Supplementary File 2: Knowledge base sourcing and preprocessing

A Bangla-language retrieval-augmented conversational agent for maternal health in a low-resource setting: design and formative evaluation

Challenges in sourcing Bangla maternal health content

The following constraints were encountered when surveying candidate sources for the knowledge base. They are reported in full here because they may inform similar efforts in other low-resource language settings.

1. **Quality of available Bangla resources.** Many online Bangla health resources were direct machine translations of English material, often outdated and lacking cultural nuance.
2. **Limited evidence base.** Few sources were clinically vetted or peer reviewed, raising questions about reliability.
3. **Absence of medical cross-validation.** Most resources showed no evidence of review by certified medical professionals.
4. **Scarcity of localized content.** Little material addressed maternal health realities specific to Bangladesh, such as region-specific nutrition practices and the substantial minority of births that still occur outside health facilities.
5. **Misinformation risk.** Social media, forums, and personal blogs frequently contained unverified claims, folk remedies, or contradictory advice.
6. **Fragmented institutional resources.** Government and nongovernmental organization sites contained useful guidance, but this was often fragmented, outdated, or not available in machine-readable form.
7. **Limited coverage of complications.** Content on perinatal mental health, postpartum recovery, and miscarriage support was particularly sparse.
8. **Absence of structured data.** No publicly available structured or machine-readable Bangla medical corpora were identified.
9. **Terminology inconsistency.** Bangla medical terms lacked standardization, with multiple spellings and colloquial alternatives that complicate retrieval.

Curated sources and permitted use

Two sources met our criteria for accuracy, relevance, and cultural applicability: *Shohay Health* (https://shohay.health), a Bangladeshi health-information initiative that states its content is reviewed by clinicians through an editorial process, and *FirstCry (Bangla)* (https://banglaparenting.firstcry.com), a South Asian digital parenting platform. Both are third-party copyrighted works. We did not seek formal permission from either publisher before crawling. Content was used solely to construct a retrieval index for non-commercial academic research; it is not redistributed in whole or in part, is not reproduced in this article or in any supplementary file, and the deployed system returns generated summaries rather than verbatim source text.

Preprocessing procedures

Articles retained from the two curated sources underwent the following preprocessing steps before embedding. Counts of items removed at each individual step were not retained; 247 articles survived the full pipeline and were segmented into 2,945 indexed passages.

- **Exclusion of irrelevant or low-quality content.** Because the maternal sections of both sites were crawled in full, each article was manually reviewed and those that were promotional or only tangentially related to maternal health were excluded. For example, articles on baby shower gift selection were discarded, while articles on managing gestational diabetes were retained.
- **Removal of promotional and redundant text.** Within retained articles, promotional phrases and repeated boilerplate disclaimers were removed so substantive content remained.
- **Removal of identifiers.** Personal references, author names, and case-specific mentions were removed or generalized.
- **Markup cleanup.** HTML tags, embedded scripts, and page metadata were stripped using regular expressions and parser libraries, retaining only human-readable text.
- **Fact-checking and cross-referencing.** Key medical claims were verified against authoritative sources and reviewed by a collaborating licensed physician. Content conflicting with clinical consensus was corrected rather than retained. For example, a source claim that papaya must be avoided entirely during pregnancy was revised to reflect that only unripe papaya carries risk.

Chunking parameter selection

Text was segmented using a recursive character splitter with a chunk size of 800 characters and an overlap of 300 characters. Smaller chunks (approximately 500 characters) narrowed the focus of individual embeddings but fragmented longer gestational-week guides into disjointed segments. Larger chunks (approximately 1000 characters) preserved broader context but reduced the relevance of retrieved segments by including extraneous material. Overlaps below 200 characters risked losing continuity between the description of a complication and its management, while overlaps above 400 characters increased redundancy and storage cost without apparent benefit. The adopted configuration reflects a balance between coherence and efficiency. Systematic benchmarking across parameter ranges was not conducted and remains a limitation.