AyurParam: A State-of-the-Art Bilingual Language Model for Ayurveda

Mohd Nauman, Sravan Gvm, Vijay Devane, Shyam Pawar, Viraj Thakur, Kundeshwar Pundalik, Piyush Sawarkar, Rohit Saluja, Maunendra Desarkar, Ganesh Ramakrishnan

BharatGen Team

Current large language models excel at broad, general-purpose tasks, but consistently underperform when exposed to highly specialized domains that require deep cultural, linguistic, and subject-matter expertise. In particular, traditional medical systems such as Ayurveda embody centuries of nuanced textual and clinical knowledge that mainstream LLMs fail to accurately interpret or apply. We introduce **AyurParam-2.9B**, a domain-specialized, bilingual language model fine-tuned from Param-1-2.9B using an extensive, expertly curated Ayurveda dataset spanning classical texts and clinical guidance. AyurParam's dataset incorporates context-aware, reasoning, and objective-style Q&A in both English and Hindi, with rigorous annotation protocols for factual precision and instructional clarity. Benchmarked on BhashaBench-Ayur, AyurParam not only surpasses all open-source instruction-tuned models in its size class (1.5–3B parameters), but also demonstrates competitive or superior performance compared to much larger models. The results from AyurParam highlight the necessity for authentic domain adaptation and high-quality supervision in delivering reliable, culturally congruent AI for specialized medical knowledge.

Date: November 5, 2025

Correspondence: kundeshwar.pundalik@tihiitb.org

1 Introduction

Large language models (LLMs) have revolutionized natural language processing by enabling unprecedented understanding, generation, and reasoning across diverse text corpora and tasks. However, despite their general-purpose prowess, these models often fall short when applied to specialized knowledge domains that require deep contextual expertise and cultural awareness. This gap is particularly pronounced in traditional medical systems such as Ayurveda — a comprehensive holistic healthcare system with roots stretching back millennia and rich embedded linguistic, clinical, and philosophical complexities.

Mainstream LLMs are typically trained on extensive but heterogeneous datasets that do not capture the linguistic nuances, semantic specificity, or culturally grounded medical knowledge embedded in Ayurvedic texts and practices. Consequently, unadapted LLMs struggle with accurate interpretation, reasoning, and generation of domain-specific clinical and wellness information, undermining trustworthiness and practical utility in sensitive healthcare contexts. Additionally, many existing generalized models lack bilingual support for Indian languages, further limiting relevance and accessibility for Ayurveda's practitioner and patient populations.

To bridge these challenges, domain-specialized LLMs have emerged as critical tools—fine-tuned or pre-trained on carefully curated, high-quality corpora rich in domain-specific terminology, structured instructional formats, and validated knowledge sources. These tailored models significantly improve accuracy, interpretability, and relevance by internalizing domain conventions, reasoning frameworks, and linguistic subtleties beyond generic LLM capabilities. Applications in clinical decision support, diagnostics, and patient communication have shown promising results, emphasizing the need for linguistic and cultural alignment in medical AI systems.

Building on this paradigm, we present AyurParam, a bilingual large language model specialized for Ayurveda that leverages Param-1-2.9B-Instruct as a base and is fine-tuned on a meticulously assembled

dataset comprising digitized classical manuscripts, clinical guidelines, objective assessments, and reasoning-driven queries in both English and Hindi. The dataset is designed with explicit supervision formats, combining dialogue-style prompt-completion pairs and domain expert annotations to enhance instruction-following and factual grounding. AyurParam excels in contextual understanding, reasoning about Ayurvedic principles such as dosha imbalances and samprapti, and delivering culturally nuanced responses across consultation, education, and research use cases.

We comprehensively evaluate AyurParam on the BhashaBench-Ayur benchmark, demonstrating state-of-the-art performance among open-source LLMs in the 1.5–3 billion parameter range and competitive results against substantially larger models. The contributions underscore the transformative potential of domain-specialized LLMs in enabling trustworthy, language-aware AI for traditional medicine, facilitating wider digital accessibility to Ayurveda's rich heritage while supporting modern clinical workflows.

2 Related Work

Domain-Specialized Large Language Models Early work on foundation models focused on broad-domain generalization, but recent advances reveal the necessity of domain adaptation for deep semantic understanding and accurate task execution in specialized fields [Ling et al., 2023]. The comprehensive survey by Ling et al. (2023) categorizes domain adaptation strategies into direct fine-tuning, targeted pre-training, and external data augmentation, each facilitating models with richer domain vocabulary, reasoning methods, and error resilience. Biomedical LLMs such as PubMedGPT and BioGPT employ massive medical corpora to internalize discipline-specific knowledge and terminology, yielding superior performance in medical QA and summarization tasks compared to generalized LLMs. Fine-tuning on well-structured domain datasets and leveraging synthetic instruction instances are critical for achieving both higher accuracy and trustworthiness in high-stakes applications.

Instruction Tuning and Task Adaptation Instruction tuning has emerged as a pivotal method to align LLMs with human expectations, safety constraints, and domain ontologies [Alaa et al., 2025; IBM, 2024]. Recent studies explore multi-task and prompt-style frameworks, including FLAN, SuperNI, and Self-Instruct, to expose models to diverse instructional templates and domain-conditional objectives [Wei et al., 2021; Wang et al., 2022]. Meta-tuning, domain-aware task conditioning, and reinforcement learning (e.g., RLHF, DPO) offer further enhancements for flexibility, factual consistency, and safer outputs. In educational and scientific fields, instruction-tuned LLMs exhibit dramatically improved reasoning and answer quality, but achieving similar success in medical and traditional knowledge domains demands curated dataset design, multilingual coverage, and expert validation.

Medical and Multilingual LLM Benchmarks Medical and healthcare adaptation of LLMs is rapidly advancing, especially where multilingual coverage and cultural grounding are needed [Qiu et al., 2024]. Benchmarks such as MMedBench and Swedish Medical LLM Benchmark have been introduced to systematically evaluate clinical reliability, multilingual QA performance, and safety in realistic environments. Adoption of domain-specific benchmarks and layered evaluation protocols is advocated to address limitations in existing medical leaderboards and guide ethical, scalable AI deployment in sensitive contexts [Alaa et al., 2025]. Studies consistently reveal that models fine-tuned with localized, expert-curated corpora outperform generalist LLMs and demonstrate practical utility in clinical consulting, wellness, and patient communication.

AI for Ayurveda and Traditional Medicine While most attention has centered on Western biomedical LLMs, a small but growing body of work addresses the challenges of adapting AI and LLMs to traditional medical systems like Ayurveda [Padia et al., 2025; Rathor et al., 2024]. Early rule-based and narrow ML applications offered limited interpretability and reasoning, but newer models—including AyurGPT and IRGPT—leverage domain-specific pre-training and multilingual instruction-tuning to support consultation and diagnosis across languages. These models process Sanskrit and regional linguistic data, enhance reasoning about dosha, dhatu, and treatment logic, and are benchmarked with domain-specific metrics, yet their scope and quality remain underexplored at scale. The work done builds on these foundations, presenting AyurParam as the first bilingual, instruction-tuned LLM extensively benchmarked for authentic, context-rich performance in Ayurveda.

3 Data Preparation

Our data preparation methodology follows a systematic pipeline encompassing taxonomy establishment, corpus collection, OCR processing, quality assurance, and knowledge-grounded Q&A generation (Figure 1) .

Each stage incorporates domain-specific constraints and validation protocols to ensure the resulting dataset meets the rigorous requirements for specialized Ayurvedic instruction tuning.

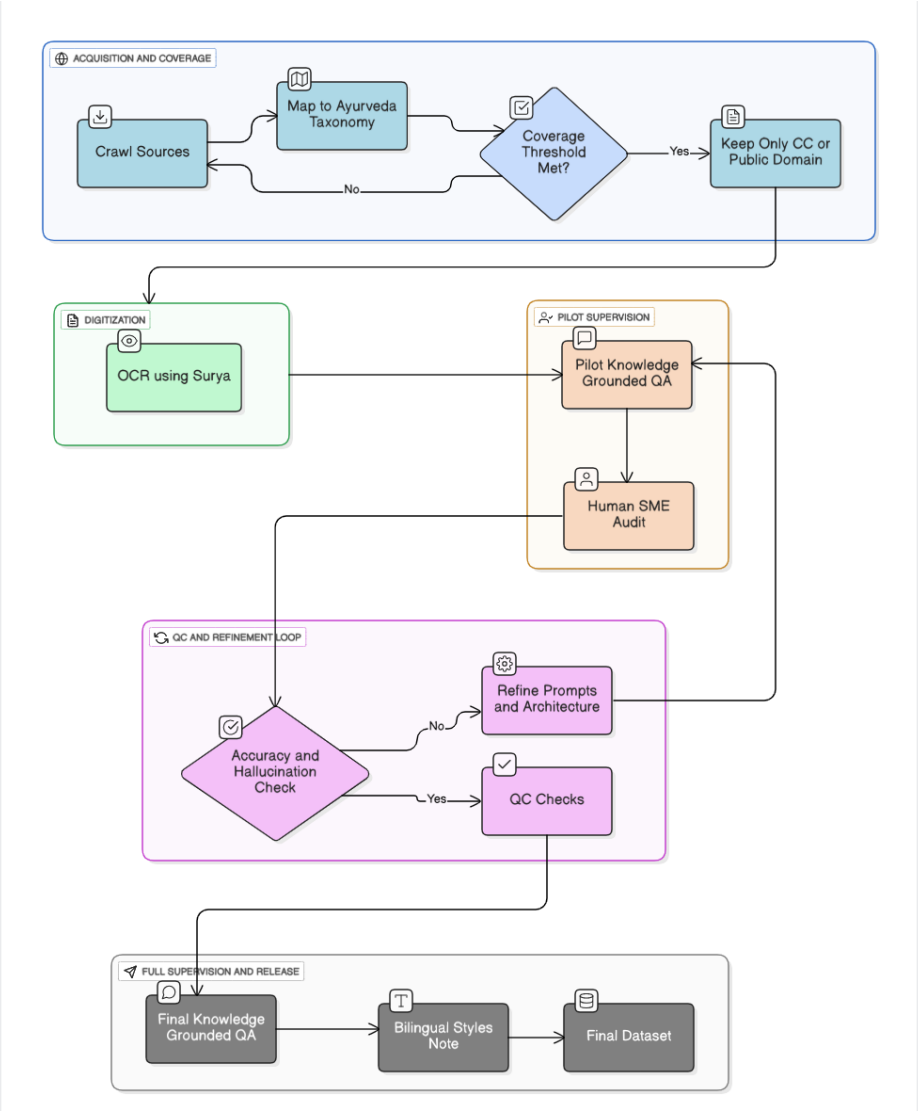

Figure 1: Data preparation pipeline

3.1 Taxonomy of Ayurvedic Domains

Before collecting or processing texts, we established a curriculum–aligned taxonomy to ensure that all major branches of Ayurveda are represented. This step prevents over-representation of easily available material (e.g., Panchakarma manuals) and ensures inclusion of essential domains that are often underrepresented in digital form.

The taxonomy was derived from the official *BAMS* undergraduate curriculum, postgraduate *MD/MS* specializations, and canonical compendia (Charaka, Sushruta, Ashtanga Hridaya, Kashyapa Samhita). It serves three purposes:

1. Acts as a retrieval lens when querying archives in Devanagari, IAST, and English transliteration.
2. Enforces per-domain quotas to maintain balance and prevent skew.
3. Defines strata for downstream evaluation and error analysis.

Domains included:

- *Foundations*: Ashtang Hridaya, Padarth Vigyan, selected Sanskrit commentaries.
- *Anatomy and Physiology*: Rachana Shaarir, Kriya Shaarir.
- *Classical Compendia*: Charaka Samhita, Sushruta Samhita.
- *Clinical Disciplines*: Kayachikitsa, Panchakarma, Shalya Tantra, Shalakya Tantra.
- *Pharmacology and Formulations*: Dravyaguna, Rasa Shastra, Bhaishajya Kalpana.
- *Pathology and Toxicology*: Rog Nidan, Agad Tantra.
- *Specialties*: Kaumarbhritya (Balrog), Strirog and Prasuti Tantra, Swasthavritta.

This taxonomy acted as the scaffold against which all subsequent collection, license filtering, OCR, and Q&A generation were aligned.

3.2 Corpus Collection and License Governance

Acquisition. We collected approximately 1,000 Books and doc spanning classical compendia, modern commentaries, and clinical manuals. Of these, ∼600 were sourced from public digital archives and open libraries (e.g., Archive.org, eGangotri), while ∼400 were drawn from institutional or government repositories. The collection spans Sanskrit originals, Hindi and Marathi translations, and English editions, ensuring bilingual coverage while preserving classical terminology. In total, the corpus comprises ∼150,000 pages (∼54.5M words), forming one of the largest curated Ayurvedic text datasets.

In total, the corpus comprises ∼150,000 pages (∼54.5M words), forming one of the largest curated Ayurvedic text datasets.

Figure 2 shows the language-wise distribution of crawled source documents, with PDFs collected in Sanskrit (Devanagari script), Hindi, Marathi, and English, ensuring comprehensive multilingual coverage of Ayurvedic literature.

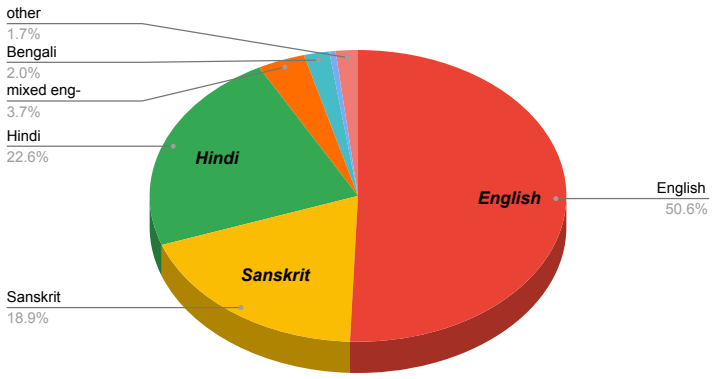

Figure 2: Language-wise distribution of the collected corpus across Sanskrit, Hindi, Marathi, and English sources

License governance. Each item was catalogued into a *license ledger* that records edition-level metadata, including title, authors, translators, publication year, source URL, language tags, and declared rights. Only texts with clear open licenses (e.g., CC0, CC-BY, public domain) were retained for training. Works with ambiguous or restrictive licenses were preserved as *shadow entries*— metadata retained for completeness but text excluded from supervision. This approach preserves reproducibility while respecting legal and ethical constraints.

Duplicate handling. Multiple editions of the same compendium are frequent in Ayurveda (e.g., Charaka Samhita with different commentaries). To avoid redundancy, we applied page-level near-duplicate detection using character n-gram hashing and MinHash signatures. Unique editions were retained, and their relationships were cross-linked in metadata to support lineage tracking.

This two-step process of broad collection followed by strict license filtering ensured that the final dataset is both comprehensive and reproducible, while remaining legally and ethically compliant.

3.3 OCR and Post-OCR Processing

Pre-processing. Many collected works were available only as scanned PDFs or images with significant variance in quality. Before OCR, each page was conditioned through deskewing, contrast normalization, and margin cropping. This ensured more consistent recognition across heterogeneous sources.

OCR engine. For Indic scripts (Devanagari: Sanskrit, Hindi, Marathi), we employed **Surya OCR**, which provides state-of-the-art performance on multilingual text lines. English pages were processed in the same pipeline for consistency. Page-level OCR confidence scores were recorded as metadata for subsequent filtering.

Normalization. Post-OCR text was standardized through several layers of cleaning:

- Unicode NFC normalization, repair of common Devanagari ligatures and numerals.
- Removal of headers, footers, and marginalia to reduce noise.
- De-hyphenation and whitespace normalization.
- Segmentation of passages with language tags (`san-Deva`, `hi-Deva`, `en-Latn`).
- Alignment with canonical divisions when detectable (e.g., Sūtrasthāna, Nidānasthāna, Vimānasthāna, Uttaratantra).

Quality assurance. We monitored OCR accuracy using a combination of automatic and heuristic checks: (i) mean and median confidence per page; (ii) crude character error rate estimation via lexicon-free sampling; (iii) Indic-specific heuristics such as akshara merges or danda detection. Low-confidence pages were flagged for exclusion or routed to stricter cleaning.

This stage yielded a normalized, linguistically tagged corpus suitable for downstream Q&A generation, with provenance and quality metadata attached at the page level.

3.4 Knowledge-Grounded Data Generation

Knowledge-grounded synthesis. Cleaned passages were transformed into supervised training examples using high-capacity LLMs (primarily **Qwen-3 235B**) under strict constraints: responses were required to be derivable from the provided span, with no external elaboration or prescriptive advice.

Human-in-the-loop calibration. To improve reliability, ten representative books were manually reviewed by Ayurveda practitioners. For each book, 50–300 pages were randomly sampled and the corresponding Q&A pairs analyzed. Experts identified over-generalization, implicit assumptions, and unsupported reasoning, which informed iterative refinements to the synthesis policy. This loop significantly reduced hallucination and improved epistemic fidelity.

Quality assurance. Reliability was enforced through a staged validation pipeline:

- *Rule-based filters:* Every item was checked for JSON schema validity, minimum/maximum answer length, banned phrases (e.g., prescriptive treatment advice), and symbol consistency.
- *Evidence anchoring:* Answers were required to cite support spans from the passage. We measured lexical overlap and coverage ratios to detect unsupported generations.
- *Selective LLM adjudication:* Only uncertain cases—those failing thresholds or flagged by overlap heuristics—were escalated to an LLM-as-judge for groundedness and contradiction checks.
- *Targeted human audits:* Stratified samples from uncertain cases, low-OCR-confidence pages, and high-stakes domains (e.g., *Śalya*, *Śālākya*) were reviewed by practitioners, with inter-annotator agreement monitored for consistency.

Final dataset. The resulting corpus contained approximately 4.75M grounded Q&A pairs, distributed as:

- **Q&A pair (EN + HI)**: ∼1.27M pairs.
- **Objective/MCQ**: ∼0.9M pairs.
- **Multi-turn reasoning**: ∼1.51M pairs.

- **Contextual Q&A (span-level comprehension)**: ∼1.07M pairs.

This balanced coverage ensures both breadth (terminology, principles, clinical applications) and depth (reasoning, multi-turn consistency). As shown in Figure 3, the dataset exhibits a balanced distribution across question types.

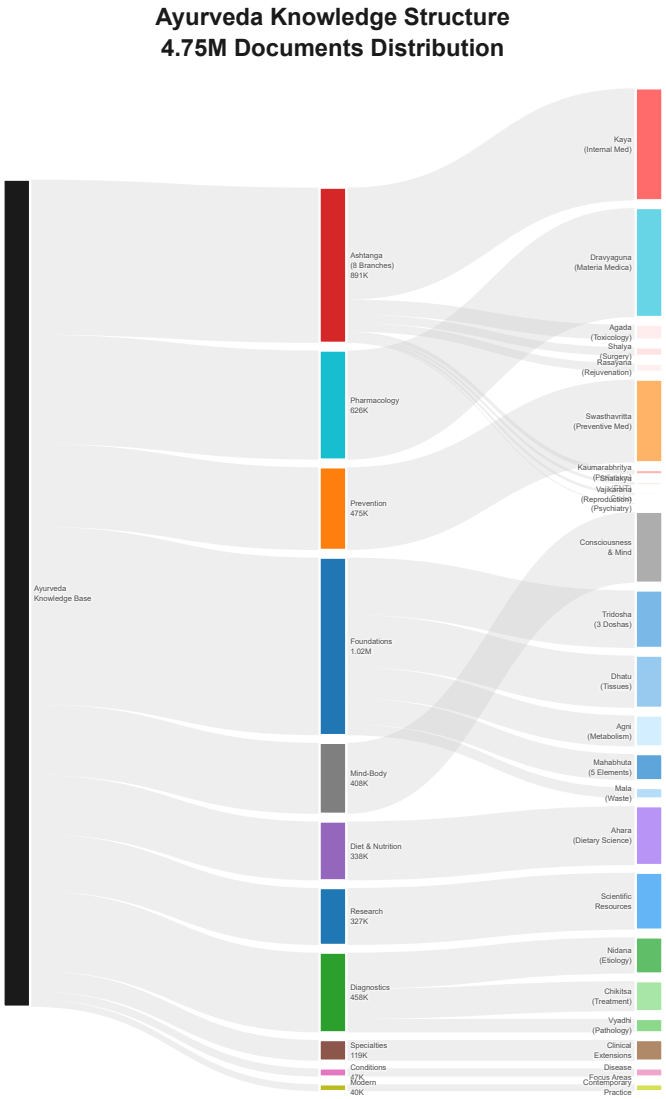

Figure 3: QA-Distribution

3.5 Ethical and Legal Considerations

All primary texts were sourced from public digital archives (e.g., Archive.org, eGangotri) and institutional repositories that explicitly provide open access or carry permissive licenses. A license filter was applied to exclude restricted or unclear materials, ensuring that the final corpus is legally redistributable for research and educational purposes.

Sensitive sources such as personal health records, unpublished manuscripts, and private clinical notes were explicitly excluded to prevent leakage of private or identifiable information. Additionally, Q&A generation was framed to support research and education rather than clinical prescription, reflecting the model's intended use.

We note that classical Ayurvedic texts themselves embed cultural and historical biases (e.g., gender roles in physician narratives). While AyurParam may reflect such tendencies, the instruction-tuning framework reduces prescriptive phrasing and emphasizes factual recall.

4 Training

4.1 Training Data

The supervised fine-tuning (SFT) dataset for AyurParam was built from the curated Ayurveda corpus discussed in Section 3. The focus during dataset construction was twofold: ensuring broad coverage across classical Ayurvedic texts, clinical subdomains, and wellness knowledge, and enabling strong instruction-following capabilities.

Conversational Alignment Data. For conversational alignment, the data was organized in a dialogue-style format using `<user>` and `<assistant>` markers. Each `<user>` tag contained the query (question, prompt, or instruction), while the `<assistant>` tag held the model's reply. To provide explicit supervision, responses were further enclosed within `<actual_response>` and `</actual_response>`, clearly indicating the ground-truth output for fine-tuning. This structure helped maintain a clean separation between prompts, user inputs, and reference answers.

The dataset integrated multiple generation methods, including:

- **Context-based Q&A:** Questions tied to page-level or passage-level content from traditional Ayurvedic manuscripts.
- **Instructional Prompts:** Generic tasks designed to elicit structured reasoning.
- **Situational Dialogue:** Conversational, scenario-driven exchanges to improve alignment.
- **Objective-style tasks:** Multiple-choice, fill-in-the-blank, and structured assessments for clinical and academic use.
- **Reasoning-driven questions:** Prompts requiring analytical thinking around Ayurvedic principles such as dosha imbalances, samprapti (disease progression), and treatment logic etc.,
- **General knowledge Q&A:** Covering terminology, principles, and practices in both English and Hindi.

Altogether, the training set contained about 4.75 million examples, spanning single-turn and multi-turn dialogues. This design not only supported factual recall but also strengthened reasoning and inference skills—key requirements for a specialized conversational agent in Ayurveda.

4.2 Training Setup

We fine-tuned the `Param-1-2.9B` [3] base model using the Hugging Face TRL framework in a supervised fine-tuning (SFT) configuration. Training was performed on a multi-node NVIDIA H100 GPU cluster with `torchrun`. The main hyperparameters and settings are summarized below:

- **Global batch size:** 1024 (micro-batch size = 4; gradient accumulation = 32)
- **Learning rate:** 5×10^{-6} with linear decay and warmup
- **Training epochs:** 3
- **Precision:** bfloat16 mixed precision
- **Vocabulary:** 256k tokens + 6 task-specific tokens (`<user>`, `<assistant>`, `<context>`, `<system_prompt>`, `<actual_response>`, `</actual_response>`)

The supervised fine-tuning corpus consisted of approximately 4.75M samples. Training required roughly two days on a single H100 node. To better support both single-turn and multi-turn Ayurvedic instruction-following, we employed custom bilingual templates (English and Hindi).

5 Model Performance

To rigorously evaluate domain specialization, we benchmarked `AyurParam-2.9B` on `BhashaBench-Ayur` (BBA), India's first large-scale evaluation suite for Ayurvedic AI systems [2]. BBA consists of 14,963 exam-style questions spanning 15+ subject domains, covering both English (9,348 questions) and Hindi (5,615 questions). The dataset integrates authentic government and institutional examinations across India, reflecting the same rigor and breadth expected of BAMS graduates and postgraduate Ayurvedic training. It evaluates factual recall, clinical reasoning, therapeutic principles, and interpretability through diverse question formats (MCQ, assertion-reasoning, fill-in-the-blanks, and match-the-column).

Table 1: Overall performance comparison on BBA dataset. Results show accuracy (%) across different model sizes. Our AyurParam-2.9B-Instruct model achieves the best performance among similar-sized models and competitive results compared to much larger models.

Similar Range Models			
Model	**BBA**	**BBA English**	**BBA Hindi**
AyurParam-2.9B-Instruct	**39.97**	**41.12**	**38.04**
Llama-3.2-3B-Instruct	33.20	35.31	29.67
Qwen2.5-3B-Instruct	32.68	35.22	28.46
granite-3.1-2B	31.10	33.39	27.30
gemma-2-2B-it	28.40	29.38	26.79
Llama-3.2-1B-Instruct	26.41	26.77	25.82
Larger Models			
AyurParam-2.9B-Instruct	**39.97**	**41.12**	**38.04**
gemma-2-27B-it	37.99	40.45	33.89
Pangea-7B	37.41	40.69	31.93
gpt-oss-20B	36.34	38.30	33.09
Indic-gemma-7B-Navarasa-2.0	35.13	37.12	31.83
Llama-3.1-8B-Instruct	34.76	36.86	31.26
Nemotron-4-Mini-Hindi-4B-Instruct	33.54	33.38	33.82
aya-23-8B	31.97	33.84	28.87

On this benchmark, `AyurParam-2.9B` achieved state-of-the-art accuracy among models in the 1.5–3B parameter range, outperforming instruction-tuned baselines such as LLaMA-3.2-3B and Qwen2.5-3B. Notably, despite being significantly smaller, `AyurParam-2.9B` demonstrated competitive performance with models in the 7B–27B range, underscoring the efficiency of targeted domain specialization (Table 1).

Table 2: Performance breakdown by question difficulty. Results demonstrate that AyurParam-2.9B-Instruct maintains strong performance across all difficulty levels, with particularly notable results on easy questions.

Similar Range Models						
Difficulty	**AyurParam-2.9B**	**Llama-3B**	**Qwen-3B**	**Granite-2B**	**Gemma-2B**	**Llama-1B**
Easy	**43.93**	36.42	35.55	33.90	29.96	27.44
Medium	**35.95**	29.66	29.57	28.06	26.83	25.23
Hard	**31.21**	28.51	28.23	26.81	24.96	25.39

Larger Models						
Difficulty	**AyurParam-2.9B**	**Gemma-27B**	**Pangea-7B**	**Llama-8B**	**Indic-7B**	**Aya-8B**
Easy	**43.93**	43.47	41.45	39.43	38.54	35.51
Medium	**35.95**	31.90	32.94	29.36	31.72	28.29
Hard	31.21	30.78	**31.77**	30.50	27.23	25.11

The model showed strong gains on clinically relevant domains such as *Kayachikitsa* and *Dravyaguna*,

while also maintaining robust performance in medium- and high-difficulty questions, where general-purpose models often struggled (Table 2). These results highlight that high-quality, domain-specific supervision and carefully designed instruction-tuning protocols enable small- and mid-scale models to achieve outsized gains in specialized benchmarks.

Evaluation on BBA also revealed limitations: performance in Hindi lagged behind English, and reasoning-intensive domains such as *Panchakarma & Rasayana* and *Ayurvedic Literature* remained challenging. This points to the need for improved multilingual coverage and deeper alignment with classical knowledge representations, which we consider essential for the next phase of research.

Table 3: Performance analysis across different question types. AyurParam-2.9B-Instruct shows strong performance on MCQ questions and competitive results across all question formats.

Similar Range Models					
Type	**Llama-1B**	**Qwen-3B**	**Llama-3B**	**AyurParam-2.9B**	**Gemma-2B**
Assert./Reason.	59.26	51.85	40.74	44.44	33.33
Fill blanks	26.97	29.21	34.83	29.78	32.02
MCQ	26.34	32.70	33.17	40.12	28.33
Match col.	26.83	29.27	29.27	24.39	36.59

Larger Models					
Type	**Pangea-7B**	**Gemma-27B**	**AyurParam-2.9B**	**Llama-8B**	**Indic-7B**
Assert./Reason.	62.96	55.56	44.44	29.63	59.26
Fill blanks	24.16	35.96	29.78	26.97	35.39
MCQ	37.53	37.98	40.12	34.83	35.10
Match col.	34.15	39.02	24.39	46.34	31.71

The question type analysis reveals nuanced performance patterns across different evaluation formats (Table 3). `AyurParam-2.9B` excels particularly in multiple-choice questions (MCQ), achieving the highest accuracy (40.12%) among all compared models, including those with significantly more parameters. This strength in MCQ performance is especially valuable for Ayurvedic assessment, as it reflects the model's ability to discriminate between closely related concepts and therapeutic approaches—a critical skill for clinical decision-making.

6 Conclusion

This work introduced **AyurParam**, a bilingual, domain-specialized large language model for Ayurveda, fine-tuned from the Param-1-2.9B-Instruct foundation model. By leveraging a rigorously curated corpus of classical texts, clinical manuals, and bilingual instructional data, AyurParam achieves high accuracy across a diverse set of question types and difficulty levels, outperforming similar-sized open-source models and remaining competitive with much larger models.

Extensive evaluation on **BhashaBench-Ayur** demonstrates that careful domain adaptation, high-quality supervision, and culturally grounded pretraining are critical for enabling small- to mid-scale models to perform reliably on complex, specialized knowledge tasks. In particular, AyurParam shows robust performance in multiple-choice questions, reasoning-intensive prompts, and multi-turn Q&A, highlighting its potential as a practical tool for Ayurvedic education, research, and clinical knowledge support.

The study underscores the value of domain-specialized, multilingual LLMs in bridging gaps between traditional knowledge systems and modern AI. Future work will focus on enhancing reasoning in high-difficulty scenarios, improving Hindi and Sanskrit understanding, and expanding coverage to underrepresented Ayurvedic subdomains, further advancing trustworthy, culturally aligned AI for traditional medicine.

7 Limitations

Despite strong performance on the BhashaBench-Ayur benchmark, AyurParam has several limitations that warrant acknowledgment:

Temporal Coverage and Contemporary Knowledge. The training corpus consists primarily of classical Ayurvedic texts and open-source educational materials digitized before 2024. As a result, AyurParam has limited exposure to recent research developments, contemporary clinical practices, and ongoing advances in Ayurveda. This temporal gap is common in domain-specialized models [4, 5] but particularly consequential in medical domains where evidence and practices evolve continuously.

Language Performance Gap. While AyurParam supports both English and Hindi, performance on Hindi queries lags behind English (38.04% vs. 41.12% accuracy, Table 1). This disparity suggests insufficient representation of Hindi content in the training corpus and highlights the ongoing challenge of building truly balanced multilingual models for specialized domains [6].

Evaluation Scope and Methodology. Our evaluation relies exclusively on structured exam-style questions from BhashaBench-Ayur. While this benchmark provides rigorous assessment of factual knowledge and reasoning, it does not capture other critical dimensions such as: (i) open-ended generation quality; (ii) clinical reasoning in realistic consultation scenarios; (iii) safety and appropriateness of generated advice; and (iv) practitioner acceptance and usability. Human evaluation by Ayurvedic experts would provide complementary insights but was beyond the scope of this work [5].

Safety and Ethical Guardrails. Although trained on factual, expert-curated content, AyurParam lacks explicit safety mechanisms to prevent generation of inappropriate, unsafe, or potentially harmful medical advice. Prior work in medical LLMs emphasizes the necessity of robust guardrails and clinical validation before deployment [6, 5]. Current responses are knowledge-grounded but not clinically validated.

Personalization and Context-Awareness. AyurParam does not account for individual patient histories, contraindications, or personalized health contexts in its responses. While this aligns with its design as an educational and reference tool, practical clinical utility would benefit from patient-specific reasoning capabilities.

Data Licensing and Diversity. The corpus is limited to open-access texts from public repositories such as Archive.org, eGangotri, and NDLI [21, 22, 23]. Incorporating licensed clinical databases, contemporary peer-reviewed literature, and modern practice guidelines would enhance both coverage and reliability but requires navigating complex licensing and copyright considerations.

8 Future Work

Several directions emerge naturally from these limitations:

Incorporating Contemporary Knowledge. Future iterations should integrate recent research publications, institutional clinical guidelines, and emerging practices to maintain temporal relevance. This may involve continual learning frameworks or periodic retraining cycles similar to approaches used in general medical LLMs [6].

Improving Multilingual Balance. Addressing the Hindi performance gap requires targeted data augmentation, improved tokenization strategies, and possibly language-specific fine-tuning. Techniques from multilingual NLP research [24, 25, 26] could be adapted to the Ayurvedic domain.

Human Evaluation and Clinical Validation. Systematic evaluation by Ayurvedic practitioners is essential to assess clinical utility, safety, and appropriateness of generated responses. This could follow frameworks established for medical AI systems [5] and include qualitative analysis of open-ended consultations.

Safety Mechanisms and Guardrails. Future versions should implement explicit safety layers to detect and prevent generation of harmful advice, incorporating lessons from instruction-tuning and alignment research [10, 9, 11]. This includes disclaimer generation, uncertainty quantification, and refusal mechanisms for out-of-scope queries.

Enhanced Data Sources. Expanding the corpus to include licensed clinical data, modern formulations, and validated treatment protocols would improve both accuracy and practical utility. Collaboration with Ayurvedic institutions and regulatory bodies could facilitate access to high-quality, authoritative sources.

Accessibility and User Adaptation. Developing mechanisms to adapt response complexity based on user expertise (practitioners vs. patients vs. students) would enhance practical utility across diverse use cases while maintaining scientific accuracy.

By addressing these limitations, future research can advance AyurParam toward a more comprehensive, safe, and clinically validated tool for Ayurvedic knowledge dissemination and education.

References

[1] BharatGenAI. (2025). *AyurParam: Domain-specialized Ayurvedic LLM*. Hugging Face. `https://huggingface.co/bharatgenai/AyurParam`

[2] BharatGen Research Team. (2025). *BhashaBench-Ayur (BBA): Pioneering India's Ayurvedic AI Benchmark*. Hugging Face Datasets. `https://huggingface.co/datasets/bharatgenai/bhashabench-ayur`

[3] BharatGen Research Team. (2025). *PARAM-1 BharatGen 2.9B Model*. arXiv preprint arXiv:2507.13390. `https://arxiv.org/abs/2507.13390`

[4] Ling, C., et al. (2023). Domain Specialization as the Key to Make Large Language Models Efficient, Reliable, and Explainable. *arXiv preprint arXiv:2305.18703*.

[5] Alaa, A., et al. (2025). Rethinking Medical Benchmarks for Large Language Models. *arXiv preprint arXiv:2508.04325*.

[6] Qiu, P., et al. (2024). Towards Building Multilingual Language Model for Medicine. *Nature Communications*. `https://doi.org/10.1038/s41467-024-52417-z`

[7] IBM Research. (2024). InstructLab: Large-scale Alignment for chatBots (LAB). *Red Hat Blog*. `https://www.redhat.com/en/topics/ai/what-is-instructlab`

[8] Wei, J., et al. (2021). Finetuned Language Models Are Zero-Shot Learners. *arXiv preprint arXiv:2109.01652*.

[9] Wang, Y., et al. (2022). Self-Instruct: Aligning Language Models with Self-Generated Instructions. *arXiv preprint arXiv:2212.10560*. `https://arxiv.org/abs/2212.10560`

[10] Ouyang, L., et al. (2022). Training Language Models to Follow Instructions with Human Feedback (InstructGPT). *NeurIPS*.

[11] Taori, R., et al. (2023). Stanford Alpaca: An Instruction-Following LLaMA Model. *Stanford CRFM Blog*.

[12] Databricks. (2023). Databricks Dolly: Democratizing the Magic of ChatGPT. *Databricks Blog*.

[13] ShareGPT contributors. (2023). ShareGPT Dataset. `https://sharegpt.com`

[14] Ding, N., et al. (2023). UltraChat: A Large-Scale Chat Dataset for Instruction Tuning. *arXiv preprint arXiv:2305.14233*.

[15] Padia, A., et al. (2025). AyurGPT: Fine-Tuning a Medical LLM for Multilingual Ayurveda Consultations. *NLP Summit Presentation*. `https://www.nlpsummit.org/ayur-gtp-fine-tuning-a-medical-llm-for-multilingual-ayurveda-consultations/`

[16] Rathor, C., et al. (2024). Artificial Intelligence in Ayurveda: A Simple Overview. *Journal of Ayurveda and Integrative Medical Sciences*. https://www.jaims.in/jaims/article/view/4040

[17] Charaka. (circa 100 AD). *Charaka Samhita*. Ancient Ayurvedic Compendium.

[18] Sushruta. (circa 200 AD). *Sushruta Samhita*. Ancient Ayurvedic Compendium.

[19] Vagbhata. (circa 600 AD). *Ashtanga Hridaya*. Classical Ayurveda text.

[20] Kashyapa. (circa 600 AD). *Kashyapa Samhita*. Pediatrics-focused Ayurvedic Compendium.

[21] Internet Archive. (2025). *Archive.org Digital Library*. https://archive.org

[22] eGangotri Foundation. (2025). *eGangotri Digital Library*. https://egangotri.in

[23] NDLI. (2025). *National Digital Library of India*. https://ndl.iitkgp.ac.in

[24] Kakwani, D., et al. (2020). IndicCorp and XL-Sum: Benchmarking Multilingual Corpora for Indian Languages. *Findings of ACL 2020*.

[25] Kunchukuttan, A. (2020). The IndicNLP Library: Natural Language Processing for Indian Languages. *EMNLP Workshop*.

[26] Xue, L., et al. (2021). mC4: A Multilingual C4 Corpus. *TACL*.

[27] Gao, L., et al. (2021). The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *NeurIPS Datasets and Benchmarks*.

[28] Laurençon, H., et al. (2022). The ROOTS Corpus: A 1.6TB Multilingual Dataset for Training Large Language Models. *ACL*.

[29] Soldaini, L., et al. (2023). Dolma: A Large-Scale, Long-Context, English Dataset. *arXiv preprint arXiv:2306.07328*.

[30] Penedo, G., et al. (2023). RefinedWeb: A Large-Scale Dataset for Web-Scale Language Models. *NeurIPS*.

[31] Raffel, C., et al. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *JMLR*.

[32] Zhu, Y., et al. (2015). BookCorpus: A Large Collection of Free Books. *ICCV Workshop*.

[33] Gokaslan, A., & Cohen, V. (2019). OpenWebText Corpus. *Open Source Release*.

[34] Mishra, A., & Sharma, R. (2023). Surya: Multilingual OCR for Indic Scripts. *Proceedings of the Indic NLP Symposium*.

[35] Gebru, T., et al. (2018). Datasheets for Datasets. *FAT* Conference*.

[36] Mitchell, M., et al. (2019). Model Cards for Model Reporting. *FAT* Conference*.

[37] Gebru, T., et al. (2021). Ethical Considerations in Dataset Collection. *FAccT Conference*.

[38] Creative Commons. (2010). Creative Commons Zero License (CC0). https://creativecommons.org/publicdomain/zero/1.0/

[39] Creative Commons. (2010). Creative Commons Attribution License (CC-BY). https://creativecommons.org/licenses/by/4.0/

[40] Alibaba DAMO Academy. (2024). Qwen3-235B: Large Language Model. *Technical Report*.

[41] Google DeepMind. (2024). Gemma: Open Language Models by Google DeepMind. *Technical Report*.