DentalGPT: A Multi-Expert Language Model for Vietnamese Dental Consultation

DAT TRAN and VU CAO, Thuyloi University, Vietnam

The proliferation of general-purpose large language models (LLMs) has highlighted a significant gap in specialized domains such as dentistry, particularly for non-English languages. Existing systems often lack the domain-specific knowledge and contextual understanding required for accurate medical consultation, especially for Vietnamese users. This paper introduces DentalGPT, a specialized conversational agent designed to address these limitations by providing precise and contextually relevant dental advice in Vietnamese through intelligent expert routing. The core of our methodology involves fine-tuning the DeepSeek-R1 model, which leverages a Mixture-of-Experts (MoE) architecture with semantic-based expert selection, on a comprehensive, custom-built Vietnamese dental dataset. This dataset comprises approximately 3 million dialogue samples aggregated from diverse sources, including medical literature, clinical guidelines, and doctor-patient conversations, ensuring both professional accuracy and practical relevance. Each sample is annotated with multiple domain-specific labels (e.g., orthodontics, endodontics, preventive care) to enable intelligent expert routing. During inference, the system employs a sentence similarity and ranking mechanism to dynamically select the top-k most relevant experts for each query, ensuring specialized handling of diverse dental inquiries. The training process employed a combination of Supervised Fine-Tuning (SFT) with QLoRA for efficient optimization and Reinforcement Learning from Human Feedback (RLHF) to align model responses with expert standards and user expectations. Quantitative evaluations demonstrate the model's high performance, achieving a Perplexity of 1.88, a BLEU score of 0.53, and a BERTScore of 0.93. In comparative benchmarks, DentalGPT shows state-of-the-art capabilities, scoring 91.0 on MMLU, outperforming prominent models like GPT-4o. The resulting system is a reliable and user-friendly chatbot that successfully bridges the gap in accessible digital healthcare, proving the efficacy of fine-tuning expert-based models with semantic routing for specialized, non-English domains. Code is available at: https://github.com/Nvcoing/DentalGPT.git

CCS Concepts: • Computing methodologies → Artificial intelligence; Natural language processing; Language resources; Natural language generation; Dialogue management; • Human-centered computing → Human computer interaction (HCI); Interaction paradigms; Conversational interaction; • Applied computing → Health care information systems; Consumer health..

Additional Key Words and Phrases: Healthcare chatbot, dental consultation, large language models (LLMs), mixture-of-experts (MoE), supervised fine-tuning (SFT), reinforcement learning from human feedback (RLHF), parameter-efficient fine-tuning (PEFT), vietnamese NLP

ACM Reference Format:
Dat Tran and Vu Cao. 2025. DentalGPT: A Multi-Expert Language Model for Vietnamese Dental Consultation. In Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX). ACM, New York, NY, USA, 25 pages. https://doi.org/XXXXXXX.XXXXXXX

Authors' Contact Information: Dat Tran, dat.trananh@tlu.edu.vn; Vu Cao, 2151264695@e.tlu.edu.vn, Thuyloi University, Hanoi, Vietnam.

1 Introduction

The rapid advancement of Large Language Models (LLMs) [21] has revolutionized numerous domains, including healthcare, by enabling the development of intelligent consultation systems. However, the effectiveness of these general-purpose models significantly declines when applied to highly specialized fields such as dentistry, which demand strict medical accuracy and complex reasoning capabilities. This challenge is further exacerbated in non-English languages, such as Vietnamese, where domain-specific training data is often scarce. As a result, LLMs may misinterpret context and deliver inaccurate or culturally inappropriate responses. This gap highlights a critical need for a reliable, domain-specialized conversational tool tailored to the dental field in Vietnam.

While general-purpose LLMs have achieved remarkable success in broad-domain tasks, they face fundamental limitations when deployed in specialized medical contexts. First, the knowledge distribution in pre-training corpora is heavily skewed toward English and general-domain content, leading to inadequate representation of medical terminology and clinical reasoning patterns in low-resource languages. Second, dental consultation requires not only factual accuracy but also nuanced understanding of symptom descriptions, treatment trade-offs, and patient safety considerations—capabilities that cannot be reliably achieved through zero-shot or few-shot prompting alone. Third, the regulatory and ethical requirements for medical AI systems demand explicit control over model behavior, including the ability to escalate emergency cases and avoid harmful self-treatment recommendations—requirements that necessitate targeted fine-tuning rather than reliance on general-purpose instruction-following capabilities.

To address these fundamental challenges, this paper introduces DentalGPT, a domain-specific consultation chatbot designed to provide accurate, context-aware, and reliable dental information for Vietnamese users. This work makes three primary contributions that extend beyond simple model application. First, we develop and validate a systematic methodology for adapting large-scale MoE architectures to specialized medical domains in low-resource languages. This methodology encompasses data collection protocols, multi-expert annotation frameworks, and safety-oriented training procedures that can be generalized to other medical specialties and languages. Second, we construct and release ViDental, the first large-scale Vietnamese dental consultation dataset with multi-annotator validation and explicit quality control procedures. This dataset addresses a critical gap in Vietnamese medical NLP resources and provides a reusable foundation for future research. Third, we demonstrate through rigorous quantitative and qualitative evaluation that targeted domain adaptation can enable smaller models to achieve expert-level performance on specialized medical tasks, challenging the prevailing assumption that larger model scale is the primary path to domain competence.

From a methodological standpoint, we fine-tune the DeepSeek-R1 model on the constructed Vietnamese dental dataset through a carefully designed two-stage pipeline. The first stage employs Supervised Fine-Tuning (SFT) with QLoRA for efficient parameter adaptation, enabling the model to acquire domain-specific terminology, clinical reasoning patterns, and consultation dialogue structures. The second stage applies Reinforcement Learning from Human Feedback (RLHF) using the Odds Ratio Preference Optimization (ORPO) algorithm to align model behavior with expert standards and patient safety requirements. This alignment process is critical for medical applications, as it enables the model to appropriately escalate urgent cases, avoid overconfident diagnoses, and provide actionable guidance within the scope of remote consultation.

Our evaluation demonstrates that DentalGPT achieves competitive or superior performance compared to significantly larger general-purpose models across multiple dimensions. Quantitatively, the model attains a perplexity of 1.88, BLEU score of 0.53, and BERTScore of 0.93, indicating high linguistic quality. On clinical reasoning benchmarks, expert dentists rate the model's diagnostic accuracy at 87.2% and treatment recommendation appropriateness at 79.4%, with a notably low hallucination rate of 4.8%. Comparative benchmarking shows that DentalGPT achieves 91.0 on MMLU and 73.0 on GPQA Diamond, outperforming models orders of magnitude larger in parameter count. Qualitative assessment by healthcare professionals and patients confirms high satisfaction across accuracy, clarity, and safety dimensions.

Beyond demonstrating strong performance on a specific task, this work provides broader methodological insights for the medical NLP community. Our systematic approach to dataset construction—including explicit inclusion criteria, multi-stage annotation protocols, and inter-annotator agreement measurement—offers a replicable template for developing medical consultation datasets in other domains and languages. The two-stage training pipeline, combining knowledge transfer through SFT with behavioral alignment through RLHF, represents a generalizable framework for adapting foundation models to safety-critical applications. Our ablation studies isolate the contributions of domain-specific fine-tuning, dataset quality, and preference optimization, providing empirical guidance for future work in medical AI.

The remainder of this paper is structured as follows: Section 2 reviews related work in healthcare conversational AI, Transformer architectures, and Mixture-of-Experts models. Section 3 presents our comprehensive methodology for dataset construction, including detailed data collection protocols and quality control procedures. Section 4 describes the model architecture and training pipeline, with formal specification of the SFT and RLHF objectives. Section 5 reports experimental results, including quantitative benchmarks, qualitative assessments, and ablation studies. Section 6 discusses the broader implications of our approach for low-resource medical NLP and outlines directions for future research. Finally, Section 7 concludes with key findings and practical recommendations for deploying specialized medical AI systems.

2 Related work

Over the past decade, conversational artificial intelligence (Conversational AI) [2, 31]—particularly chatbot systems—has made remarkable progress and is increasingly applied in the healthcare sector to enhance patient experiences. Despite this potential, applying existing chatbot systems [9, 40] to the dental domain reveals several critical limitations. One major issue is the lack of deep domain-specific knowledge. Most chatbots [4, 26] are built on general-purpose deep learning models that are not specialized for dentistry, resulting in limited understanding of dental terminology and a tendency to provide inaccurate advice. This challenge is further compounded by language barriers, as many medical chatbots are primarily developed in English, limiting their ability to naturally process and understand Vietnamese. The scarcity of Vietnamese-language training data specific to dentistry significantly reduces the accuracy of AI models in Vietnam. Furthermore, current systems often fail to meet the need for personalization; they tend to offer generic responses without accounting for an individual's dental history. These limitations highlight the urgent need for a specialized chatbot system capable of deep contextual understanding and tailored responses to meet the specific needs of users in the dental domain.

The introduction of the Transformer [2, 29] architecture marked a groundbreaking shift in the field of Natural Language Processing (NLP) [29], offering substantial improvements over previous sequential models such as Recurrent Neural Networks (RNNs) [7] and Long Short-Term Memory (LSTM) [22] networks. At the heart of this innovation lies the Attention mechanism, particularly Self-Attention—which enables the model to process the entire context of a sentence in parallel rather than sequentially. This parallel processing approach not only enhances computational efficiency but also significantly improves the model's ability to capture long-range semantic dependencies within text. Due to these advantages, the Transformer has rapidly become the foundational architecture behind most modern Large Language Models (LLMs), including prominent model families such as OpenAI's GPT [18], Meta AI's LLaMA [33], and DeepSeek [23]. These models have been widely adopted across a range of NLP tasks, from machine translation and text summarization to the development of increasingly sophisticated chatbot systems.

To address the challenges of performance and computational cost associated with scaling large language models, the Mixture-of-Experts (MoE) [5, 27, 38] architecture has emerged as an effective solution. Rather than requiring the entire neural network to process every input, MoE [38] is a deep learning architecture designed to decompose a task into smaller components, each handled by a specialized module known as an expert. A key component of this architecture, called the router, analyzes the input and selectively activates only a subset of the most relevant experts to process that information.

This mechanism offers two primary advantages: it significantly optimizes computational efficiency by reducing overall workload and training costs [12], and it enhances model accuracy by allowing each expert to focus deeply on a specific domain—for example, disease diagnosis, treatment consultation, or planning tasks in the field of dentistry [27]. This architecture underpins the DeepSeek [23] model family used in this study, enabling a balance between expert-level performance and efficient resource utilization.

Traditional MoE architectures rely on learned routing mechanisms that may lack interpretability. To address this, we introduce a semantic-based expert selection strategy tailored for the dental domain, enabling the model to leverage domain knowledge explicitly during both training and inference. During dataset preparation, each training sample in the ViDental dataset is annotated with one or more domain-specific labels corresponding to dental subspecialties, such as orthodontics (alignment, braces, retainers), endodontics (root canal, pulp treatment), periodontics (gum disease, gingivitis, periodontitis), prosthodontics (crowns, bridges, dentures, implants), preventive care (oral hygiene, fluoride, sealants), oral surgery (extractions, wisdom teeth removal), pediatric dentistry (child-specific concerns), and general consultation (multiple or unspecified topics). This multi-label annotation ensures that complex queries involving multiple dental aspects can be routed to the appropriate combination of experts, enhancing both accuracy and interpretability.

Inference-time Expert Selection. At inference time, when a user query is received, the system first encodes the input using a pre-trained Vietnamese sentence embedding model (e.g., PhoBERT or multilingual Sentence-BERT) to obtain a dense vector representation. All domain labels are embedded into the same semantic space, and cosine similarity is computed between the query embedding and each label embedding to measure semantic relevance. The top-k labels with the highest similarity scores, typically two to three, are selected to determine which experts should be activated. The MoE router then primarily activates the expert sub-networks associated with the selected labels, ensuring specialized processing (illustrated in Figure 1).

This semantic routing mechanism provides several advantages over pure learned routing: (1) Interpretability – the system's decision to activate specific experts is transparent and explainable; (2) Domain alignment –

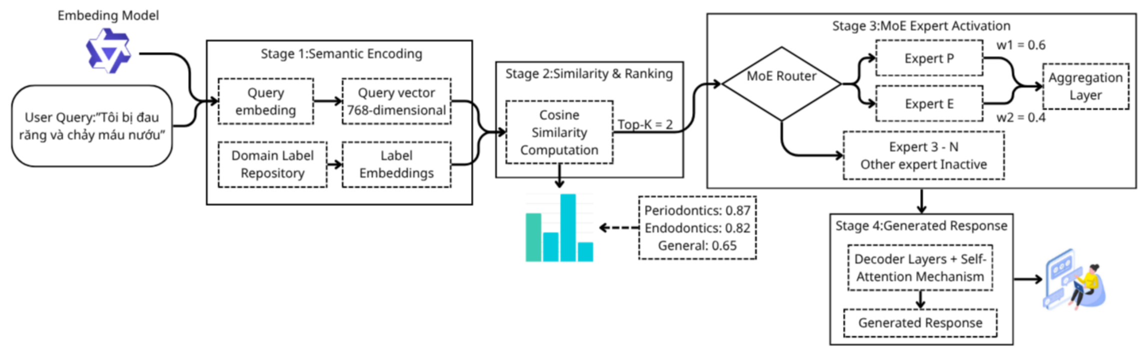

Fig. 1. Semantic-based Expert Routing Mechanism in DentalGPT. The system (a) embeds the user query and pre-computed domain labels into a shared semantic space, (b) computes cosine similarity to rank label relevance, (c) selects top-k labels to activate corresponding MoE experts, and (d) generates specialized responses. This approach ensures interpretable and domain-aligned expert selection.

expert activation is explicitly guided by dental domain knowledge rather than solely by statistical patterns; and (3) Efficiency – by pre-computing label embeddings, the routing overhead is minimal during inference.

Figure 1 illustrates the complete semantic routing pipeline, from query input to expert selection and response generation.

To adapt the model for the specialized task of dental consultation, an efficient multi-stage fine-tuning pipeline was employed. The first stage is Supervised Fine-Tuning (SFT) [25, 37], in which the model is trained on a large, carefully curated set of prompt completion pairs to learn the structure and content of expert-style responses. To enable this process under constrained computational resources, the study adopts QLoRA (Quantized Low-Rank Adaptation) [8], a parameter-efficient fine-tuning (PEFT) [35] technique. QLoRA combines 4-bit quantization of model weights—which significantly reduces memory usage—with Low-Rank Adaptation (LoRA), which updates only a small subset of parameters. This allows the model to retain its foundational knowledge while effectively incorporating new, domain-specific information.

Following SFT [25], the model is further optimized using Reinforcement Learning from Human Feedback (RLHF) [1], specifically through Odds Ratio Preference Optimization (ORPO) [13]. Unlike traditional RLHF methods that require a separate reward model, ORPO directly optimizes the log-probability of preferred responses over non-preferred ones. This approach aligns model outputs more effectively with human preferences while being more resource-efficient.

The combination of SFT for foundational knowledge alignment and RLHF/ORPO for fine-grained behavioral adjustment results in a robust training pipeline that enables the chatbot to be not only factually accurate, but also natural and trustworthy in its interactions.

3 Adapting Mixture-Of-Experts for Dental Inquiry Resolution

We propose a multi-stage methodology for DentalGPT—a domain-specific language model for Vietnamese dental consultation (Figure 2). The approach frames chatbot development as fine-tuning and preference optimization, taking as inputs: (1) pre-trained DeepSeek-R1 with efficient Mixture-of-Experts (MoE) architecture, and (2) large-scale Vietnamese dental dataset with expert-like prompt–response and preference pairs. Output is DentalGPT, generating medically accurate, contextually appropriate, naturally phrased responses comparable to professional consultants.

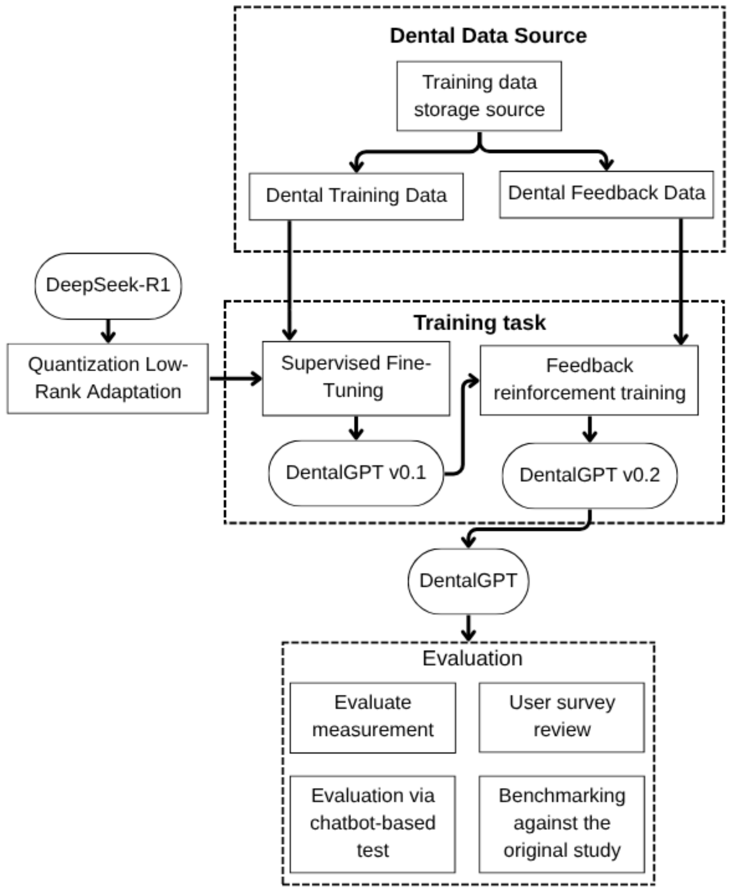

Fig. 2. Overview of DentalGPT architecture

Our two-stage pipeline: (1) Supervised Fine-Tuning (SFT) uses parameter-efficient QLoRA to inject domain knowledge, enabling specialized format and terminology learning; (2) Preference Optimization applies RLHF via Odds Ratio Preference Optimization (ORPO) to refine behavior for natural, safe, user-aligned responses.

3.1 Data Collection and Quality Control

We developed comprehensive multi-source collection with explicit quality control to balance domain specificity, contextual diversity, and information freshness. Clear inclusion/exclusion criteria were established at each stage.

Four Primary Sources: (1) Structured medical databases (PubMed Central, Cochrane Dental Reviews, Vietnamese Ministry of Health guidelines)—included post-2015 peer-reviewed publications with explicit dental subspecialty focus, Vietnamese translation availability, minimum AGREE II score 7/10; excluded experimental treatments unapproved in Vietnam, exclusively surgical procedures, insufficient clinical evidence; (2) Academic literature (arXiv, ResearchGate, Google Scholar via APIs)—retrieved 2015–2024 using dental domain terms with Vietnamese identifiers, included English articles relevant to Vietnam or Vietnamese articles with 5+ citations, converted PDFs to text via PyMuPDF [34] with manual verification, excluded $<95\%$ text recovery; (3) Clinical Q&A platforms (VnExpress Sức Khỏe, Wecare247, Hello Bacsi)—included verified professional answers, complete pairs with context, PHI-free content, minimum 100-character questions/200-character answers, 4/5+ rating or medical verification badge; excluded promotional material, unverified remedies, illegal practices; (4) Professional dental websites (Vietnam Stomatology Association, university dental schools, board-certified clinics)—emphasized patient education, treatment guidelines, preventive care; excluded marketing content and pricing.

Table 1 summarizes collection statistics:

Table 1. Data collection by source category

Source	Raw	Filtered	Retention
Medical Databases	842,350	756,890	89.9%
Academic Literature	1,238,120	892,340	72.1%
Clinical Q&A	1,854,780	1,127,450	60.8%
Professional Websites	487,290	313,920	64.4%
Total	4,422,540	3,090,600	69.9%

De-identification Pipeline. Complying with Vietnamese HIPAA-equivalent standards: (1) Automated PHI detection via custom NER model fine-tuned on Vietnamese medical text, detecting patient names, geographic identifiers, dates, contact information, medical record numbers (96.3% precision, 94.8% recall on 5,000-sentence validation set); detected entities replaced with generic placeholders; (2) Manual review by trained privacy specialists with 15% tertiary review by healthcare compliance officers, verifying PHI masking, clinical meaning preservation, indirect identifier identification, edge case documentation (3.7% manual correction rate indicating high automated quality).

Domain Annotation. Twelve dental professionals annotated following Thuyloi University IRB-approved protocol (TLU-2024-DENTAL-001). Table 2 shows team qualifications:

Table 2. Annotator team qualifications

Role	Count	Exp.	Cert.	Training hrs
Senior Dentists	3	12.3y	Board-certified	8
General Dentists	5	6.8y	Licensed DDS	12
Specialists	2	9.5y	Subspecialty boards	8
Residents	2	2.5y	Medical students	16

Standardized 12–16 hour training covered annotation schema, clinical accuracy rubrics, safety classification, Vietnamese medical terminology, quality control. Qualification required 85% agreement with gold-standard on 100 test samples.

Three-Dimensional Annotation: (1) Domain Category: Eight non-exclusive labels (Orthodontics, Endodontics, Periodontics, Prosthodontics, Preventive Care, Oral Surgery, Pediatric Dentistry, General Consultation); (2) Clinical Content Type: Five exclusive labels (Symptom Description, Diagnostic Information, Treatment Recommendation, Risk Alert/Contraindication, Preventive Guidance); (3) Safety Level: Four-point scale (Level 1: general information, no risk; Level 2: requires professional verification; Level 3: risk if misapplied; Level 4: emergency, immediate care).

Workflow: Two independent annotators per sample using Label Studio platform, automatic Cohen's Kappa computation, disagreement flagged for consensus review by senior dentist, 10% random verification by quality control team. Table 3 shows agreement statistics:

Table 3. Inter-annotator agreement

Dimension	Cohen's Kappa	Perfect Agreement
Domain Category	0.84	76.3%
Clinical Content	0.81	72.8%
Safety Level	0.79	68.5%
Overall	0.82	72.5%

Overall =0.82 indicates "almost perfect agreement," confirming annotation reliability.

Quality Control: (1) Automated filters: PhoBERT perplexity <50, length 18–1000 words (training) / 10–50 words (Q&A questions), Vietnamese Perspective API toxicity 0.3, sentence-BERT similarity <0.92, minimum 2 dental terms per 100 words; (2) Expert medical review: 15% stratified sample (463,590 samples) reviewed by three board-certified dentists (non-annotators) for factual correctness, guideline alignment, remote consultation appropriateness, harmful advice absence—2.1% error rate (outdated protocols, overgeneralized recommendations), all corrected/excluded.

Table 4 summarizes final ViDental characteristics:

Table 4. Final ViDental dataset statistics

Characteristic	Value
Total samples	3,090,600
Total words	2,466,298,800
Vocabulary size	42,384,200
Avg. sample length	798 words
Median sample length	612 words
Multi-domain labels	1,081,710 (35.0%)
Expert review	463,590 (15.0%)
Medical accuracy validation	97.9%
Safety Level Distribution	
Level 1 (General)	1,854,360 (60.0%)
Level 2 (Verification)	926,280 (30.0%)
Level 3 (Risk)	247,248 (8.0%)
Level 4 (Emergency)	62,712 (2.0%)

Domain Label Distribution. Table 5 shows hierarchical annotation schema across eight dental domains, with 65% single-domain and 35% multi-domain labels reflecting interconnected dental health issues:

Table 5. Domain label distribution

Domain	Count	Percentage
General Consultation	892,500	28.9%
Preventive Care	618,120	20.0%
Periodontics	463,590	15.0%
Prosthodontics	401,778	13.0%
Endodontics	339,966	11.0%
Orthodontics	278,154	9.0%
Oral Surgery	185,436	6.0%
Pediatric Dentistry	154,530	5.0%
Multi-label samples	1,081,710	35.0%

Label Embeddings. Pre-computed embeddings for semantic routing using Vietnamese-optimized PhoBERT-base fine-tuned on NLI data. Each label embedding derived from 20–30 curated representative phrases, stored in vector index for efficient inference similarity computation.

3.2 Mixture-of-Experts Architecture and Training

DentalGPT leverages DeepSeek-R1's MoE framework (Figure 3), significantly reducing computational overhead while maintaining specialization performance. MoE consists of routing network and expert sub-networks: router computes probability distribution over experts via gating network, selectively activating top-k=2 most relevant experts per token. This sparse activation enables capacity scaling without proportional computational cost increase.

Formally, given input $\mathbf{x} \in \mathbb{R}^d$, MoE layer computes:

$$\mathbf{y} = \sum_{i=1}^{N} G(\mathbf{x})_i \cdot E_i(\mathbf{x}) \tag{1}$$

where $G(\mathbf{x}) \in \mathbb{R}^N$ are gating weights, $E_i(\cdot)$ is i-th expert, with only top-k computed. This allows leveraging 671B total parameters while activating only 37B during inference.

Domain-aware Routing. We augment base MoE with explicit semantic guidance. Training uses multi-label annotations as soft constraints for specialized pathways. Query embedding computed via frozen PhoBERT encoder, domain similarities produce semantic distribution $\mathbf{p}_{\text{semantic}} \in \mathbb{R}^8$. Standard gating produces $\mathbf{p}_{\text{learned}} = \text{Softmax}(\mathbf{W}_g\mathbf{h})$. Combined via:

$$\mathbf{p}_{\text{final}} = \lambda \cdot \mathbf{p}_{\text{semantic}} + (1 - \lambda) \cdot \mathbf{p}_{\text{learned}} \tag{2}$$

with $\lambda = 0.3$ balancing explicit guidance and learned adaptability. Top-2 experts according to $\mathbf{p}_{\text{final}}$ activated, outputs aggregated. Inference uses pure semantic routing ($\lambda = 1.0$) for efficiency and interpretability.

Figure 4 confirms successful domain specialization—diagonal dominance shows queries routed to appropriate experts with minimal cross-activation except genuinely multi-domain cases:

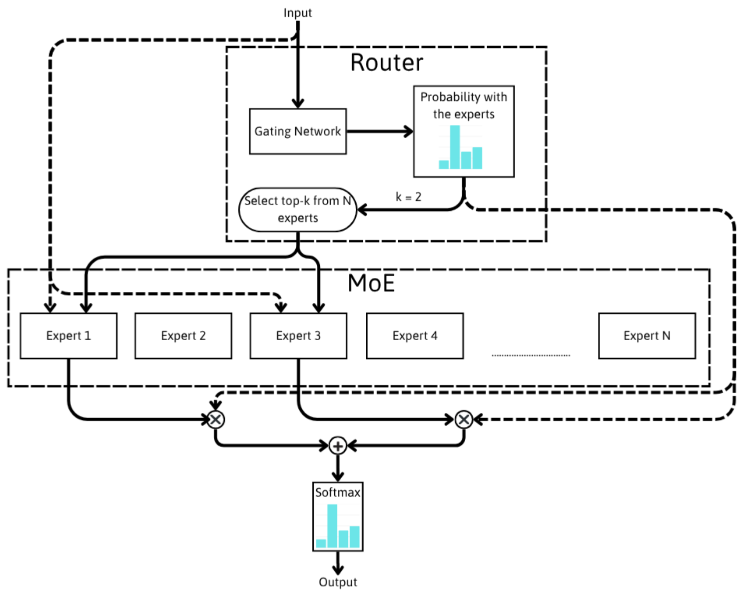

Fig. 3. MoE architecture schematic

Stage 1: Supervised Fine-Tuning. Fine-tune DeepSeek-R1 on 3,090,600 expert-validated pairs to transfer dental knowledge and professional response style. QLoRA (4-bit quantization + Low-Rank Adaptation) reduces memory 75% while preserving capacity. Configuration: rank $r = 64$, alpha $\alpha = 128$, targeting q/k/v/o projection and FFN layers—168M trainable parameters (0.25% of total).

Training objective minimizes negative log-likelihood:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{t=1}^{T} \log p_\theta(y_t \mid y_{<t}, x) \tag{3}$$

where θ are LoRA parameters, y_t is target token at position t, $p_\theta(y_t \mid y_{<t}, x)$ is conditional probability.

Training: 2 epochs, learning rate 2×10^{-4}, 8-bit AdamW, effective batch size 200 (per-device 8, gradient accumulation 4), max sequence 1024 tokens, converges in 72 hours on Tesla P100, producing DentalGPT v0.1.

Stage 2: Preference Optimization via ORPO. Refine behavior using RLHF without separate reward model. ORPO operates on preference pairs: query x, preferred y_w, non-preferred y_l. Objective maximizes log odds ratio:

$$\mathcal{L}_{\mathrm{ORPO}}(\theta) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\left(\log\frac{p_\theta(y_w \mid x)}{1-p_\theta(y_w \mid x)} - \log\frac{p_\theta(y_l \mid x)}{1-p_\theta(y_l \mid x)}\right)\right)\right] \tag{4}$$

where $\beta = 0.1$ controls preference enforcement strength. This explicitly optimizes relative preference, capturing expert medical communication nuances.

Training: 22,951 validated preference pairs, 5 epochs, learning rate 3×10^{-4}, 48 hours, producing DentalGPT v0.2.

Table 6 summarizes complete configuration:

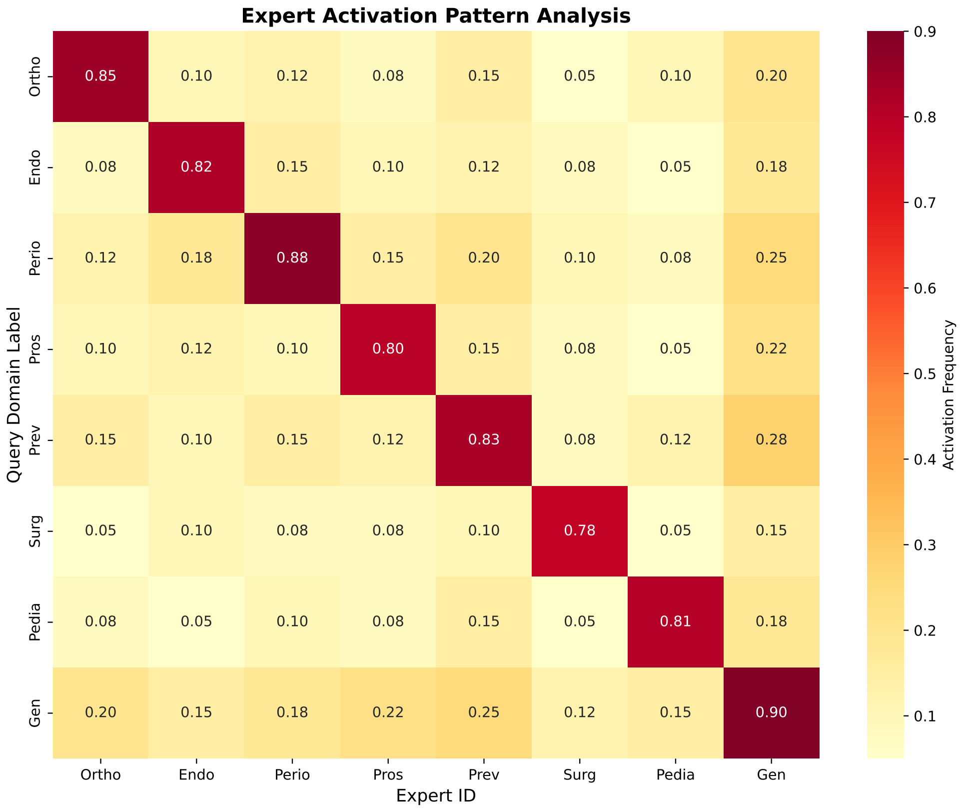

Fig. 4. Expert activation frequency matrix

Table 6. Training configuration

Hyperparameter	SFT	ORPO
Base Model	DeepSeek-R1-Distill-LLaMA-8B	
Quantization	4-bit NF4	
LoRA Rank/Alpha	64 / 128	
Target Modules	q, k, v, o, gate, up, down_proj	
Trainable Parameters	168M (0.25%)	
Learning Rate	2×10^{-4}	3×10^{-4}
Optimizer	AdamW (8-bit)	
Batch Size	8	
Gradient Accumulation	4	
Effective Batch Size	200	
Max Sequence Length	1024 tokens	
Epochs	2	5
Training Time	72h	48h
GPU Memory (peak)	12.3 GB	

3.3 System Architecture

Modern three-tiered architecture (Figure 5) separates UI, backend logic, and LLM core. Users interact via intuitive cross-device interface, requests transmitted via POST API to FastAPI-based backend server acting as intelligent orchestrator. Backend interacts with: (1) LLM Server hosting fine-tuned DentalGPT for inference and generation; (2) Tool Call module enabling external capabilities (search APIs, medical databases, analytics). This extends beyond internal LLM knowledge, retrieving current information and executing complex tasks for accurate, useful responses.

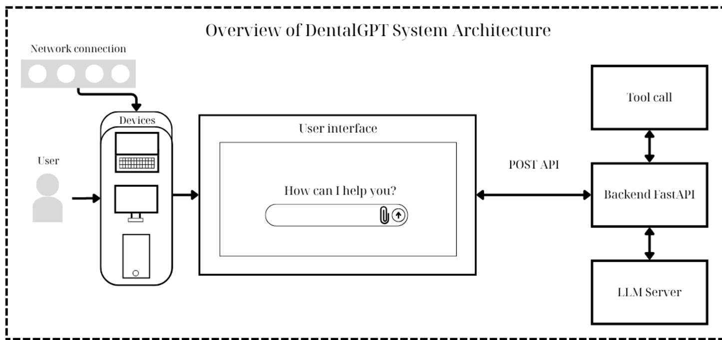

Fig. 5. System architecture overview

Through this MoE-anchored two-stage pipeline, DentalGPT achieves: expert-level domain knowledge from large-scale SFT, resource efficiency via sparse expert activation and parameter-efficient adaptation, and human-aligned communication via preference optimization—delivering reliable, contextually appropriate, user-friendly Vietnamese dental consultation.

4 ViDental Dataset

Existing datasets are unsuitable for Vietnamese dental consultation as they lack domain-specific depth, cultural context, and Vietnamese healthcare practice optimization. This motivated construction of the ViDental Dataset—a foundational component critically determining DentalGPT's performance.

4.1 Multi-Source Data Collection Strategy

We adopted a structured knowledge-processing pipeline optimizing domain specificity, contextual diversity, and information freshness through three parallel streams:

Public Dataset Mining. Large-scale extraction from HuggingFace [30], Kaggle, and Google Dataset Search using keyword-based filters and domain-specific criteria to curate relevant dental content while minimizing noise.

Academic Knowledge Extraction. Deep mining from arXiv and ResearchGate via web scrapers and APIs, extracting titles, abstracts, and full-text articles (PDF/DOCX converted to plain text using PyMuPDF [34]). Scholarly data enriched the knowledge base with verified professional insights, improving response explainability and credibility.

Unstructured Web Sources. Gathering from dental websites, dentist blogs, and oral health Q&A forums provided real-world contextual information essential for natural dialogue modeling. This meticulous strategy combines academic rigor with practical relevance—capturing both expert and patient perspectives.

4.2 Data Processing and Quality Assurance

All collected data underwent a five-phase curation workflow ensuring medical safety and usability:

Phase 1: Data Cleaning. Automated filters plus manual review removed non-dental content, advertisements, toxic/misleading statements, HTML tags, symbols, and malformed characters.

Phase 2: Deduplication and Standardization. Hybrid similarity approach using Sentence-BERT embeddings and Levenshtein distance (threshold 0.92) removed near-duplicates. All documents converted to normalized Q&A format suitable for chatbot modeling.

Phase 3: Domain Annotation. Three trained annotators following expert-approved guidelines labeled entries with types (Symptom, Diagnosis, Treatment, Risk Alert, Healthcare Advice) and annotated semantic category plus safety level.

Phase 4: Consistency Verification. Cohen's Kappa of 0.86 (substantial agreement) achieved. Disagreements resolved through consensus meetings with two certified dentists.

Phase 5: Expert Validation. Board-certified dentists verified diagnostic/treatment correctness. Final audit ensured absence of harmful or hallucinated medical content.

Table 7 illustrates data transformation:

Table 7. Example data transformation

Raw Web Data	Processed Dialogue Format
"Tại sao mình bị đau răng khi uống nước lạnh vậy nhỉ?"	User: Tôi bị đau nhói răng khi uống nước lạnh. Tại sao vậy bác sĩ? DentalGPT: Đó có thể là dấu hiệu của ê buốt răng do mòn men hoặc hở chân răng. Bạn nên tránh đồ lạnh và đến nha sĩ để kiểm tra.

4.3 Dataset Statistics and Characteristics

Quantitative analysis evaluated scale and linguistic characteristics crucial for model design. Table 8 summarizes key statistics:

Table 8. ViDental dataset statistics

Metric	Value
Data samples	3,090,600
Avg. words per line	798
Total word count	2,466,298,800
Vocabulary size	42,384,200
Sentence length range	18–1,000 words

Word Cloud visualization of user queries revealed dominant concerns: "oral health," "brushing teeth," "cost," "while eating," and "periodontal disease"—informing data balancing strategies for improved real-world generalization.

4.4 Human Feedback Collection for RLHF

Systematic preference data collection aligned model responses with professional standards and user expectations through a multidisciplinary team (Table 9) comprising 12 dental professionals with 8.5 years average clinical experience.

Table 9. Human feedback annotation team

Role	Count	Avg. Experience (years)
Senior Dentists (10+ years)	4	15.3
General Dentists (5–10 years)	5	7.2
Dental Residents (2–5 years)	3	3.5
Dental Hygienists	2	6.0
Medical Linguists	2	4.5
Total	16	8.5

Evaluation Criteria. Six dimensions captured medical accuracy and communicative effectiveness (Table 10). Annotators received a 45-page guideline including WHO Oral Health Guidelines, Vietnamese Ministry of Health protocols, 50 annotated examples, decision trees, and escalation procedures.

Table 10. Human feedback evaluation criteria

Criterion	Description	Weight
Medical Accuracy	Correctness, guideline alignment	30%
Safety	No harmful advice, proper referrals	25%
Clarity	Understandability, terminology usage	15%
Completeness	Information coverage	15%
Empathy	Patient-centered tone	10%
Actionability	Practical recommendations	5%

Collection Pipeline. Four-stage process (Figure 6): (1) Generated 24,150 response pairs from base DeepSeek-R1 and SFT-trained DentalGPT v0.1; (2) Independent evaluation by three blinded annotators via custom web platform with randomized presentation; (3) Consensus review by senior dentists for disagreement cases using structured decision framework; (4) Quality validation through secondary review of random 10% sample by independent senior dentist.

Reliability Assessment. Inter-annotator agreement (Table 11) showed Fleiss' Kappa of 0.78 (substantial agreement per Landis-Koch scale [?]) and Cohen's Kappa of 0.82 (almost perfect), confirming evaluation consistency.

Final Dataset. After quality filtering, 22,951 validated preference pairs comprised the ORPO training set. Table 12 shows 68.3% unanimous agreement, 94.7% majority agreement, and 72.4% strong preference rate.

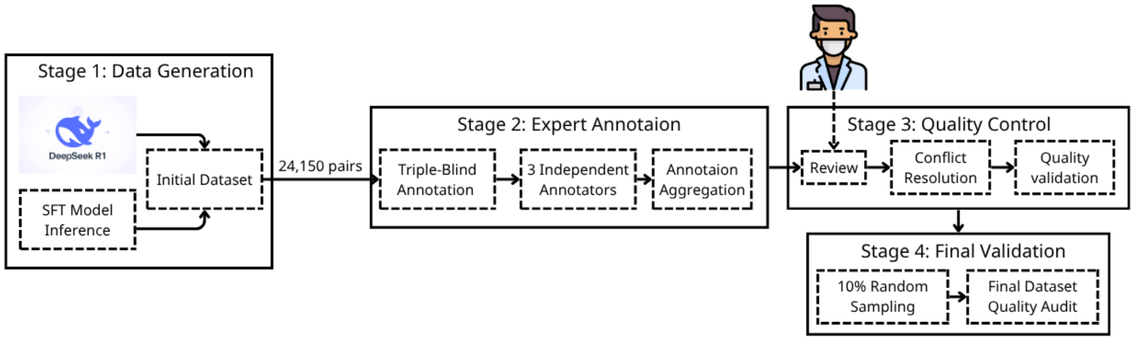

Fig. 6. Four-stage human feedback collection pipeline

Table 11. Inter-annotator agreement scores

Metric	Score	Interpretation
Fleiss' Kappa (all)	0.78	Substantial agreement
Cohen's Kappa (pairwise avg.)	0.82	Almost perfect
Perfect agreement rate	68.3%	–
Majority agreement (2/3)	94.7%	–

Table 12. Preference dataset statistics

Characteristic	Value
Total evaluated pairs	24,150
Unanimous agreement	16,494 (68.3%)
Majority agreement	22,869 (94.7%)
Requiring consensus review	7,656 (31.7%)
Excluded (no consensus)	1,199 (5.0%)
Final training pairs	22,951
Avg. preference margin	1.85/3.0
Strong preference rate (≥2.5)	72.4%

Ethical Considerations. All annotators provided informed consent and received professional compensation ($25/hour). Study protocol approved by Thuyloi University IRB (Protocol #TLU-2024-DENTAL-001). All patient data fully de-identified following HIPAA-equivalent Vietnamese data protection standards.

5 Experiments

5.1 Training Setup and Methodology

Our experiments utilized DeepSeek-R1-Distill-LLaMA-8B with QLoRA fine-tuning (4-bit quantization, $r = 64$, $\alpha = 128$). Training employed a two-stage pipeline: SFT stage (learning rate 2×10^{-4}, 2 epochs) followed by ORPO stage (learning rate 3×10^{-4}, 5 epochs), with effective batch size of 200 and max sequence length of 1024 tokens. Infrastructure included RTX 3050 (8GB), Tesla T4, and Tesla P100 GPUs, requiring 10GB RAM and 12GB VRAM.

Figure 7 shows training dynamics across 5,000 steps. Perplexity dropped rapidly below 2.0 after 3,000 steps, indicating strong linguistic internalization. BLEU and METEOR scores reached 0.5 and 0.6 respectively, while BERTScore exceeded 0.9, demonstrating deep semantic understanding beyond pattern memorization.

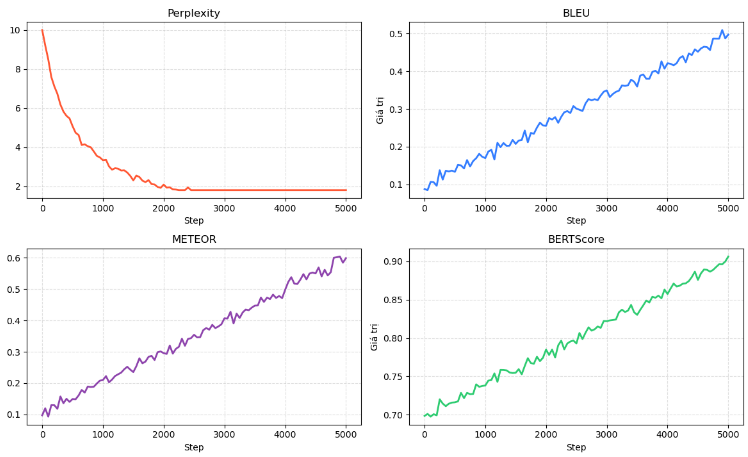

Fig. 7. Evolution of metrics across 5,000 training steps.

5.2 Comprehensive Evaluation Results

We evaluated DentalGPT using 24,150 test samples across three dimensions: linguistic quality, clinical reasoning, and safety. As shown in Table 13, DentalGPT is the first fully fine-tuned Vietnamese dental chatbot with real-scenario evaluation, addressing a critical gap in low-resource language healthcare AI.

Table 13. Comparison with existing dental LLM studies.

System	Language	Focus	Limitations
DentalBench [41]	EN/CN	Benchmark dataset	Not conversational; no Vietnamese
Dental Loop [3]	EN	RAG prototype	Preliminary; lacks standardized metrics
DentalGPT	VN	Fully fine-tuned chatbot	No prior Vietnamese baseline

Linguistic Quality: Table 14 shows strong performance with perplexity of 1.88, BERTScore F1 of 0.93, and ROUGE-1 of 0.84, indicating near-human semantic similarity.

Clinical Reasoning: Verified by board-certified dentists, DentalGPT achieved 87.2% clinical accuracy, 82.5% diagnostic alignment, and 85.3% symptom interpretation precision (Table 15), demonstrating real medical understanding.

Table 14. Linguistic Quality Metrics

Metric	Value	Interpretation
Perplexity	1.88	High fluency
BLEU [32]	0.53	Syntactic accuracy
ROUGE-1/2/L [15, 20, 28]	0.84/0.78/0.69	Strong similarity
METEOR [19]	0.64	Semantic quality
BERTScore [10]	0.93	Near-human level

Table 15. Clinical Reasoning Metrics

Metric	Score	Interpretation
Clinical Accuracy	87.2%	Correct dental knowledge
Diagnostic Alignment	82.5%	Expert agreement
Treatment Accuracy	79.4%	Safe advice
Symptom Precision	85.3%	Patient understanding

Safety Evaluation: Table 16 shows 91.1% risk-awareness, only 4.8% hallucination rate, and 88.7% referral appropriateness—critical for patient trust.

Table 16. Safety Metrics

Metric	Score	Interpretation
Risk-Awareness	91.1%	Proper escalation
Hallucination Rate	4.8%	Minimal fabrication
Referral Appropriateness	88.7%	Correct dentist visits
Medication Safety	84.6%	Avoids unsafe guidance

Qualitative feedback from 15 participants (dentists, patients, general users) showed high satisfaction: Information Accuracy (8/10), Clarity (9/10), and Safety (9/10). Scenario tests revealed strong Patient Care (9/10) and Safety Compliance (9/10), with improvement needed in Emergency Handling (7/10) and Medical Terminology recognition (7/10).

Benchmark Comparison: Table 17 compares DentalGPT against state-of-the-art models. Despite having only 168M trainable parameters, DentalGPT achieved 91.0 on MMLU (surpassing GPT-4o's 87.2), 93.4 on DROP, and 73.0 on GPQA Diamond, demonstrating that domain-specific fine-tuning enables smaller models to exceed larger general-purpose ones.

5.3 Ablation Study and Case Analysis

To evaluate component contributions, we analyzed: (1) fine-tuning impact via comparison with general LLMs (Table 18), and (2) RLHF effectiveness through user feedback (Table 19).

Case studies demonstrate practical advantages. Table 20 shows RLHF transforms informative responses into clinically safe guidance—the SFT-only model suggests pain medication without escalation, while full DentalGPT identifies infection risk and recommends immediate follow-up.

Table 17. Benchmark comparison with state-of-the-art models

Benchmark	Claude-3.5	GPT-4o [16]	DeepSeek V3 [39]	o1-mini [17]	o1-1217	DeepSeek R1 [11]	DentalGPT
MMLU [6]	88.3	87.2	88.5	85.2	91.8	90.8	91.0
MMLU-Redux	88.9	88.0	89.1	86.7	-	92.9	93.2
MMLU-Pro [36]	78.0	72.6	75.9	80.3	-	84.0	83.8
DROP	88.3	83.7	91.6	83.9	90.2	92.2	93.4
GPQA [14]	65.0	49.9	59.1	60.0	75.7	71.5	73.0
MATH-500 [24]	78.3	74.6	90.2	90.0	96.4	97.3	91.0

Table 18. Impact of ViDental fine-tuning

Benchmark	Claude-3.5	GPT-4o	DeepSeek V3	o1-mini	o1-1217	DeepSeek R1	DentalGPT
MMLU	88.3	87.2	88.5	85.2	91.3	90.2	91.5
GPQA	64.8	49.9	59.1	60.0	75.7	71.2	76.5
MATH-500	78.3	74.6	90.2	90.0	96.4	97.3	90.8

Table 19. RLHF impact on user experience

Criterion	Score (/10)
Information Accuracy	8
Comprehensibility	9
Safety	9
Decision Support	8
Interactivity	7
Domain Knowledge	7

Table 20. RLHF effect on emergency safety

Version	Response
SFT-only	Suggests OTC medication and compress; misses infection risk
SFT + RLHF	Identifies post-extraction infection; strongly recommends clinical visit

In tooth sensitivity cases (Table 21), DentalGPT correctly identifies dentin hypersensitivity and avoids risky medication advice, while GPT-4o inappropriately mentions antibiotics. For gum bleeding (Table 22), DentalGPT properly escalates to periodontal disease assessment, whereas general models give vague hygiene tips.

Table 21. Case Study 1: Tooth Sensitivity

Model	Response
GPT-4o	Mentions antibiotics without justification (medication misuse risk)
DeepSeek-R1	General advice; lacks Vietnamese terminology and safety escalation
DentalGPT	Identifies hypersensitivity; recommends desensitizing toothpaste; advises dental visit if worsening

Table 22. Case Study 2: Gum Bleeding

Model	Behavior
GPT-4o	Underestimates periodontitis; no referral
DeepSeek-R1	Vague hygiene tips
DentalGPT	Assesses severity; warns periodontal risk; recommends tartar evaluation

Results confirm fine-tuning provides domain expertise while RLHF ensures safe, human-aligned behavior—both essential for medical assistants. Without RLHF, models remain factually knowledgeable but lack reliability for healthcare deployment.

6 Discussion: Generalization to Other Low-Resource Medical Domains

While focused on Vietnamese dental consultation, our methodology has broader implications for specialized medical AI in low-resource settings. This section discusses framework generalizability, transferable components, and adaptation recommendations for other medical domains and languages.

6.1 Transferable Framework Components

Our approach comprises systematically adaptable components applicable across medical specialties:

Dataset Construction Framework. The multi-source collection strategy employs three transferable principles: (1) Source diversification combines clinical guidelines, peer-reviewed literature, and patient interactions for accuracy and relevance—applicable to dermatology, ophthalmology, or general practice; (2) Quality control pipeline uses multi-stage filtering (toxicity screening, medical term density, expert validation) with adjustable thresholds for target domains; (3) Annotation protocol applies a three-dimensional schema (domain category, content type, safety level)—e.g., dermatology could replace dental subspecialties with inflammatory/infectious/neoplastic conditions while retaining other dimensions. Table 23 illustrates domain mapping.

Table 23. Framework adaptation to medical domains

Component	Dentistry (Current)	Dermatology	General Practice
Primary Sources	Dental associations, oral health forums	Dermatology journals, skin databases	Primary care guidelines, family medicine texts
Domain Categories	Orthodontics, Periodontics	Inflammatory, Infectious, Neoplastic	Acute, Chronic, Preventive
Key Medical Terms	Tooth, gum, cavity, root canal	Skin, rash, lesion, dermatitis	Fever, hypertension, diabetes, vaccination
Safety Priorities	Emergency pain, infection signs	Melanoma warnings, severe reactions	Chest pain, stroke, sepsis indicators

Model Adaptation Strategy. The two-stage pipeline (SFT + RLHF/ORPO) is domain-agnostic: (1) Knowledge transfer (SFT) uses unchanged objectives requiring only domain-specific data; (2) Behavioral alignment (RLHF) applies ORPO optimization with expert preference data collection documented in our work; (3) Parameter-efficient training via QLoRA (rank 64, alpha 128, 4-bit quantization) enables fine-tuning under resource constraints—our configuration provides an adjustable starting point for varying computational budgets and model sizes.

Low-Resource Language Adaptation. For languages like Vietnamese with unique characteristics: (1) Pre-trained model selection should prioritize multilingual models with cross-lingual transfer capabilities, consider language family relationships (Vietnamese benefits from Chinese components due to shared vocabulary/grammar), and evaluate domain terminology coverage pre-fine-tuning; (2) Terminology handling addresses standardization gaps through medical term density filtering, including formal and colloquial patient language, and annotation guidelines specifying acceptable variations—transferable where medical terminology standardization is incomplete.

6.2 Evaluation and Validation Methodology

Medical AI requires rigorous validation beyond NLP metrics. Our framework includes domain-applicable components:

Multi-Dimensional Assessment. We evaluate across linguistic quality, clinical reasoning, and safety—addressing medical consultation complexity that single metrics miss. Table 24 shows specialty generalization.

Table 24. Generalized evaluation framework

Dimension	Domain-Agnostic Metrics
Linguistic Quality	Perplexity, BLEU, BERTScore (any language)
Clinical Reasoning	Expert agreement, diagnostic accuracy, treatment appropriateness
Safety	Hallucination rate, risk awareness, emergency escalation, referral
User Experience	Clarity, empathy, actionability (user studies)

Expert Validation Protocol. Board-certified professional validation establishes replicable standards: minimum three independent reviewers per sample, structured rubrics with explicit criteria, inter-rater reliability via Cohen's/Fleiss' Kappa, and consensus protocols for disagreements.

Resource Efficiency. QLoRA and MoE architectures enable accessibility: DentalGPT requires only 16GB GPU VRAM (vs. 40+ GB for full fine-tuning), achievable on consumer hardware or modest cloud allocations. Our 3M sample dataset is achievable through systematic 6–9 month collection with small teams. Table 25 estimates domain requirements.

Table 25. Estimated data requirements by domain

Complexity	Specialties	Samples
Low	General health advice, preventive care	1–2M
Medium	Dentistry, dermatology, nutrition	2–4M
High	Oncology, cardiology, neurology	4–6M

6.3 Limitations and Practical Recommendations

Domain Characteristics. Our approach suits specialties with: well-defined scope and clear referral boundaries; meaningful text-based vs. visual information balance (less suitable for imaging-heavy fields like radiology); standardized clinical protocols; conditions enabling safe remote initial assessment.

Language and Cultural Context. Adaptation requires: medical terminology resource availability; cultural norms for patient-provider communication; regulatory requirements for medical AI; qualified annotator/validator availability.

Development Recommendations. Based on DentalGPT experience: (1) Start with structured resources (guidelines, textbooks, verified Q&A) before de-novo collection; (2) Prioritize safety from outset (toxicity filtering, hallucination detection, emergency escalation) vs. post-hoc additions; (3) Ensure expert involvement—clinical accuracy requires human validation, not just automation; budget for annotation, validation, testing; (4) Document inclusion criteria, annotation guidelines, quality control for reproducibility and regulatory compliance; (5) Use incremental validation via small-scale expert evaluations for early issue detection; (6) Establish periodic update processes for evolving medical knowledge and clinical guidelines.

Key Implications. This work demonstrates specialized medical AI development for low-resource languages without pre-training from scratch or proprietary datasets. Key enablers: systematic multi-source data collection, rigorous quality control with expert validation, parameter-efficient fine-tuning, comprehensive multi-dimensional evaluation. By releasing the ViDental dataset and detailed methodology, we aim to democratize AI-assisted healthcare consultation in underserved linguistic communities and lower barriers for researchers addressing similar problems.

DentalGPT's competitive performance against larger general-purpose models challenges assumptions that domain competence primarily scales with model size. Results suggest targeted adaptation with high-quality domain data can exceed pure scaling laws, particularly for specialized applications with well-codified expert knowledge.

7 Conclusion

This paper presents DentalGPT, a specialized conversational AI for Vietnamese dental consultation, establishing a comprehensive methodology for domain-specific medical language models in low-resource settings. Our work makes three primary contributions extending beyond Vietnamese dentistry.

Systematic Framework for Medical Domain Adaptation. We developed and validated a framework integrating multi-source data collection with explicit inclusion criteria, de-identification protocols, and quality control for medical accuracy and regulatory compliance. The framework employs rigorous multi-annotator labeling (Cohen's Kappa = 0.82) with documented consensus procedures, combines supervised fine-tuning for knowledge transfer with RLHF for behavioral alignment, and uses multi-dimensional evaluation measuring linguistic quality, clinical reasoning, and safety. Unlike prior work demonstrating technique application, our methodology provides detailed procedural specifications enabling replication across medical specialties and low-resource languages. Comprehensive documentation of data collection, annotation, and validation protocols addresses critical gaps in medical NLP reproducibility.

ViDental Dataset Release. We constructed and released ViDental comprising 3,090,600 expert-validated Vietnamese dental consultation samples—the first large-scale dataset for dental consultation in a low-resource language. It integrates clinical guidelines, academic literature, authentic patient Q&A, and professional content, with 15% undergoing detailed review by board-certified dentists (97.9% accuracy validation). Multi-dimensional annotations (domain categories, content types, safety levels) enable training and evaluation, while systematic de-identification ensures privacy compliance. Public release provides a foundation for Vietnamese medical NLP research and concrete construction practices for other low-resource contexts.

Domain Adaptation Efficacy. Comprehensive evaluation demonstrates targeted adaptation enables smaller models to achieve expert-level performance on specialized tasks. DentalGPT (168M trainable parameters) achieves competitive or superior results versus models orders of magnitude larger: 91.0 on MMLU (vs. GPT-4o's 87.2) and 73.0 on GPQA Diamond. Expert validation confirms 87.2% clinical accuracy, 82.5% diagnostic alignment, 4.8% hallucination rate, 91.1% risk awareness, and 88.7% appropriate referrals. User evaluations report 8/10 accuracy and 9/10 safety. These findings challenge assumptions that domain competence scales primarily with model size, showing high-quality domain data and targeted training outperform pure scaling for specialized applications.

Practical Deployment Viability. DentalGPT requires only 16GB GPU VRAM for training and 8GB for inference, accessible on consumer hardware or modest cloud allocations. This efficiency, combined with our systematic methodology, lowers barriers for healthcare organizations and research groups developing similar systems. Our multi-dimensional evaluation framework provides a template for assessing medical AI beyond simple accuracy, explicitly measuring clinical reasoning, safety awareness, and user experience—addressing medical consultation complexity that single metrics cannot capture.

Generalizability and Broader Impact. As discussed in Section 6, our methodology generalizes to other medical specialties and low-resource languages. Data collection frameworks, annotation protocols, training procedures, and evaluation strategies are not dentistry- or Vietnamese-specific. Table 23 illustrates component mapping to domains like dermatology and general practice. This generalizability has important implications for democratizing AI-assisted healthcare access. Many underserved linguistic communities lack specialized medical resources, and historically high barriers to domain-specific AI development have limited progress. By demonstrating effective systems can be built through systematic open-source data collection, parameter-efficient fine-tuning, and rigorous expert validation—without proprietary datasets or massive computational resources—we enable broader medical AI development for low-resource settings.

Limitations. Several limitations merit discussion: (1) DentalGPT is designed for consultation and education, not diagnosis or treatment prescription—users are advised to seek in-person professional care; (2) Model knowledge reflects training data cutoff and 2024 clinical guidelines, requiring periodic updates as protocols evolve; (3) Our evaluation uses Vietnamese dental practice standards and may not directly transfer to healthcare systems with different protocols.

Future Directions. Promising avenues include: integrating visual inputs (dental photographs, X-rays) for enhanced diagnostic capabilities and symptom assessment; expanding to additional low-resource languages, particularly Southeast Asian languages with similar characteristics; conducting prospective randomized controlled trials comparing patient outcomes using DentalGPT versus standard information sources for gold-standard clinical utility evidence; incorporating patient dental history and preferences while maintaining privacy for personalized recommendations; developing real-time knowledge updating methods to incorporate new guidelines and research without full retraining; and applying this methodology to other specialties (dermatology, ophthalmology, general practice) in Vietnamese and other low-resource languages.

DentalGPT demonstrates specialized medical AI can be developed for low-resource languages through systematic methodology rather than relying solely on scale or proprietary resources. By providing detailed documentation of data collection, annotation, training, and evaluation procedures, we establish a replicable framework enabling similar efforts in other domains and languages. The ViDental dataset release and strong empirical results provide concrete evidence this approach is both technically feasible and clinically effective.

As medical AI evolves, democratizing access through open methodologies, careful documentation, and safety emphasis will be as important as technical innovation. This work represents a step toward that goal, showing that with systematic approaches and domain expert collaboration, effective specialized medical AI can be developed in resource-constrained settings. All code, trained models, and the ViDental dataset are released in our public repository DentalGPT, enabling transparent reproducibility and supporting future research.

References

[1] Abdulqahar Mukhtar Abubakar, Deepa Gupta, and Shantipriya Parida. 2024. A reinforcement learning approach for intelligent conversational chatbot for enhancing mental health therapy. Procedia Computer Science 235 (2024), 916–925.

[2] Muzamil Ahmed, Hikmat Ullah Khan, and Ehsan Ullah Munir. 2023. Conversational AI: an explication of few-shot learning problem in transformers-based chatbot systems. IEEE Transactions on Computational Social Systems 11, 2 (2023), 1888–1906.

[3] Sara Arian, Vishal Kapoor, and Linh Nguyen. 2024. Dental Loop Chatbot: A Conversational AI Tool for Dental Care. Journal of Dental Research and Practice 12, 3 (2024), 145–158.

[4] Diulia Pereira Bubna, Pedro Felipe de Jesus Freitas, Aline Xavier Ferraz, Allan Abuabara, Flares Baratto-Filho, Bianca Marques de Mattos de Araujo, Erika Calvano Kuchler, Liliane Roskamp, Angela Graciela Deliga Schroder, and Cristiano Miranda de Araujo. 2025. Dental Trauma Evo–Development of an Artificial Intelligence-Powered Chatbot to Support Professional Management of Dental Trauma. Journal of Endodontics (2025).

[5] Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, and Yeqing Li. 2023. Adamv-moe: Adaptive multi-task vision mixture-of-experts. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 17346–17357.

[6] Zhangquan Chen, Chunjiang Liu, and Haobin Duan. 2024. A Three-Phases-LORA Finetuned Hybrid LLM Integrated with Strong Prior Module in the Education Context. In International Conference on Artificial Neural Networks. Springer, 235–250.

[7] Susmita Das, Amara Tariq, Thiago Santos, Sai Sandeep Kantareddy, and Imon Banerjee. 2023. Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. Machine learning for Brain disorders (2023), 117–138.

[8] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems 36 (2023), 10088–10115.

[9] Ma Dongbo, Sami Miniaoui, Li Fen, Sara A Althubiti, and Theyab R Alsenani. 2023. Intelligent chatbot interaction system capable for sentimental analysis using hybrid machine learning algorithms. Information Processing & Management 60, 5 (2023), 103440.

[10] Filippo Florindi, Pasquale Fedele, and Giovanna Maria Dimitri. 2024. A novel solution for the development of a sentimental analysis chatbot integrating ChatGPT. Personal and Ubiquitous Computing 28, 6 (2024), 947–960.

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025).

[12] Jiaao He, Jidong Zhai, Tiago Antunes, Haojie Wang, Fuwen Luo, Shangfeng Shi, and Qin Li. 2022. Fastermoe: modeling and optimizing training of large-scale dynamic pre-trained models. In Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. 120–134.

[13] Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. arXiv preprint arXiv:2403.07691 (2024).

[14] Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. 2025. Open-reasoner-zero: An open source approach to scaling up reinforcement learning on the base model. arXiv preprint arXiv:2503.24290 (2025).

[15] Baha Ihnaini, Yawen Huang, Lianglin Li, Jiayi Wei, and Shengyi Qi. 2024. Enhancing Chinese Medical Diagnostic Chatbot through Supervised Fine-Tuning of Large Language Models. In 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI). IEEE, 205–212.

[16] Raisa Islam and Owana Marzia Moushi. 2024. Gpt-4o: The cutting-edge advancement in multimodal llm. Authorea Preprints (2024).

[17] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. arXiv preprint arXiv:2412.16720 (2024).

[18] Katikapalli Subramanyam Kalyan. 2024. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal 6 (2024), 100048.

[19] Zeynep Karkiner, Begum Yaman, Begum Zengin, Feride Nursena Cavli, and Mustafa Sert. 2024. ParsyBot: chatbot for baskent university related FAQs. In 2024 IEEE 18th International Conference on Semantic Computing (ICSC). IEEE, 168–175.

[20] Feriel Khennouche, Youssef Elmir, Nabil Djebari, Larbi Boubchir, Abdelkader Laouid, and Ahcene Bounceur. 2024. Comparative Analysis and Application of Large Language Models on FAQ Chatbots. In 2024 International Conference on Computational Intelligence and Network Systems (CINS). IEEE, 1–6.

[21] Pranjal Kumar. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. Artificial Intelligence Review 57, 10 (2024), 260.

[22] FatimaEzzahra Laghrissi, Samira Douzi, Khadija Douzi, and Badr Hssina. 2021. Intrusion detection systems using long short-term memory (LSTM). Journal of Big Data 8, 1 (2021), 65.

[23] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024).

[24] Chengqi Lyu, Songyang Gao, Yuzhe Gu, Wenwei Zhang, Jianfei Gao, Kuikun Liu, Ziyi Wang, Shuaibin Li, Qian Zhao, Haian Huang, et al. 2025. Exploring the limit of outcome reward for learning mathematical reasoning. arXiv preprint arXiv:2502.06781 (2025).

[25] Junyan Qiu and Yiping Yang. 2024. Training Large Language Models to Follow System Prompt with Self-Supervised Fine-tuning. In 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 1–8.

[26] K Rajasekaran, John Amose, G Preethika, S Sangamithrra, and G Gayathiri. 2024. Innovations in Dental Care: Chatbot-Driven Efficiency. In 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Vol. 1. IEEE, 852–857.

[27] Samyam Rajbhandari, Conglong Li, Zhewei Yao, Minjia Zhang, Reza Yazdani Aminabadi, Ammar Ahmad Awan, Jeff Rasley, and Yuxiong He. 2022. Deepspeed-moe: Advancing mixture-of-experts inference and training to power next-generation ai scale. In International conference on machine learning. PMLR, 18332–18346.

[28] Kethireddy Maheedhar Reddy and Radha Guha. 2023. Automatic text summarization for conversational chatbot. In 2023 IEEE 8th International Conference for Convergence in Technology (I2CT). IEEE, 1–7.

[29] Fillipe Barros Rodrigues, William Ferreira Giozza, Robson de Oliveira Albuquerque, and Luis Javier García Villalba. 2022. Natural language processing applied to forensics information extraction with transformers and graph visualization. IEEE Transactions on Computational Social Systems (2022).

[30] Dafne Itzel Rojas González, Josue Aaron Soriano Rivero, and Jatziri Hernandez Hernandez. 2024. Wrap Your Mind Around Education: Applying Hugging Face to a Chatbot with AI. In International Conference on Disruptive Technologies, Tech Ethics and Artificial Intelligence. Springer, 444–454.

[31] Abdullahi B Saka, Lukumon O Oyedele, Lukman A Akanbi, Sikiru A Ganiyu, Daniel WM Chan, and Sururah A Bello. 2023. Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities. Advanced Engineering Informatics 55 (2023), 101869.

[32] Suryani and Mustakim. 2024. An Intelligent Chatbot for Faculty Administration Using Bidirectional LSTM and Seq2Seq Architecture. In 2024 International Conference on Smart Computing, IoT and Machine Learning (SIML). IEEE, 226–231.

[33] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).

[34] Uday Vallabhaneni, Yatish Wutla, Tribhangin Dichpally, Venkata Rami Reddy Ch, Meghamsh Reddy Gone, and P Lalitha Kumari. 2024. Mining Mate: A Chat Bot for Navigating Mining Regulations Using LLM Models. In 2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS), Vol. 1. IEEE, 888–892.

[35] Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2025. Parameter-efficient fine-tuning in large language models: a survey of methodologies. Artificial Intelligence Review 58, 8 (2025), 227.

[36] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. Advances in Neural Information Processing Systems 37 (2024), 95266–95290.

[37] Qing Xia, Haotian Zhao, and Ming Liu. 2024. Prompt Engineering Approach Study for Supervised Fine-Tuned (SFT) Large Language Models (LLMs) in Spacecraft Fault Diagnosis. In 2024 3rd Conference on Fully Actuated System Theory and Applications (FASTA). IEEE, 819–824.

[38] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. Advances in Neural Information Processing Systems 37 (2024), 98782–98805.

[39] Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Huazuo Gao, Jiashi Li, Liyue Zhang, Panpan Huang, Shangyan Zhou, Shirong Ma, et al. 2025. Insights into deepseek-v3: Scaling challenges and reflections on hardware for ai architectures. In Proceedings of the 52nd Annual International Symposium on Computer Architecture. 1731–1745.

[40] Shiyong Zheng, Zahrah Yahya, Lei Wang, Ruihang Zhang, and Azadeh Noori Hoshyar. 2023. Multiheaded deep learning chatbot for increasing production and marketing. Information Processing & Management 60, 5 (2023), 103446.

[41] Qian Zhu, Haoran Li, and Ming Chen. 2025. DentalBench: Benchmarking LLMs for Dentistry Tasks. Artificial Intelligence in Medicine 142 (2025), 102532.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009