MenstLLaMA: A specialized large language model for menstrual health education in India

Prottay Kumar Adhikary, Isha Motiyani, Gayatri Oke, Maithili Joshi, Kanupriya Pathak, Salam Michael Singh, Tanmoy Chakraborty*

Indian Institute of Technology Delhi, India

*Corresponding author: tanchak@iitd.ac.in

Summary

Background

The quality and accessibility of menstrual health education in developing nations, including India, remain inadequate due to challenges such as poverty, social stigma, and gender inequality. While community-driven initiatives aim to raise awareness, artificial intelligence (AI) offers a scalable solution for disseminating accurate information. However, existing general-purpose large language models (LLMs) are ill-suited for this task, suffering from low accuracy, cultural insensitivity, and overly complex responses. To address these limitations, we developed *MenstLLaMA*, a specialized LLM tailored to the Indian context, designed to deliver menstrual health education empathetically, supportively, and accessible.

Methods

We curated a novel, domain-specific dataset and benchmarked state-of-the-art LLMs to develop *MenstLLaMA*, an empathic companion model. The evaluation employed an open-label benchmark design with a four-stage framework: (1) overlap with ground truth, (2) clinical relevance, (3) response diversity, and (4) user satisfaction. A panel of clinical experts (n=118) conducted expert evaluations, while participants (n=1,200) interacted with chatbots, including *MenstLLaMA*, in 15–20-minute randomized sessions for user satisfaction assessment.

Findings

MenstLLaMA was compared against state-of-the-art general-purpose LLMs such as GPT-4o, Claude-3, and Mistral during the evaluation period using automated and human-based metrics. *MenstLLaMA* achieved the highest BLEU score (0.059) and BERTScore (0.911), outperforming competitors without requiring few-shot learning. Clinical experts consistently rated its responses superior to gold-standard answers. User case studies revealed high ratings in Understandability (4.7/5) and Relevance (4.3/5), with a moderate rating in Context Sensitivity (3.9/5).

Interpretation

MenstLLaMA demonstrates exceptional accuracy, empathy, and user satisfaction in menstrual health education, bridging critical gaps left by general-purpose LLMs. Its potential for integration into broader health education platforms positions it as a transformative tool for menstrual well-being. Future research may explore its long-term impact on public perception and menstrual hygiene practices.

Keywords

Menstrual Health Education, Artificial Intelligence (AI), Large Language Models (LLMs), Cultural Sensitivity, Health Equity

Introduction

Menstruation, a biological process experienced by nearly half the global population, continues to be shrouded in stigma and misinformation, particularly in low-income and rural areas.[1] This accentuates the lack of awareness about menstrual hygiene, resulting in health risks and contributing to the spread of misinformation. In India, menstrual health remains a deeply ingrained taboo across both rural and urban settings.[2] Women are often segregated during menstruation, barred from entering places of worship, and labelled as “impure” or “unclean” during this natural cycle.[3,4] This societal ostracism and inadequate menstrual health education (MHE) have far-reaching consequences. For instance, nearly 24% of school-aged girls in India skip or drop out of school altogether due to menstruation-related issues,[5] reflecting the persistent discrimination they face from a young age.[2] These challenges underscore the urgent need for nationwide initiatives to make MHE accessible, inclusive, and culturally sensitive.

MHE can empower women and benefit society in multiple ways. It can impart essential information about maintaining menstrual hygiene and accessing supplies and sanitation facilities to prevent menstrual health disorders.[6] Simultaneously, it can address the social stigma surrounding menstruation and help dismantle cultural taboos that restrict women from fully participating in various spheres of life while they are menstruating.[6] While India’s education system has begun incorporating MHE into school curricula, these efforts often fall short, failing to address cultural sensitivities or challenge entrenched norms that perpetuate menstrual stigma.[7] Non-governmental organizations (NGOs) have stepped in to bridge this gap, conducting community-based workshops and awareness campaigns across the country.[7] However, the advent of Generative Artificial Intelligence (GenAI) and Large Language Models (LLMs) offers an unprecedented opportunity to scale these efforts and reshape public understanding of menstrual health.

Most of the Indian population, particularly the youth, has internet access and relies on it as a primary knowledge source. However, while the internet offers a vast repository of information, it is also prone to spreading misinformation and adding to the stigma associated with menstruation.[8] The introduction of LLMs offers new possibilities for disseminating accurate menstrual health information and combating misinformation.[9] While LLMs hold promise as tools

for disseminating accurate health information, existing general-purpose models often fall short in addressing this specialized domain. They struggle with low accuracy, lack of cultural sensitivity, and the tendency to generate verbose or overly complex responses that fail to resonate with diverse user groups.[10]

To tackle these challenges, we introduce MenstLLaMA, a specialized LLM tailored to the menstrual health domain. Built upon the LLaMA-3-8B-Instruct model[11] and fine-tuned using our custom MENST dataset comprising 24,167 question-answer pairs, MenstLLaMA bridges critical gaps in existing LLMs. Our comprehensive evaluation demonstrated its superior performance compared to leading models, including GPT-4o[12] and Claude-3,[13] across both automated metrics and clinical expert assessments. To evaluate MenstLLaMA by medical practitioners and to conduct a user study, we further developed ISHA (Intelligent System for MenstrualHealth Assistance), an interactive chatbot powered by MenstLLaMA. This study highlights the potential of *MenstLLaMA* as an empathetic, culturally aware AI companion for advancing menstrual health education in India.

Research in context

Evidence before this study

LLMs have shown promise in various healthcare domains, but their application to menstrual health education has been limited. Before initiating this study, we conducted a systematic review of AI applications in menstrual health education, searching PubMed, Google Scholar, and arXiv for articles published between 2020 and 2024 using the keywords “artificial intelligence,” “language models,” “menstrual health,” and “health education.” While general-purpose LLMs like GPT-4 and Claude-3 have proven effective in medical knowledge tasks, no specialized models exist for menstrual health education. Previous efforts leveraging AI in menstrual health have primarily focused on period-tracking applications and symptom prediction, with limited emphasis on comprehensive menstrual health education or counseling supported by specialized LLMs.

Added value of this study

This study provides three key contributions to the field. First, we developed the MENST dataset, a comprehensive collection of 24,167 question-answer pairs specifically curated for menstrual health education. Second, we introduced MenstLLaMA, the first open-source large language model specifically fine-tuned for menstrual health education. Third, we demonstrated through rigorous evaluation that a domain-specific model can outperform general-purpose LLMs in providing accurate, culturally sensitive, and empathetic menstrual health information. Our evaluation methodology, combining automatic metrics with clinical expert assessment and user studies, provides a framework for evaluating AI systems in sensitive healthcare domains. The strong performance of MenstLLaMA, particularly in user comprehension and empathetic communication, suggests that domain-specific fine-tuning can effectively address healthcare challenges.

Implications of all the available evidence

This study demonstrates the potential of specialized language models to transform menstrual health awareness and education through accurate, culturally sensitive, and accessible information delivery.

Methods

Study design and setting

This study followed a structured approach, encompassing three main stages: dataset creation, model development, and evaluation. The first stage involved the creation of the MENST dataset, a comprehensive collection of 24,167 question-answer pairs covering various facets of menstrual health. This dataset formed the foundation for developing and evaluating MenstLLaMA, a novel generative AI model built upon the LLaMA-3 architecture.[11]

In the model development phase, we fine-tuned the LLaMA-3 model using the MENST dataset to improve its capacity to generate accurate, culturally sensitive, and empathetic responses for menstrual health education. To assess MenstLLaMA's performance, we conducted a comparative analysis involving other state-of-the-art GenAI models (e.g., GPT-4o and Claude-3) and evaluated its effectiveness against expert clinical assessments.

Model fine-tuning and analysis were performed using Python-based machine learning frameworks, including PyTorch and Hugging Face's Transformers library.[14] The fine-tuning process leveraged high-performance computing resources to enhance model performance and efficiency. We employed the SciPy[15] and NLTK[16] libraries for comprehensive statistical analysis, calculating various evaluation metrics to ensure both reproducibility and robustness in our assessments. In addition to these computational evaluations, human-based assessments were conducted by clinical experts and end-users. These assessments were focused on evaluating the model's relevance, understandability, and cultural sensitivity, particularly in the context of menstrual health education. This dual approach ensured a thorough and well-rounded evaluation process.

Dataset creation

Dataset sources

The MENST dataset was compiled from various reputable sources, including health information portals, medical institutions, government websites, global organizations, and educational platforms. Most of the dataset was sourced from official medical documents, where we extracted and curated relevant FAQs. Additionally, we incorporated an existing question-answer dataset on menstrual health, specifically the Menstrual Health Awareness Dataset.[17] This dataset, consisting of 562 question-answer pairs, and all the collected QA pairs were manually annotated with metadata (see *Metadata creation* for the detailed metadata description).

To enrich the dataset further, we employed prompting techniques with advanced language models like GPT-4 and Gemini 1.5 Pro, generating additional question-answer pairs from relevant menstrual health documents. This process helped to create a robust and comprehensive dataset tailored to menstrual health education, ensuring accuracy, cultural relevance, and empathy in the content. In the following subsections, we provide a detailed breakdown of the dataset's structure, metadata creation, and the paraphrasing strategies employed to augment its coverage.

Metadata creation

We created comprehensive metadata for all the documents to facilitate efficient data management and detailed catalogues of menstrual health topics across various demographics and contexts. We tagged each document with a unique document ID starting with 'D' or 'F,' where unstructured documents or paragraphs were prefixed with 'D', which stands for documents. In contrast, 'F' or Frequently Asked Questions (FAQs) was prefixed to the already structured or question-answer pair documents. In addition, we included the source and link of each document, along with the topics covered by the document, as part of the metadata.

The dataset must be structured as question-answer pairs to fine-tune an LLM for tasks such as question-answering or conversational interactions. To achieve this, we processed unstructured documents using LLMs to generate corresponding question-answer pairs. The preprocessing methodology for transforming unstructured documents into question-answer format is detailed in the following section. These LLM-generated question-answer pairs were combined with pre-existing pairs to construct the final MENST dataset. Table 1 outlines the metadata elements for the original collected documents, while Table 2 provides the metadata elements for the final processed question-answer dataset.

Table 1: Metadata schema for source documents in the MENST dataset.

Metadata Labels	**Description**
Document ID	Each document is assigned a unique ID, starting with 'D' or 'F'. This naming convention helps distinguish documents with only paragraph texts and question answers.
Document Name	Name of the document, such as title or main heading.
Source	Source of the document containing the name of the website or organisation.
Link	URL taking to the document
Topic	The topic on which the document is based, for example, is irregular menstruation.

Table 2: Metadata schema for question-answer pairs in the MENST dataset.

Metadata Labels	**Description**
Document ID	ID of the document from which the question has been created or taken.
Question	Questions related to the topic.
Answer	Answer to the question.
Age group	Age group to which the question is related. Different age groups considered are adolescents - those below 19, young adults - those between 20 and 40, and older adults - those above 40.
Region	The region the question is specific about. It could be rural, urban or both.
Topic	The topic on which the question is based.
Subtopic	This specifies the exact subdomain of the question under the topic.
Keywords	Tags contained in the query (i.e. Medication, Therapy, Surgery, Doctor)
LLM Used	Name of the LLM used for post-processing the question.

Collaborating with expert gynaecologists, we created a taxonomy for the topics covered in the menstrual health documents. Primary categories include *Anatomy*, *Normal Menstruation*, *Abnormal Menstruation*, *Pregnancy*, *Lifestyle*, *Support*, and *Society*. Each primary category was further subdivided into specific subtopics. For instance, *Normal Menstruation* includes subtopics like *Menarche*, *Menopause*, *Normal Flow*, and *Normal Cycle*, while *Abnormal Menstruation* covers issues such as *Abnormal Bleeding*, *Irregular Periods*, *Menstrual Pain*, *PCOS*, and *PMS*. A detailed breakdown of the dataset's taxonomy is provided in Figure 1.

Question-answer pair creation

Our primary dataset included 117 documents, of which 14 were question-answer documents (FAQ). These FAQ documents, sourced from official medical portals, were used as a gold test set (Set-1). For the remaining 103 unstructured documents, we used GPT-4 and Gemini 1.5 Pro to generate question-answer (QA) pairs. These generated pairs were then validated by domain experts.

We employed a standardized prompt template to ensure the consistency of the QA generation process. The prompt template used three random sample QA pairs from Set-1 to provide context for the language models, as shown in Figure 2.

Figure 1: Taxonomy of menstrual health topics in the MENST dataset. The taxonomy consists of seven primary categories (Anatomy, Normal Menstruation, Abnormal Menstruation, Pregnancy, Lifestyle, Support, and Society) with corresponding subtopics (Menarche, Menopause, Normal Flow, Normal Cycle, Menstruation Bleeding, Abnormal Bleeding, Irregular Periods, Menstrual Pain, Polycystic ovary syndrome [PCOS], Premenstrual syndrome [PMS], Diet, Attitude, Hygiene, Lack of resources, Government Scheme, Medical Diagnosi, Myths, Taboo)

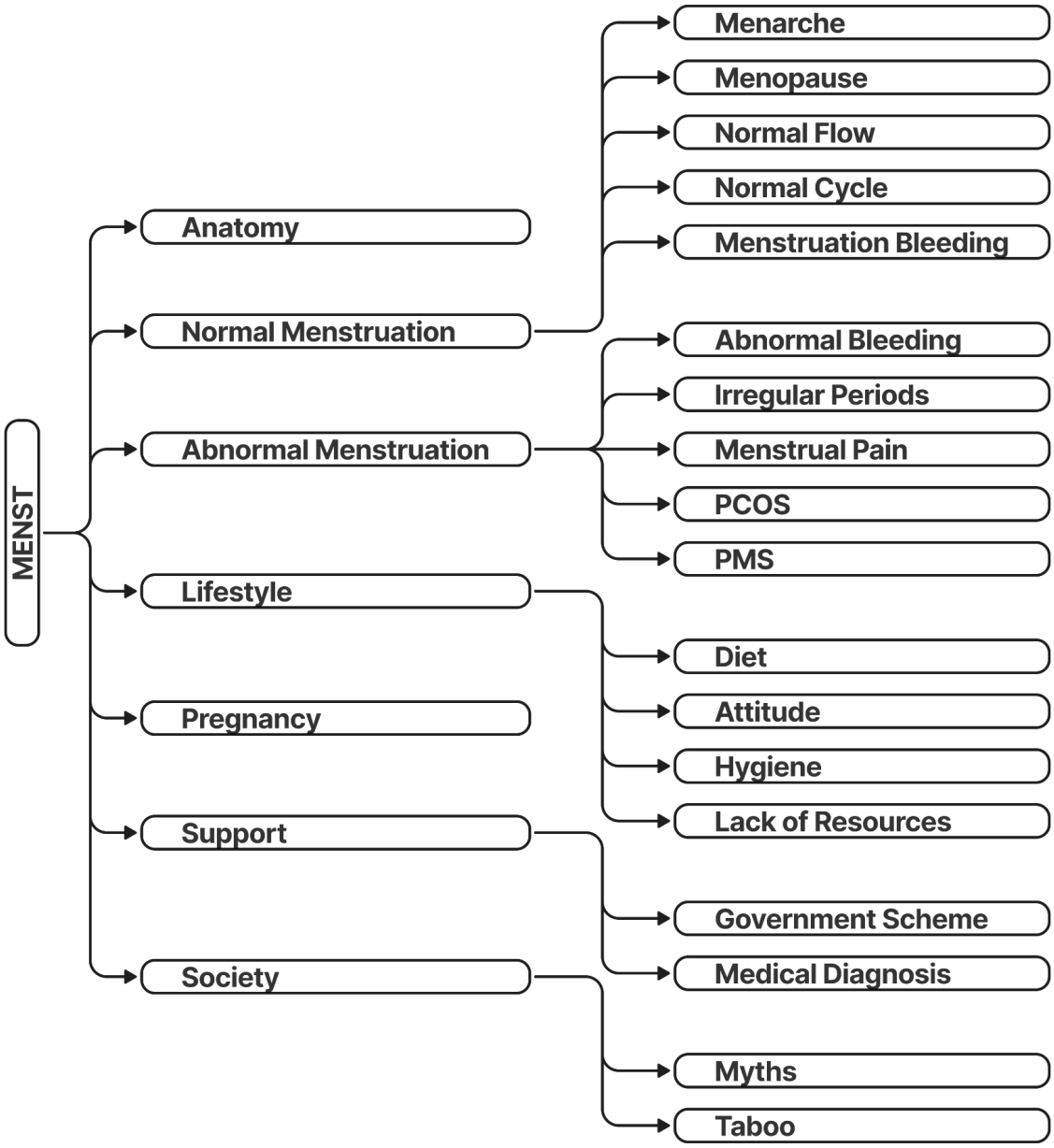


Figure 2: Prompt template structure for curating question-answer pairs from unstructured medical documents. The template includes four components – (1) a task description specifying the generation of relevant questions from document content, (2) the unstructured document from which the QA pair is to be generated, (3) instructions for creating concise, well-structured questions focusing on critical aspects, and (4) example format showing three QA pairs from gold standard data to guide the generation process.

Task: Generate Question-Answer Pairs from a Given Document

Document: {document_text}

Instructions: Carefully read the provided Document and create questions that directly relate to the content, focusing on important details, definitions, implications, and other critical aspects. Ensure the questions are concise and the answers are well-structured to align perfectly with the questions.

Question-Answer Pair Examples:
Q: {q1}
A: {a1}

Q: {q2}
A: {a2}

Q: {q3}
A: {a3}

Use the following format and return each possible question-answer pair from the given document:
Q: {question}
A: {answer}

Paraphrasing for dataset augmentation

To enhance the diversity of the dataset, we employed paraphrasing techniques. This approach improves the model's handling of varied user expressions and contexts. By prompting GPT-3.5 with instructions shown in Figure 3, we implemented four distinct paraphrasing strategies:

1. **Scenario-based paraphrasing** helps questions include real-life situations, making the model's learning data more responsive to human input. For example, "Can I cook while I am on my period?" could be paraphrased as "Today while I was in the kitchen, my mother scolded me as I entered the kitchen on my period. Is it a sin to cook during this phase?" This helps the model better understand the queried question and respond in ways more aligned with human conversational patterns.

2. Rephrasing from a **male point of view** gives a male's perspective on the question, often lacking direct experience with menstruation. This perspective is essential as it expands the model's understanding and ability to engage with questions from individuals indirectly affected by the issue. For example, "I am experiencing painful periods. Is this normal?" becomes "My wife is experiencing painful periods. Is this normal?". This ensures that the model can adequately address queries from a broader audience.

3. **Changing the sentence structure** introduces linguistic diversity in the question, which is crucial for developing syntactic flexibility in a model. For example, "Is it normal to start menstruating at the age of 12?" can be rephrased as "I am 12 years old and just got my

first period. Is this okay?". This diversity in data helps the model learn the flexibility needed to understand personalized questions.

4. Paraphrasing from the **rural women's perspective** captures the language and concerns of women in rural settings, who may have different cultural and educational backgrounds. For example, "I don't have access to sanitary pads. What can I use instead?" is paraphrased as "In my village, there are no sanitary pads. What can I use instead?" This helps the model respond appropriately to a diverse demographic, especially village students.

Figure 3: Example of the prompt template used for dataset augmentation using paraphrasing the questions. The template demonstrates the male perspective paraphrasing strategy, showing (1) the input instruction that specifies the paraphrasing task from a male's perspective, (2) a sample input question about menstrual pain, and (3) three generated paraphrased versions that maintain clinical accuracy while incorporating different male familial relationships (wife, teenage daughter, growing woman). This template generated diverse yet contextually relevant question variations for the MENST dataset, enhancing the MenstLLaMA's ability to handle queries from different perspectives and social contexts.

Instruction: Suppose you are a male and lack proper knowledge of menstruation. Generate 3 paraphrased versions of the given input question from a man's perspective on menstruation separated with #.

Input Question: I am experiencing painful periods. Is this normal?

Output Answers: My wife is experiencing painful periods. Is this normal? # My teenage daughter is having painful periods. Is it expected? # Is a painful period for a growing woman normal?

Each question from the raw dataset was paraphrased across four distinct scenarios, resulting in 24,167 question-answer pairs. A breakdown of the dataset's structure is provided in Table 3.

Table 3: Distribution of QA pairs in the MENST dataset across different data sources before and after paraphrasing augmentation

Set No.	Source	RawQA	ParaphrasedQA
Set-1	FAQs	117	1521
Set-2	Text Documents on Menstruation	1180	15340
Set-3	Menstrual Health Awareness Dataset	562	7306
	Total	1859	24167

Expert-compiled test dataset

In collaboration with medical professionals, we developed an expert-curated test dataset to evaluate the model. We asked experts to compile questions they commonly encounter in clinical practice and provide appropriate answers. Using our metadata framework (7 topics, 18

subtopics), we obtained 200 QA pairs reflecting real-life clinical scenarios. This expert-curated set was used as a gold test set for evaluating the model's performance.

Model development

We developed MenstLLaMA, a specialized language model for menstrual health, by fine-tuning the Meta-LLaMA-3-8B-Instruct model[11] on our curated MENST dataset, which comprises 24,167 question-answer pairs. Fine-tuning was conducted using a parameter-efficient fine-tuning (PEFT) strategy,[18] with a specific focus on Low-Rank Adaptation (LoRA)[19] to align the model with the MENST dataset. PEFT methods optimize the fine-tuning of LLMs by adjusting a smaller subset of parameters while keeping most of the model's weights frozen. LoRA introduces a low-rank matrix into the attention mechanism, enabling the model to effectively learn task-specific patterns with reduced computational overhead. This approach allows for efficient adaptation across diverse conversational contexts while maintaining performance comparable to full fine-tuning.

Instruction fine-tuning was performed on the LLaMA-3-8B-Instruct model using a structured prompt format to distinguish between instructions and responses. Each QA pair was reformatted, as shown in Figure 4. This format ensured that the model could clearly distinguish between instructions and responses. This improved its ability to handle complex context-sensitive conversations by enabling the model to effectively capture the nuances of dialogue.

Figure 4: Instruction fine-tuning format for MenstLLaMA, illustrating the standardized structure for QA pairs. The format demonstrates the conversion of the QA pairs into LLaMA instruction syntax (**<s>[INST]** *question* **[/INST]** *answer* **<s>**). This example illustrates a dietary question during menstruation, with the response structured to provide clear, scientific information about nutritional needs. The format enables the model to distinguish between the question and their corresponding responses.

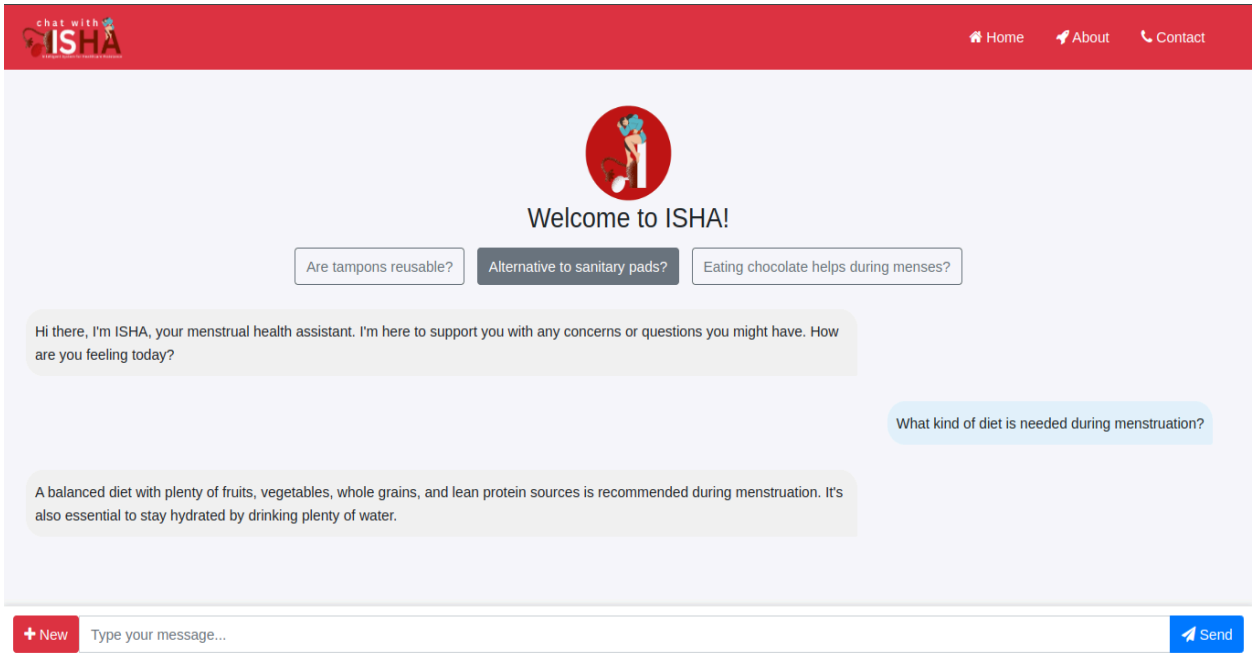
<s>[INST] What kind of diet is needed during menstruation? **[/INST]** During menstruation, it is better to maintain a well-balanced diet with iron-rich foods and green leafy vegetables since they help combat fatigue and replenish lost blood. Eating bananas helps as they are high in potassium and can help regulate blood pressure and reduce bloating.**</s>**

We used Meta-Llama-3-8B-instruct[20], an 8 Billion parameter model, as the base model for our model configuration. We finetuned the model with a maximum sequence length of 2048 tokens and leveraged LoRA for efficient training. We applied 4-bit quantization using the NF4 type to optimise memory usage. We used Paged AdamW as the optimizer with 32-bit precision for efficient gradient updates. The training hyperparameters were fine-tuned to balance learning rate, batch size, and gradient accumulation steps, allowing us to achieve optimal performance with minimal computational overhead. The training process involved five epochs with a learning rate of 2e-4, a warmup ratio of 0.03 and a maximum gradient norm of 0.3. We utilised the NVIDIA A100 GPU with 80GB of memory for the finetuning process.

Development of an interactive chatbot

To enable qualitative evaluations by medical practitioners and facilitate case studies, we developed an interactive chatbot powered by the fine-tuned MenstLLaMA model, ISHA (**I**ntelligent **S**ystem for Menstrual **H**ealth **A**ssistance). ISHA was designed to ensure a seamless conversational flow by leveraging contextual information from previous utterances within the same session. During the fine-tuning and inference stages, we incorporated tailored instructions into the model's prompts to enhance the chatbot's empathy and emotional awareness. This approach aimed to create a contextually accurate chatbot that is emotionally attuned to the nuances of the menstrual health domain. Figure 5 presents the ISHA interface.

Figure 5: Interface of the ISHA chatbot powered by MenstLLaMA.



Baseline models

We compared MenstLLaMA against nine state-of-the-art LLMs including three closed-source models (GPT-4,[12] Gemini 1.5 Pro,[21] Claude-3[13]) and six open-source models (Mistral,[22] LLaMA-3,[11] GPT-2,[23] Orca2,[24] Falcon,[25] Phi3[26]).

Evaluation

The evaluation of MenstLLaMA and other state-of-the-art closed and open-sourced models followed a comprehensive four-step approach to assess the model's performance across technical accuracy, clinical relevance, and user satisfaction. These steps included **automatic evaluation** using standard NLP metrics, **clinical expert evaluation**, **medical practitioner feedback** in a simulated setting, and a **user-centred evaluation**. This multi-faceted approach

provided a balanced understanding of the model's strengths and limitations, particularly in the sensitive domain of menstrual healthcare.

Automatic evaluation

MenstLLaMA was benchmarked against leading conversational models, including proprietary systems (GPT-4o, Gemini 1.5 Pro, Claude-3) and open-source models (Mistral, LLaMA-3, GPT-2, Orca2, Falcon, Phi3). Given the generative nature of the task, we employed standard Natural Language Processing (NLP) evaluation metrics such as BLEU,[27] METEOR,[28] ROUGE-L,[29] and BERTScore[30] to assess the models against the expert-compiled gold test set. These metrics assessed fluency, relevance, overlap, and semantic similarity with reference responses. More details on the metrics are presented in Table 4. Following this evaluation, the top four models were selected for further analysis. Detailed results from this comparative analysis are presented in the Results section.

Evaluation by clinical experts

To assess the clinical relevance of our model against the top-performing baseline models based on the automatic evaluation, we engaged six experienced gynaecological experts experienced in patient interactions. The panel comprised three Bachelor of Medicine trainees (n=3), two physician assistants (n=2), and one sexual health therapist (n=1). These professionals were presented with the 200 questions from the expert-compiled test set, each accompanied by five unlabeled responses: the top four models and one gold-standard response curated by domain experts. The experts selected the most appropriate responses from a set of five, where four responses were generated by the top four models, and the remaining response was sourced from the expert-compiled test set. To mitigate potential bias, we did not disclose the source of the responses.

Evaluation by medical practitioners

Following the evaluations that verified the MenstLLaMA's performance against the baseline models, we extended the qualitative evaluation by assigning the gynaecological practitioners to rate MenstLLaMA based on some metrics. For this purpose, we used ISHA, our chatbot interface for MenstLLaMA. We instructed all 12 practitioners to interact with ISHA, act like patients, and then evaluate the model's response based on qualitative metrics. Among the practitioners, five were experienced doctors (n=5), and seven were postgraduate trainees in medicine (n=7), all handling patients regularly in their clinical practice. We compiled these metrics to assess various aspects of the model's performance in a clinical context. Each metric was rated on a five-point Likert scale, where 1 signifies the lowest score and 5 the highest. More details on these metrics are reported in Table 4. In this way, we evaluated MenstLLaMA with the experts' involvement in a simulated clinical setting, providing insights beyond quantitative measures. The results of this evaluation are discussed in the Results section.

Table 4: Evaluation metrics, definition and scoring criteria for MenstLLaMA performance. Notes: automatic evaluation metrics (1-4) where all the metrics' scores range from 0 to 1, with 0 being the least and 1 being the maximum, medical practitioners' assessment criteria (5-9), and user evaluation criteria (5-11).

Metric Type		**Metric**	**Definition**
Automatic evaluation		1) BLEU	Measures the precision of word overlap between the model's response and gold reference answers, providing insight into response accuracy.
		2) METEOR	Evaluates semantic similarity by considering word stems, synonyms, and paraphrases, offering a more flexible assessment of meaning preservation.
		3) ROUGE-L	Measures the longest co-occurring subsequences between the generated response and the gold answer.
		4) BERTScore	Leverages contextual embeddings to gauge the semantic coherence between the gold and the generated response.
User evaluation	Medical practitioner evaluation	5) Relevance	How closely the responses aligned with the information users sought, with ratings ranging from 1 (least relevant) to 5 (most relevant).
		6) Understandability	Measures the clarity and ease with which the responses could be comprehended, ensuring that the information was presented in a way that was straightforward and accessible, rated from 1 (not at all understandable) to 5 (completely understandable).
		7) Preciseness	The extent to which responses were concise and directly addressed the query, avoiding unnecessary details or digressions. Rated from 1 (responses were random) to 5 (responses were precise).
		8) Correctness	The factual accuracy of the information provided. Rated from 1 (not at all correct) to 5 (absolutely correct).

		9) Context Sensitivity	The model's ability to recognize and respond appropriately to the emotional tone of the interaction. Rated from 1 (robotic) to 5 (human-like).
		10) Tone	Assessed whether the responses were delivered in a polite and empathetic manner. Rated from 1 (not at all empathetic) to 5 (empathetic).
		11) Flow	States that the flow of the conversation was smooth or hard to relate to. Rated from 1 (hard to relate) to 5 (smooth flow).

User evaluation

For a more comprehensive evaluation of the MenstLLaMA, a user study was conducted with 200 participants, comprising 181 females (n=181) and 19 males (n=19), representing a diverse range of socio-demographic and professional backgrounds. The cohort included engineering and medical students, researchers, practising doctors, allied health professionals, and industry professionals from sectors such as technology, finance, and consulting. Conducted primarily across India, Bangladesh, Nepal and Bhutan, the study facilitated a comprehensive evaluation by integrating perspectives from varied domains and disciplines. The participants were instructed to interact with MenstLLaMA through the ISHA chatbot interface and provide user feedback. The participants were asked to evaluate the model's response based on the same metrics (see Table 4 for the full set of metrics) as those used in the above evaluation by medical practitioners in simulated evaluation, with two additional metrics: *tone*, and *flow*. These additional metrics were particularly significant in menstrual and gynaecological health, where sensitivity and compassion are paramount. The results of this user study are presented in the Results section.

Results

We evaluated the MenstLLaMA against the baseline LLMs using both automatic and human-based approaches.

Table 5: Automatic scores of the models using the metrics: BLEU, METEOR, ROUGE-L, and BERTScore. We employed zero-shot and few-shot (one and two shots) prompting techniques for the baseline models. The scores are computed with respect to the expert-compiled test set.

Technique	Models	BERTScore	BLEU	METEOR	ROUGE-L
Zero-Shot	GPT4o	0.896	0.052	0.256	**0.253**

	Gemini1.5-Pro	0.887	0.034	0.210	0.215
	Claude-3	0.888	0.054	**0.321**	0.227
	LLaMA 3	0.881	0.046	0.293	0.201
	GPT 2	0.874	0.045	0.219	0.219
	Phi 3	0.885	0.047	0.281	0.207
	Orca 2	0.882	0.040	0.259	0.197
	Mistral	0.888	0.043	0.301	0.201
	Falcon	0.865	0.021	0.213	0.143
One-Shot	GPT4o	0.891	0.056	0.256	0.242
	Gemini1.5-Pro	0.879	0.043	0.223	0.221
	Claude-3	0.904	0.040	0.305	0.230
	LLaMA 3	0.881	0.044	0.298	0.202
	GPT 2	0.888	0.035	0.237	0.211
	Phi 3	0.898	0.046	0.273	0.205
	Orca 2	0.879	0.047	0.265	0.183
	Mistral	0.905	0.051	0.292	0.182
	Falcon	0.857	0.031	0.197	0.159
Two-shot	GPT4o	0.892	0.056	0.260	0.246
	Gemini1.5-Pro	0.878	0.045	0.224	0.220
	Claude-3	0.900	0.045	0.305	0.226
	LLaMA 3	0.879	0.044	0.302	0.200
	GPT 2	0.886	0.032	0.236	0.215
	Phi 3	0.901	0.051	0.276	0.204
	Orca 2	0.875	0.042	0.265	0.179
	Mistral	0.907	0.054	0.296	0.180
	Falcon	0.855	0.032	0.195	0.160
Proposed Model	MenstLLaMA	**0.911**	**0.059**	0.290	0.224

Table 5 reports the automatic scores for the models computed using BLEU, METEOR, ROUGE-L, and BERTScore. In our experimentation, we conducted various prompting strategies for the baseline LLMs viz, zero-shot (no example given in the prompt), one-shot (single example) and two-shot (two examples). MenstLLaMA achieved the highest scores in the BLEU (0.059) and BERTScore (0.911) metrics. While Claude-3 and GPT-4o achieved the highest for

the METEOR (0.321) and ROUGE-L (0.253) metrics. Interestingly, Gemini1.5-Pro and LLaMA 3 demonstrate reliable but slightly lower scores in comparison, reflecting their capabilities in specific tasks. The newer models like Mistral and Orca 2 provide strong BLEU and METEOR scores but occasionally underperform in ROUGE-L compared to others. Falcon, while showing potential, falls behind across most metrics. In the zero-shot scenario, the closed-source models (GPT-4o, Claude-3 and Gemini1.5) outperformed all the open-source models. Introducing a single example in the one-shot scenario improved most models. Notably, Mistral showed a significant jump in performance, achieving the highest BERTScore (0.905), surpassing even the closed-source models in this one-shot setting. A similar trend is observed in the two-shot setting, where the Mistral showed a gain over the zero-shot and two-shot settings and even outperformed the closed-source models in the two-shot scenario. Overall, the closed-source models consistently performed well across all scenarios. They showed strong performance even in zero-shot settings, indicating robust generalization capabilities. However, with a few shot prompting, some open-source models, such as Mistral and Phi 3, showed remarkable improvements. Mistral, in particular, achieved competitive performance with closed-source models in one-shot and two-shot settings. While most models showed improvement with additional examples, the degree of improvement was not uniform. Open-source models generally benefited more from few-shot examples compared to closed-source models. Our proposed model, MenstLLaMA, demonstrated strong performance across all metrics without requiring a few-shot examples. This indicates that fine-tuning the MENST dataset has resulted in a model well-adapted to the menstrual health domain.

Based on these automatic scores, we selected the top four performing models for further assessment: MenstLLaMA, GPT-4o, Claude-3-opus, and Mistral. These models underwent another round of evaluation by the clinical experts on the expert-compiled gold test. To ensure an unbiased assessment, we engaged experts who were not involved in curating the gold test set. MenstLLaMA consistently outperformed other models, including the gold-standard responses. This indicates that our model can generate contextually appropriate and medically reliable responses. The fact that MenstLLaMA's outputs were often preferred over the gold-standard responses shows the model's effectiveness in interpreting and applying domain-specific knowledge.

Subsequently, we also carried out medical practitioners' evaluations in a simulated clinical setting using ISHA, MenstLLaMA's chatbot interface. In this assessment, MenstLLaMA scored 3.5 in *Relevance*, indicating that while the responses were generally pertinent, there were gaps in addressing specific queries fully. *Understandability* was rated at 3.6, reflecting that most responses were clear and easy to comprehend, though there is room for improvement in conveying complex information more effectively. *Preciseness* scored 3.1, suggesting that although the chatbot provides relevant information, there is some inconsistency in maintaining concise and directly relevant answers. *Correctness* was rated at 3.5, underscoring that while the chatbot delivered accurate responses for common issues, additional updates are needed, particularly regarding recent guidelines and less frequently addressed topics. Notably, the model excelled in *Context Sensitivity* with a score of 4, demonstrating a strong ability to respond empathetically and understand the emotional nuances in interactions.

Finally, in the user case study, the users rated the model across seven key metrics on a 5-point scale. The users rated relatively high in *Relevance* (4.3), indicating that users generally found the responses pertinent to their queries. *Understandability* received the highest score (4.7), reflecting the clarity and ease with which users could comprehend the information provided. *Preciseness* was rated 4.28, showing that while responses were generally concise and to the point, there remains a need to fine-tune the output to avoid verbosity and redundancy. *Correctness* was rated 4.1, affirming that users found the chatbot's information mostly accurate, though some responses could benefit from more specific and detailed updates. The evaluation also highlighted positive scores in *Flow* (4.2) and *Tone* (4.6), with users appreciating the polite and empathetic nature of the interaction. However, *Context Sensitivity* scored lower at 3.9, suggesting room for improvement in how well ISHA adapts to and understands subtle emotional cues in conversation.

Discussion

In this study, we introduced the MENST dataset and MenstLLaMA. The MENST dataset was used to train the MenstLLaMA, a novel open-source large language model designed explicitly for the menstrual health domain. Our comprehensive evaluation highlighted MenstLLaMA's effectiveness in delivering accurate, accessible, and culturally sensitive menstrual health information.

Our domain-specific MenstLLaMA demonstrated competitive performance compared to state-of-the-art general-purpose LLMs, including open-source and closed-source LLMs. Notably, it achieved superior results in automatic evaluation metrics, such as BLEU (0.059) and BERTScore (0.911), without requiring few-shot examples. This underscores the potential of domain-specific fine-tuning to generate fluent, adequate and contextually relevant responses. While closed-source models like GPT-4o and Claude-3 showed robust performance in zero-shot settings, MenstLLaMA's consistent performance indicates successful adaptation to the menstrual health domain.

Clinical expert evaluations showed encouraging results, with MenstLLaMA's responses often preferred over gold-standard answers. This preference suggests that the model captures domain knowledge effectively and aligns with clinical expertise. The model's ability to generate responses that the experts rated more appropriate than the curated answers highlights its potential as a reliable source of menstrual health information. The simulated clinical setting evaluation using the ISHA chatbot revealed both strengths and areas for improvement. The medical practitioners interacted with ISHA, acted like patients, and evaluated the model's response based on qualitative metrics. While the model demonstrated strong empathetic capabilities (Context Sensitivity: 4.0), its performance in preciseness (3.1) suggests room for improvement in generating concise, focused responses. These findings indicate that while domain-specific training successfully improved the model's emotional intelligence, further work may be needed to enhance its ability to provide more precise information.

User evaluation results were particularly promising for real-world applications. High scores in understandability (4.7) and relevance (4.3) suggest that MenstLLaMA effectively bridges the gap

between clinical accuracy and user comprehension. The strong performance in tone (4.6) and flow (4.2) demonstrates the model's ability to maintain engaging, empathetic conversations while generating accurate and relevant health information.

Our study has several limitations. First, while the MENST dataset covers a broad range of topics, it does not fully represent all cultural perspectives and demographics. Second, although our evaluation included diverse user groups, longitudinal studies would be valuable to assess the model's impact on the overall outcomes and behaviour changes. Third, the current implementation focuses primarily on text-based interactions, potentially limiting accessibility for users with different communication needs. Looking forward, several directions for future research emerge. First, expanding the dataset to include more diverse demographics, age groups, and cultural contexts which could enhance the model. Second, investigating methods to improve the model's context sensitivity while maintaining its strong performance in other areas could lead to more personalized interactions. Third, exploring the integration of MenstLLaMA with other healthcare interventions could help establish its role in comprehensive menstrual health education programs.

In conclusion, MenstLLaMA represents a significant advancement in applying artificial intelligence to menstrual health education. Its performance demonstrates that domain-specific language models can effectively bridge the gap between clinical accuracy and accessible health communication. As we continue to refine and expand this approach, such models could play an increasingly important role in improving access to quality menstrual health information and reducing associated stigma globally.

Contributors

The contributions of individual authors are as follows: Conceptualisation: TC; Data curation: IM, PKA, MJ, KP, GO; Formal analysis: PKA, IM, SMS; Funding acquisition: TC; Investigation and Methodology: PKA, IM, SMS, TC; Project administration: TC, PKA, SMS; Resources, software, validation, visualisation: PKA, SMS; Writing – review & editing: TC, PKA, SMS, GO.

Data sharing

The data supporting the findings of this study are publicly available at Hugging Face. The dataset can be accessed at https://huggingface.co/datasets/proadhikary/MENST. Additionally, the pre-trained models and code used in this study are hosted on Hugging Face Model Hub and can be accessed at https://huggingface.co/proadhikary/Menstrual-LLaMA-8B. For further details or to report any issues, please contact the corresponding author.

Acknowledgements

TC would like to thank the financial support of the Tower Research Capital Markets toward using machine learning for social good, Rajiv Khemani Young Faculty Chair Professorship in AI, and the equipment support of central HPC facility (Padum).

Conflict of interest

The authors have no conflict of interest.

References

1 Mohd Tohit NF, Haque M. Forbidden Conversations: A Comprehensive Exploration of Taboos in Sexual and Reproductive Health. *Cureus* 2024; **16**: e66723.
2 Tuli A, Dalvi S, Kumar N, Singh P. “It’s a girl thing”: Examining Challenges and Opportunities around Menstrual Health Education in India. *ACM Trans Comput-Hum Interact* 2019; **26**: 29:1-29:24.
3 Bhartiya A. Menstruation, Religion and Society. *Int J Soc Sci Humanity* 2013; : 523–7.
4 Tan DA, Haththotuwa R, Fraser IS. Cultural aspects and mythologies surrounding menstruation and abnormal uterine bleeding. *Best Pract Res Clin Obstet Gynaecol* 2017; **40**: 121–33.
5 van Eijk AM, Sivakami M, Thakkar MB, *et al.* Menstrual hygiene management among adolescent girls in India: a systematic review and meta-analysis. *BMJ Open* 2016; **6**: e010290.
6 Babbar K, Martin J, Ruiz J, Parray AA, Sommer M. Menstrual health is a public health and human rights issue. *Lancet Public Health* 2022; **7**: e10–1.
7 Tuli A, Chopra S, Kumar N, Singh P. Learning from and with Menstrupedia: Towards Menstrual Health Education in India. *Proc ACM Hum-Comput Interact* 2018; **2**: 174:1-174:20.
8 Ghosh M, Ghosh I. ICT and information strategies for a knowledge economy: the Indian experience. *Program* 2009; **43**: 187–201.
9 Vengala K, Cherukuri NK, Annavarapu P, Dodla P. Artificial Intelligence in Menstrual Health: A Chatbot Approach to Personalized and Stigma-Free Conversations. In: 2024 International Conference on Computational Intelligence for Green and Sustainable Technologies (ICCIGST). 2024: 1–6.
10 Fall A, Fall J. A domain-specific language for models of landscape dynamics. *Ecol Model* 2001; **141**: 1–18.
11 Grattafiori A, Dubey A, Jauhri A, *et al.* The Llama 3 Herd of Models. 2024; published online Nov 23. DOI:10.48550/arXiv.2407.21783.
12 OpenAI, Achiam J, Adler S, *et al.* GPT-4 Technical Report. 2024; published online March 4. DOI:10.48550/arXiv.2303.08774.
13 The Claude 3 Model Family: Opus, Sonnet, Haiku. https://www.semanticscholar.org/paper/The-Claude-3-Model-Family%3A-Opus%2C-Sonnet%2C-Haiku/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627 (accessed Dec 31, 2024).
14 Wolf T, Debut L, Sanh V, *et al.* HuggingFace’s Transformers: State-of-the-art Natural Language Processing. 2020; published online July 14. DOI:10.48550/arXiv.1910.03771.
15 Virtanen P, Gommers R, Oliphant TE, *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* 2020; **17**: 261–72.
16 Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with

the natural language toolkit. O'Reilly Media, Inc., 2009.
17 Menstrual Health Awareness Dataset. https://www.kaggle.com/datasets/gurjotkrr/menstrual-health-awareness-dataset (accessed Jan 1, 2025).
18 Ding N, Qin Y, Yang G, *et al.* Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat Mach Intell* 2023; **5**: 220–35.
19 Hu EJ, shen yelong, Wallis P, *et al.* LoRA: Low-Rank Adaptation of Large Language Models. In: International Conference on Learning Representations. 2022. https://openreview.net/forum?id=nZeVKeeFYf9.
20 llama3/MODEL_CARD.md at main · meta-llama/llama3. GitHub. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md (accessed Dec 31, 2024).
21 Team G, Georgiev P, Lei VI, *et al.* Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. 2024; published online Dec 16. DOI:10.48550/arXiv.2403.05530.
22 Jiang AQ, Sablayrolles A, Mensch A, *et al.* Mistral 7B. 2023; published online Oct 10. DOI:10.48550/arXiv.2310.06825.
23 Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. 2019.
24 Mitra A, Corro LD, Mahajan S, *et al.* Orca 2: Teaching Small Language Models How to Reason. 2023; published online Nov 21. DOI:10.48550/arXiv.2311.11045.
25 Almazrouei E, Alobeidli H, Alshamsi A, *et al.* The Falcon Series of Open Language Models. 2023; published online Nov 29. DOI:10.48550/arXiv.2311.16867.
26 Abdin M, Aneja J, Awadalla H, *et al.* Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. 2024; published online Aug 30. DOI:10.48550/arXiv.2404.14219.
27 Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a Method for Automatic Evaluation of Machine Translation. In: Isabelle P, Charniak E, Lin D, eds. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002: 311–8.
28 Banerjee S, Lavie A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In: Goldstein J, Lavie A, Lin C-Y, Voss C, eds. Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, Michigan: Association for Computational Linguistics, 2005: 65–72.
29 Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In: Text Summarization Branches Out. Barcelona, Spain: Association for Computational Linguistics, 2004: 74–81.
30 Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. 2020; published online Feb 24. DOI:10.48550/arXiv.1904.09675.