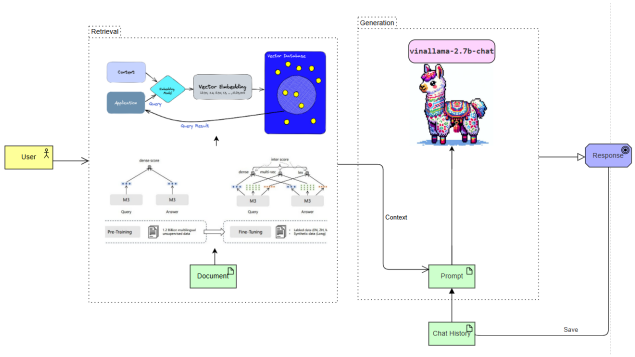
other models.

Xuan-phi Nguyen et al. [4] SEALLMS is proposed, a family of large language models specifically designed for Southeast Asian languages (SEA). These models address the language imbalance found in major LLMs by incorporating extended vocabularies, region-specific fine-tuning, and guidance links tailored to maritime languages. SEALLMS outperforms TATGPT-3.5 in several low-resource languages such as Lao, Khmer, and Burmese, demonstrating strong capabilities in multilingual guided applications and cultural awareness. We decided to include SEALLM in our research due to their flexible parameter configurations (1.5B, 7B, and 13B), as well as their continuous updates aimed at enhancing performance and language coverage over time.

Within the scope of this research, we adopt a Retrieval-Augmented Generation (RAG) architecture in combination with LangChain [5] to enhance contextual understanding for Vietnamese mental health-related queries. To enable efficient and semantically meaningful retrieval, we employ the BGE-M3 embedding model [6] to generate dense vector representations of textual data. These vectors are indexed using FAISS [7], which allows for rapid and accurate similarity search during the retrieval phase. The retrieved context is subsequently passed to one of the selected fine-tuned LLMs (e.g., Vistral-7B-Chat, VinaLLaMA-7B, or SeaLLM-7B), enabling the system to produce contextually relevant and domain-specific responses tailored for pediatric development and mental health consultations. In addition, the research team employs evaluation metrics such as METEOR, ROUGE, and Cosine Similarity [8] [9] to assess the contextual relevance and response quality of each model. These benchmarks allow for a comprehensive comparison, enabling the selection of the most suitable model for real-world deployment in mental health and pediatric development consultation.

III. System Architecture

A. Overall System Design



Hình 1: System architecture of the proposed RAG-based consulting assistant using LangChain and vinallama-2.7b-chat.

The overall architecture of our system is based on the Retrieval-Augmented Generation (RAG) framework, combining a retriever and a large language model (LLM) for more accurate and context-aware responses. The system consists of two main modules: Retrieval and Generation, as illustrated in Figure 1.

In the **Retrieval** module, when a user submits a query, the system encodes it into a vector embedding and compares it against a pre-built FAISS vector database. This database contains vectorized embeddings of documents that were preprocessed and selected from sources related to child development and mental health. The retrieval process employs a search method using `search_type = "similarity"` to identify documents that exhibit a high degree of relevance to the input query. Only documents that meet a predefined similarity score threshold (`score_threshold`) are selected. These documents are then compiled into a context that is passed to the Generation module for further processing.

The **Generation** module uses a Vietnamese LLM—`vinallama-2.7b-chat`—which has been fine-tuned on domain-specific instruction data to enhance its ability to generate context-aware and helpful responses in the field of child development and mental health. The retrieved documents, along with the current prompt, context from the retrieval module, and chat history, are passed to the model as input. The model then produces a response, which is returned to the user and optionally saved for continuous interaction.

B. Component Implementation Details

1) Vector Database Construction: This component is responsible for transforming raw documents into vector representations that can be efficiently queried using similarity search.

The process begins by loading and parsing a collection of PDF documents containing curated knowledge related to child development and mental health. Each document is then segmented into smaller chunks using a recursive text splitter with a chunk size of 1400 characters and an overlap of 200 characters. This overlapping strategy ensures that semantic context is preserved across boundaries and improves embedding consistency.

For vectorization, we employ the BGE-M3 model from the BAAI research group. It is a multilingual, fine-tuned model optimized for dense semantic retrieval. We utilize the SentenceTransformer interface to embed both the query and the document chunks.

All embeddings are normalized and stored in a FAISS vector database. This structure enables fast approximate nearest-neighbor (ANN) search. Prior to saving, the system verifies the integrity of the index by checking the alignment between document chunks and vector entries.

To further ensure data quality and minimize retrieval noise, we test the resulting vector store with a sample query. Any failure during the process triggers an automatic cleanup of the invalid store to prevent corrupted data from affecting future interactions.

2) LLM Fine-Tuning: This section details the methodology employed for fine-tuning various Large Language Models (LLMs) on a Vietnamese language task. The process encompasses model selection, quantization, data preprocessing, fine-tuning strategy, and evaluation.

a) Model Selection and Quantization: Pretrained base models, specifically VinaLLaMA (`vilm/vinallama-7b-chat`), Vistral-7B (`Viet-Mistral/Vistral-7B-Chat`), SeaLLMs (`SeaLLMs/SeaLLMs-v3-7B-Chat`), and Dama-2 (`vietgpt/dama-2-7b-chat`), were loaded from their respective Hugging Face repositories. To mitigate memory consumption and accelerate the training process, a strategic Post-Training Quantization (PTQ) approach was adopted using the `bitsandbytes` library. For initial single-run evaluations (Section III-B2d), models were quantized to 8-bit (int8) precision. However, for the more extensive 5-fold cross-validation (Section III-B2e), 4-bit (int4) quantization, specifically the Normalized Float 4-bit (NF4) format with `bnb_4bit_compute_dtype=torch.bfloat16` and `bnb_4bit_use_double_quant=True`, was applied. This INT4 PTQ configuration reduced the memory footprint by up to 50% compared to FP16, enabling multiple training iterations on standard GPU hardware (e.g., NVIDIA A100). The impact of quantization was rigorously evaluated using ROUGE, METEOR, and Cosine Similarity metrics, confirming minimal degradation in response quality.

b) Low-Rank Adaptation (LoRA) Configuration: The fine-tuning process leverages Low-Rank Adaptation (LoRA) [7] for parameter-efficient fine-tuning. A LoRA rank (`r`) of **32** was chosen as a balance between model adaptability and computational overhead, a common practice providing sufficient capacity for adaptation without an excessive increase in trainable parameters. The LoRA alpha (`lora_alpha`) was also set to 32, and a dropout rate (`lora_dropout`) of 0.5 was applied to LoRA layers to prevent overfitting.

To effectively adapt the models, LoRA was applied to a comprehensive set of target modules within the transformer architecture. These include:

- **Attention Mechanism Modules**: `q_proj` (query projection), `k_proj` (key projection), `v_proj` (value projection), and `o_proj` (output projection). Targeting these modules allows the model to refine how it weighs and combines information from different parts of the input sequence, crucial for understanding context and relationships.
- **Feed-Forward Network (FFN) Modules**: `gate_proj` (gating projection in SwiGLU/GeGLU variants), `up_proj` (upscaling projection in FFN), and `down_proj` (downscaling projection in FFN). Adapting these layers enables the model to learn more complex, task-specific feature transformations.

This selection aims to imbue the model with task-specific knowledge while largely preserving its pretrained general language understanding capabilities by only updating the low-rank decomposition matrices.

c) Data Preprocessing: Rigorous data preprocessing was conducted to ensure the quality and consistency of the training dataset:

1) **Text Normalization**: All text data (both instructions and responses) was converted to lowercase. Punctuation was removed to reduce noise and focus on lexical content. Vietnamese word tokenization was performed using the `underthesea` library.
2) **Duplicate and Near-Duplicate Removal**: To prevent data leakage and improve model generalization, a fuzzy matching technique was employed. The `thefuzz` library (specifically, `fuzz.ratio`) was used to calculate the similarity between all pairs of input instructions and, separately, all pairs of output responses. Records where either the input or output exhibited a similarity score greater than a specified threshold (0.9, i.e., 90% similarity) with another record were removed, keeping only the first occurrence. This step helps ensure a more diverse and less redundant training set.
3) **Instruction-Response Coherence (Optional but mentioned in original thought process)**: Initially, consideration was given to filtering instruction-response pairs based on their cosine similarity using a pretrained embedding model. Pairs with similarity scores below a threshold (e.g., 0.9) would be removed to eliminate noisy or inconsistent samples. However, the primary implemented filtering focused on fuzzy matching of inputs and outputs separately as described above.

d) Single-Run Evaluation: Before proceeding to full cross-validation, an initial single-run fine-tuning and evaluation was performed for each model. The preprocessed dataset was split into a training set (80%) and an evaluation set (20%) using a fixed `random_state` for reproducibility. Models were fine-tuned using the LoRA configuration and int8 quantization as described above. The primary purpose of this step was to obtain a preliminary assessment of each model's performance and to debug the training pipeline.

e) 5-Fold Cross-Validation: To robustly assess generalization performance and mitigate potential biases from a single train-test split, a 5-fold cross-validation strategy was employed. The entire preprocessed dataset (after fuzzy duplicate removal) was divided into five mutually exclusive folds. For each iteration, one fold served as the evaluation set, while the remaining four folds were combined to form the training set. Fine-tuning was performed independently for each of the five iterations using the LoRA configuration and int4 quantization. This approach provides a more reliable estimate of model performance across different subsets of the data.

f) Evaluation Metrics: The performance of the fine-tuned models was evaluated using a suite of standard metrics for text generation tasks:

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation)** [11]: ROUGE-1, ROUGE-2, and ROUGE-L scores were calculated to measure n-gram overlap between the generated and reference texts.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering)** [12]: METEOR considers synonymy and stemming along with precision and recall.
- **Cosine Similarity**: To assess semantic similarity, sentence embeddings were generated for both predicted and reference texts using the **BAAI/bge-m3** model from Sentence

Transformers [9]. The cosine similarity between these embedding vectors was then computed, providing a measure of how semantically close the generated output is to the target.

ROUGE-1

Formula:

$$\text{ROUGE-1} = \frac{\sum_{g \in G_1(R)} \min\left(\text{Count}_C(g), \text{Count}_R(g)\right)}{\sum_{g \in G_1(R)} \text{Count}_R(g)} \quad (1)$$

Explanation:

- $G_1(R)$ is the set of all *unigrams* (single words) in the reference text R.
- $\text{Count}_C(g)$ is the number of times unigram g appears in the candidate text C.
- $\text{Count}_R(g)$ is the number of times unigram g appears in the reference text R.
- The numerator sums the number of overlapping unigrams between C and R, using the minimum count from each to avoid overcounting.
- The denominator is the total number of unigrams in the reference, including duplicates.

Interpretation:

ROUGE-1 measures the unigram recall — how much of the reference content is covered by the candidate. The score ranges from 0 to 1:

- ROUGE-1 = 1 means all reference words appear in the candidate.
- ROUGE-1 = 0 means there is no unigram overlap between candidate and reference.

ROUGE-2

Formula:

$$\text{ROUGE-2} = \frac{\sum_{g \in G_2(R)} \min\left(\text{Count}_C(g), \text{Count}_R(g)\right)}{\sum_{g \in G_2(R)} \text{Count}_R(g)} \quad (2)$$

Explanation:

- $G_2(R)$ is the set of all *bigrams* (pairs of consecutive words) in the reference text R.
- $\text{Count}_C(g)$ is the number of times bigram g appears in the candidate text C.
- $\text{Count}_R(g)$ is the number of times bigram g appears in the reference text R.
- The numerator counts the number of overlapping bigrams between C and R, taking the minimum count to avoid duplication.
- The denominator is the total number of bigrams in the reference (including repeated ones).

Interpretation:

ROUGE-2 measures the bigram-level recall — how many bigram phrases from the reference appear in the candidate. It focuses more on fluency and local word ordering compared to ROUGE-1. The score ranges from 0 to 1, where:

- ROUGE-2 = 1 means every bigram in the reference also exists in the candidate.
- ROUGE-2 = 0 means there is no bigram overlap at all.

ROUGE-L

Formula:

$$P_{\text{LCS}} = \frac{\text{LCS}(C, R)}{|C|}, \quad R_{\text{LCS}} = \frac{\text{LCS}(C, R)}{|R|} \quad (3)$$

$$\text{ROUGE-L} = \frac{(1 + \beta^2) \cdot P_{\text{LCS}} \cdot R_{\text{LCS}}}{\beta^2 \cdot P_{\text{LCS}} + R_{\text{LCS}}} \quad (4)$$

Explanation:

- $\text{LCS}(C, R)$ is the length of the *Longest Common Subsequence* between the candidate text C and the reference text R.
- $|C|$ and $|R|$ are the total number of words in the candidate and reference texts, respectively.
- P_{LCS} is the LCS-based precision: the proportion of the candidate covered by the LCS.
- R_{LCS} is the LCS-based recall: the proportion of the reference covered by the LCS.
- β is a weighting factor to control the balance between recall and precision. Commonly, $\beta = 1$ to compute the F1-score.

Interpretation:

ROUGE-L captures the longest shared in-sequence word overlap between the candidate and the reference, regardless of contiguity. It is particularly useful for evaluating fluency and sequence alignment. A score of:

- 1 indicates perfect in-order match between candidate and reference,
- 0 indicates no sequence overlap.

METEOR

Formula:

$$P = \frac{m}{|C|}, \quad R = \frac{m}{|R|} \quad (5)$$

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9 \cdot P} \quad (6)$$

$$\text{Penalty} = \gamma \left(\frac{ch}{m}\right)^{\beta} \quad (7)$$

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \quad (8)$$

Explanation:

- m is the number of matched unigrams between the candidate C and reference R.
- $|C|$ and $|R|$ are the total number of unigrams in the candidate and reference texts, respectively.
- P and R are unigram-level precision and recall.

- F_{mean} is a harmonic mean of P and R, with recall weighted 9 times more than precision.
- ch is the number of *chunks* (i.e., contiguous matched subsequences in order).
- γ and β are hyperparameters that control how harshly disordered matches are penalized. Common values: $\gamma = 0.5$, $\beta = 3$.
- Penalty increases as the number of chunks grows (i.e., when matches are more fragmented).

Interpretation:

METEOR measures both content matching and word order alignment. It rewards matches at the unigram level but penalizes disordered or fragmented alignments. A score of:

- 1 indicates a perfect match in content and order,
- 0 indicates no match at all.

Cosine Similarity

Formula:

$$\text{CosineSim}(X, Y) = \frac{\sum_{i=1}^{n} X_i \cdot Y_i}{\sqrt{\sum_{i=1}^{n} X_i^2} \cdot \sqrt{\sum_{i=1}^{n} Y_i^2}} \tag{9}$$

Explanation:

- $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_n)$ are the vector representations of the candidate and reference texts, respectively.
- X_i and Y_i are the i-th components of vectors X and Y (e.g., embedding values).
- The numerator is the dot product between X and Y.
- The denominator is the product of the Euclidean norms (magnitudes) of X and Y.

Interpretation:

Cosine similarity measures the angle between two vectors in a high-dimensional space. It captures the semantic closeness between two texts regardless of their length. The score ranges from:

- 1: the vectors point in the same direction (perfect match),
- 0: the vectors are orthogonal (no similarity),
- -1: the vectors are in opposite directions (rare in non-negative text embeddings).

All metrics were calculated using the `evaluate` library. For cross-validation, metrics were averaged across the five folds.

g) Best Model Selection and Merging: Based on the average cross-validation results, particularly prioritizing Cosine Similarity, the best performing fold's adapter weights were selected for each base model. These LoRA adapters were then merged into their respective full-precision (BF16) base models to create a fine-tuned, standalone model. This merged model represents the final fine-tuned version intended for subsequent deployment or conversion.

h) GGUF Conversion for Ollama Deployment: As a final step towards practical deployment, the merged fine-tuned models (in BF16 format) are intended to be converted to the GGUF (GPT-Generated Unified Format). This format is optimized for efficient inference on CPU and GPU, and is widely supported by platforms such as Ollama, facilitating local execution and broader accessibility of the fine-tuned models. The conversion process typically involves scripts provided by projects like `llama.cpp`.

3) Integration with RAG and LangChain: The final system was fully implemented by integrating the RAG (Retrieval-Augmented Generation) architecture with the LangChain framework, and deploying it as a production-ready chatbot using Ollama and FastAPI. Workflow of the system is shown in Figure 1.

User Interaction via API Endpoint: The user submits a query to the API endpoint. This query is passed into the retrieval pipeline, which uses FAISS to identify the most semantically similar document chunks from the knowledge base, previously encoded using the BGE-M3 model.

Retrieval Phase: The retriever searches using `search_type = similarity` with k and a `score_threshold`. Only documents that satisfy the semantic similarity threshold are used to construct the context for the model.

Prompt Construction: The prompt is constructed by concatenating the user query, the retrieved context, and the system message (including the conversation history).

Prompt Vietnamese Template Used in the System

<im_start|>|system
Bạn là trợ lý AI chuyên về sức khỏe và tâm thần nhi khoa. Mục tiêu của bạn là cung cấp thông tin chính xác, dễ hiểu và hữu ích cho phụ huynh. **Hãy luôn trả lời bằng tiếng Việt. Đặc biệt khi có câu hỏi liên quan đến thông tin liên lạc và đặt lịch phòng khám, hãy sử dụng những thông tin sau:**

- **Thông tin liên hệ / đặt lịch phòng khám private:**
- Số điện thoại/Zalo: **private**
- Địa chỉ: **private**

QUAN TRỌNG: Tất cả các câu trả lời, lời khuyên, và thông tin về mốc phát triển phải dựa vào độ tuổi so với ngày hiện tại (`{current_date}`), được điều chỉnh và phù hợp CHÍNH XÁC với ĐỘ TUỔI của trẻ được đề cập trong câu hỏi hoặc lịch sử trò chuyện. Nếu không có thông tin độ tuổi rõ ràng, bạn có thể hỏi lại một cách lịch sự để làm rõ.
`{asq_guidance_placeholder}`
Hãy tuân thủ các hướng dẫn sau:

1) **HIỂU ĐÚNG NGỮ NGHĨA:** Phân tích kỹ lưỡng toàn bộ câu hỏi để hiểu đúng ý định, đặc biệt với các từ có nhiều nghĩa (ví dụ: "bập bẹ" có thể là tập nói hoặc tập đi). Nếu không chắc chắn, hãy hỏi lại người dùng để làm rõ.
2) **Ưu tiên Context:** Lấy thông tin từ phần **Context** được cung cấp làm nguồn chính để trả lời câu hỏi.
3) **Bổ sung từ kiến thức của bạn:** Nếu Context không có thông tin, không đầy đủ, hoặc bạn cảm thấy kiến thức đã được huấn luyện có thể làm rõ hoặc bổ sung giá trị cho câu trả lời, hãy sử dụng nó. Thông tin bổ sung phải liên quan trực tiếp đến câu hỏi và lĩnh vực chuyên môn của bạn, **phù hợp với độ tuổi của trẻ.**
4) **Chính xác và không bịa đặt:** Dù thông tin lấy từ Context hay từ kiến thức của bạn, nó phải chính xác, dựa trên cơ sở khoa học, **và hoàn toàn phù hợp với độ tuổi của trẻ đang được hỏi đến**. Tuyệt đối không bịa đặt thông tin hoặc đưa ra lời khuyên không phù hợp lứa tuổi.
5) **Tập trung vào câu hỏi:** Luôn trả lời trực tiếp vào câu hỏi `{question}`.
6) **Khi không có thông tin:** Nếu cả Context và kiến thức đã huấn luyện của bạn đều không có thông tin để trả lời câu hỏi, hãy thông báo một cách lịch sự, ví dụ: "Tôi rất tiếc, hiện tại tôi không có đủ thông tin về chủ đề này từ cả tài liệu được cung cấp lẫn kiến thức của mình."
7) **Diễn đạt:** Tự nhiên, không trích dẫn nguyên văn từ Context trừ khi đó là một định nghĩa quan trọng hoặc trích dẫn ngắn cần thiết.
8) **Định dạng:** Rõ ràng, dùng gạch đầu dòng (-) hoặc số (1., 2.) nếu phù hợp, câu đầy đủ, đúng ngữ pháp.
9) **Giọng điệu:** Ngôn ngữ đơn giản, thân thiện, cảm thông và hỗ trợ.
10) **Lịch sử hội thoại:** Sử dụng lịch sử `{chat_history}` để hiểu ngữ cảnh các câu hỏi trước đó, nhưng câu trả lời phải tập trung vào CÂU HỎI HIỆN TẠI, hạn chế sử dụng câu trả lời trước đó trong câu trả lời hiện tại.

Context: `{context}`
Lịch sử hội thoại: `{chat_history}`
<im_end|>| <im_start|>|user
`{question}`
<im_end|>| <im_start|>|assistant

C. Dataset

The dataset used in this research comprises two main types:

1) A conversational dataset used to train chatbot models.
2) A document dataset used for knowledge retrieval and context injection within the RAG pipeline.

1. Conversational Dataset for Chatbot: This dataset was designed to support instruction-based training of Vietnamese LLMs. It consists of natural language dialogues relevant to concerns raised by parents at pediatric mental health clinics.

1) **Internal Dataset:** This dataset was manually collected from real conversations on social platforms such as TikTok, Zalo, and Facebook, between parents and psychologists at a children's psychology clinic in Ho Chi Minh City. Key topics include:
   - Greetings and consultation openings
   - Developmental milestones
   - ASQ-3 screening instructions
   - Solutions and concerns regarding ASD or developmental delays

   All entries were reviewed and filtered by clinical psychologists for relevance and quality.
2) **Collected Dataset:** This dataset was constructed by extracting user queries related to child development and mental health using the Google Suggest API. The resulting keywords were clustered into topics matching the internal dataset and were validated by psychologists.
3) **Merged Dataset:** The internal and collected datasets were combined to form high-quality instruction–response pairs, allowing the chatbot to generate both natural and professional replies.

(a) Example Dataset: An example data entry follows the seq2seq JSON format:

```
{
  "instruction": "Câu chào hỏi",
  "input": "Chào, cho mình hỏi?",
  "output": "Bạn cần mình giúp gì nào?"
}
```

(b) Number of Attributes: Each record contains three key-value pairs:
   - `instruction`: the purpose or theme of the conversation
   - `input`: the user's question
   - `output`: the chatbot's expected response

(c) Dataset Size:
   - Number of files: 4 (General Chatbot, ASD, Development, Delayed Speech)
   - File sizes: 100KB, 45KB, 30KB, and 12KB respectively
   - Total records: 2,000 dialogue pairs

(d) Usage: This dataset is used for fine-tuning and evaluating multiple Vietnamese LLMs, including:
   - `Viet-Mistral/Vistral-7B-Chat`
   - `SeaLLMs/SeaLLMs-v3-7B-Chat`
   - `vilm/vinallama-7b-chat`
   - `vietgpt/dama-2-7b-chat`

2. Document Dataset for Knowledge Retrieval: This dataset comprises curated documents obtained from both public and

proprietary sources, including pediatric guidelines, diagnostic manuals, and intervention documents from clinical psychologists. It supports the RAG framework by providing factual context to enhance chatbot responses.

(a) Example Dataset: A representative document is a 10–12 page PDF outlining intervention strategies for ASD in children aged 10–12 months. The file contains textual content, tables, and clinical diagrams.
(b) Number of Attributes: This dataset contains unstructured PDF documents; no structured attributes are defined.
(c) Dataset Size:
- Number of documents: 53 PDF files
- Total size: approximately 124 MB

(d) Usage: The documents are split into text chunks, embedded using the BGE-M3 model, and stored in a FAISS vector database. During inference, the retriever fetches relevant documents, which are injected as context into the LLM using the RAG architecture.

IV. Results and Discussion

This section presents the performance evaluation of four fine-tuned large language models-SeaLLM, VinaLLaMA, Vistral, and Dama-2 on the Vietnamese mental health dialogue dataset. The evaluation was conducted using both single run testing and 5 fold cross validation, with standard metrics including ROUGE-1, ROUGE-2, ROUGE-L, METEOR, and Cosine Similarity. These metrics assess various aspects of text generation quality such as lexical overlap, fluency, semantic alignment, and vector-based similarity.

A. Single-Run Evaluation

Table I, II summarizes the evaluation scores for a single inference run across all models.

Table I: Single-run ROUGE Scores for All Models

Model	ROUGE-1	ROUGE-2	ROUGE-L
Vistral	**0.6409**	**0.3247**	**0.4375**
Dama-2	0.6079	0.2874	0.3984
VinaLLaMA	0.6076	0.2652	0.3891
SeaLLM	0.6210	0.2679	0.3887

Table II: Single-run Semantic Similarity Metrics

Model	METEOR	Cosine Similarity
Vistral	**0.3567**	**0.7993**
Dama-2	0.3269	0.7793
VinaLLaMA	0.2958	0.7726
SeaLLM	0.2989	0.7167

B. Cross-Validation Evaluation

To assess the models' robustness and generalization capabilities, we applied 5-fold cross-validation. Table III, IV shows the average scores across the folds.

Table III: Cross-Validation ROUGE Scores (K=5)

Model	ROUGE-1	ROUGE-2	ROUGE-L
Vistral	**0.6399**	**0.3158**	**0.4259**
Dama-2	0.6084	0.2806	0.3935
VinaLLaMA	0.6104	0.2735	0.3905
SeaLLM	0.6184	0.2664	0.3839

Table IV: Cross-Validation Semantic Similarity Scores (K=5)

Model	METEOR	Cosine Similarity
Vistral	**0.3549**	**0.7972**
Dama-2	0.3360	0.7757
VinaLLaMA	0.3075	0.7746
SeaLLM	0.2986	0.7079

C. Expert Opinion

An expert evaluation was conducted using 21 questions collected from real-world consultations at a pediatric mental health clinic. These questions were reviewed by a specialist in child development and mental health, emphasizing clinical relevance and accuracy of responses generated by four fine-tuned AI models deployed on the Ollama platform. All models were evaluated under uniform conditions, using identical hyperparameters: temperature set at 0.5, context window size (`num_ctx`) of 12288, and maximum prediction length (`num_predict`) of 8192 tokens. Furthermore, each model utilized a standardized prompt clearly defining their role as an AI assistant specialized in pediatric health and mental health, tasked with providing gentle, accurate, and scientifically-based advice.

The specialist noted that the **Vistral** and **SeaLLM** models generally produced responses aligned closely with clinical expectations, effectively capturing the intended clinical nuances and specificity required by the provided documents and fine-tuning dataset. Both models demonstrated consistent semantic coherence, offering clinically relevant guidance suitable for parents and healthcare providers.

Conversely, the **VinaLLaMA** and **Dama-2** models exhibited inconsistencies in the quality of their responses. Specifically, the expert identified frequent syntactic errors and semantic inaccuracies, resulting in responses often deviating from the primary clinical inquiries. These models struggled with accurately interpreting and effectively addressing the nuances of pediatric mental health consultations.

A significant limitation identified across all models was their reliance on outdated medical knowledge, not fully compliant with the latest ICD-10 and ICD-11 standards. Despite extensive fine-tuning and retrieval augmentation, all models occasionally provided recommendations based on obsolete or replaced clinical guidelines. The expert highlighted the importance of continuous updates to the medical knowledge integrated into these AI systems, ensuring ongoing clinical validity and reliability of AI-generated recommendations.

D. Discussion

The comparative evaluation of the four fine-tuned models—**Vistral**, **SeaLLM**, **VinaLLaMA**, and **Dama-2**—revealed distinct differences in performance across both quantitative

metrics and practical usability. According to the ROUGE, METEOR, and Cosine Similarity scores, the **Vistral** model consistently achieved the highest results, indicating its strong capability for generating semantically accurate and contextually relevant responses in Vietnamese pediatric mental health consultations. **Dama-2** and **VinaLLaMA** followed, while **SeaLLM** ranked lowest across most metrics.

However, practical expert assessment highlighted that **SeaLLM**, despite its relatively lower evaluation scores, performed admirably in real-world scenarios. The model provided coherent and relevant guidance that often met or exceeded clinical expectations, suggesting that standard text-generation metrics may not fully capture the model's effectiveness in nuanced, conversational pediatric healthcare settings.

Another important observation is that the knowledge base of the models—especially regarding up-to-date clinical standards—can be further enhanced. Missing or outdated information can be supplemented through the integration of additional high-quality and domain-specific documents, which can be ingested into the RAG retrieval system. This approach allows for continuous improvement of the assistant's knowledge without requiring complete retraining of the base language model.

In the future, the system can be further improved by combining advanced machine learning techniques and deep learning, especially techniques related to image processing and multimodal understanding. For example, taking advantage of vision large language model (Vision LLM) may allow assistance to explain and respond to visual data such as development charts or diagnostic imaging. In addition, the implementation of RAG agents, a generation strategy to enhance improvement with the decision -based decision -making.

REFERENCES

[1] Yang, K., Zhang, T., Kuang, Z., Xie, Q., Huang, J., & Ananiadou, S. (2024). "MentaLLaMA: Interpretable Mental Health Analysis on Social Media with Large Language Models." arXiv preprint arXiv:2309.13567v3.

[2] Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., & El Sayed, W. (2023). "Mistral 7B: Towards Efficient and High-Performance Language Models." arXiv preprint arXiv:2302.13971.

[3] Nguyen, Q., Pham, H., & Dao, D. (2023). "VinaLLaMA: LLaMA-based Vietnamese Foundation Model." arXiv preprint arXiv:2312.11011v1.

[4] Nguyen, X.-P., Zhang, W., Li, X., Aljunied, M., Hu, Z., Shen, C., Chia, Y. K., Li, X., Wang, J., Tan, Q., Cheng, L., Chen, G., Deng, Y., Yang, S., Liu, C., Zhang, H., & Bing, L. (2023). "SeaLLMs: Large Language Models for Southeast Asia." arXiv preprint arXiv:2312.16934v2.

[5] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). "Retrieval-Augmented Generation for Large Language Models: A Survey." arXiv preprint arXiv:2312.10997v5.

[6] Hebert, L., Golab, L., Poupart, P., & Cohen, R. (n.d.). "FedFormer: Contextual Federation with Attention in Reinforcement Learning." [Unpublished manuscript].

[7] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). "LoRA: Low-Rank Adaptation of Large Language Models." arXiv preprint arXiv:2106.09685.

[8] Lakatos, R., Pollner, P., Hajdu, A., & Joó, T. (2024). "Investigating the Performance of Retrieval-Augmented Generation and Fine-Tuning for the Development of AI-Driven Knowledge-Based Systems." arXiv preprint arXiv:2403.09727v1.

[9] Steck, H., Ekanadham, C., & Kallus, N. (2024). "Is Cosine-Similarity of Embeddings Really About Similarity?" arXiv preprint arXiv:2403.05440v1.

[10] Li, H., Li, X., Hu, P., Lei, Y., Li, C., Zhou, Y., ... & Zhou, Y. (2023). "Boosting Multi-modal Model Performance with Adaptive Gradient Modulation." arXiv preprint arXiv:2308.07686.

[11] Lin, C.-Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries." In *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Post-Conference Workshop of ACL 2004*, Barcelona, Spain.

[12] Banerjee, S., & Lavie, A. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, Ann Arbor, Michigan, pp. 65–72.