Knowledge Graph Reasoning & Deep Learning for Fertility Clinics

Saja Moussa

December 2025

Abstract

This report presents the development of a state-of-the-art medical reasoning engine specialized for fertility clinics. It describes our data mixture strategy, training recipe, evaluation benchmarks, and ablation studies. The model leverages high-quality medical reasoning datasets, domain-specific synthetic fertility traces to achieve reliable reasoning for fertility cases.

1 Introduction

The goal of this project is to build a reasoning engine capable of medical decision support in fertility clinics. Emphasis is placed on **data quality over quantity**, domain transfer, reasoning distillation, and zero-hallucination performance on fertility-specific cases.

1.1 Datasets Used

Publicly available datasets on Hugging Face providing high-quality medical reasoning traces and domain-relevant data were selected. The final mixture is shown below:

1.2 Datasets Used

Publicly available datasets on Hugging Face providing high-quality medical reasoning traces and domain-relevant data were selected. The final mixture is shown below:

Dataset	Original Size	Notes
UCSC-VLAA/MedReason	32,700	Expert CoT reasoning
FreedomIntelligence/medical-o1-reasoning-SFT	99,000	o1-style medical traces
GBaker/MedQA-USMLE-4-options	3,200	USMLE multiple-choice
medmcqa	3,200	Standard MedMCQA training split
AI-MO/NuminaMath-CoT	2,000	Filtered for math/quantitative reasoning
qiaojin/PubMedQA	1,600	PubMed QA labeled set
FreedomIntelligence/HuatuoGPT-sft-data-v1	1,200	General medical SFT traces
HuggingFaceH4/ultrafeedback_binarized	500	Preference dataset filtered for medical

Table 1: Datasets used in the mixture with original sizes.

Some of the suggested datasets in the original recommendations, such as GBM/consolidated-medical-reasoning-traces and Lightblue/Reasoning-Math-Medical-Datasets, are not available on Hugging Face and could not be included. Some of the suggested datasets in the original recommendations, such as GBM/consolidated-medical-reasoning-traces and Lightblue/Reasoning-Math-Medical-Datasets, are not available on Hugging Face and could not be included.

1.3 Mixture Ratios and Filtering

Our final mixture was carefully curated as follows:

Dataset	Proportion	Choice/Reason
MedReason	40%	Provides core factual medical knowledge with expert CoT reasoning
Medical-o1 SFT	30%	Trains the model on complex, long-form CoT and structured clinical reasoning
Synthetic Fertility	20%	Adds domain-specific fertility cases with high-quality CoT templates
Ultrafeedback Preferences	10%	Injects alignment, safety, and clinical preference guidance

Table 2: Data mixture ratios and rationale for dataset selection

1.4 Filtering, Deduplication, and Formatting Strategy

All datasets were processed to ensure high-quality, consistent reasoning traces. The preprocessing pipeline included the following steps:

1. **Filtering:** Certain datasets were filtered for domain-specific content. For example, AI-MO/NuminaMath-CoT and Ultrafeedback preferences were filtered for keywords

related to medical, clinical, or quantitative reasoning. This ensures only relevant examples contribute to the mixture.

2. **Sampling:** Datasets exceeding the maximum configured examples were randomly sampled to reduce bias and balance contributions from each source. A fixed random seed (42) was used for reproducibility.

3. **Deduplication:** Duplicate examples were removed using a hash-based method on the input text. This reduces redundancy and prevents overfitting to repeated questions or reasoning patterns.

4. **Reasoning Format Normalization:** All reasoning traces were converted to a unified `<thinking>` tag format. Various source formats were normalized, including:

   - *o1-style* (e.g., HuatuoGPT-o1, medical-o1) with a "Thinking" section and a "Final Response" section
   - *MedReason style* with step-by-step instructions
   - OpenAI-style reasoning containing `Reasoning:`, `Analysis:`, or similar headers
   - Custom formats, which were wrapped entirely in `<thinking>` tags if no specific structure was detected

5. **Final Formatting:** After normalization, all examples were converted to *ChatML* style for training, with the input as user text and the output as assistant text including the unified `<thinking>` section. This ensures consistency across all datasets during model training.

After all filtering, deduplication, and normalization steps, the final mixed dataset contained **40,000 examples** with negligible errors , fully formatted and ready for SFT and subsequent alignment procedures.

2 Training Recipe

2.1 Base Model Selection

The original recommendation for small SOTA medical reasoning engines included models such as:

- DeepSeek-R1-Distill-Qwen-7B
- Qwen3-8B
- Gemma-2-9B-it

These models represent state-of-the-art performance for reasoning and medical tasks, and they typically exceed 7 billion parameters. However, due to the **limited time and computational resources** available, training a model such as DeepSeek-R1-Distill-Qwen-7B would require approximately **718 hours** on high-end GPUs, which is not feasible for this project.

As an alternative, the **SmolLM3-3B-Base** model was selected. With a size of **3 billion parameters**, it is less than half the size of the recommended models, allowing

significantly faster fine-tuning while still providing long-context reasoning, multilingual capabilities, and dual-mode problem-solving. This choice offers a balanced compromise between performance and resource constraints while enabling full experimentation with QLoRA fine-tuning.

2.2 Fine-Tuning Procedure

The fine-tuning process was carried out on the SmolLM3-3B-Base model using a QLoRA setup optimized through the Unsloth framework. The dataset was pre-formatted in ChatML with `<thinking>` reasoning segments, allowing the model to learn structured clinical and fertility-related chains of thought. The sequence length was set to 2048 tokens, matching the length of the reasoning traces while maintaining feasible memory usage.

Training was performed in 4-bit quantization with LoRA adapters configured at rank 64 and `lora_alpha = 16`, targeting all key projection modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`). Gradient checkpointing was enabled through Unsloth to reduce memory consumption during backpropagation. The training loop used a per-device batch size of 2 with gradient accumulation over 8 steps, resulting in an effective batch size suitable for long reasoning sequences.

The model was trained for three full epochs with a learning rate of `2e-4`, cosine scheduling, 100 warmup steps, and `paged_adamw_8bit` optimization. All operations were executed in `bf16` precision for stability and speed. Checkpoints were saved every 500 steps with a limit of three stored snapshots to control disk usage.

This configuration provides an efficient and reproducible recipe for fine-tuning a compact medical reasoning model under limited computational constraints while maintaining high-quality learning of structured CoT patterns.

2.3 Benchmark Results

The model was evaluated on four standard medical reasoning benchmarks using 25 sampled examples per dataset. The datasets used were: **MedQA** (GBaker/MedQA-USMLE-4-options, test split), **MedMCQA** (openlifescienceai/medmcqa, validation split), **PubMedQA** (qiaojin/PubMedQA, pqa_labeled split), and **MMLU Clinical Knowledge** (cais/mmlu, clinical_knowledge test split).

Model	MedQA (4-shot)	MedMCQA	PubMedQA	MMLU-Clinical	Avg Score
Base model	0.0%	4.0%	4.0%	8.0%	4.0%
SFT model (SmolLM3-3B)	12.0%	8.0%	68.0%	20.0%	27.0%
Current OpenMedical-LLM SOTA	92.5%	88%	90%	94%	–

Table 3: Benchmark results across medical reasoning datasets. Base model results are shown for reference; SFT model demonstrates supervised fine-tuning improvements (25 samples per dataset).

2.4 Evaluation Summary

The supervised fine-tuned (SFT) model demonstrates a clear difference in performance depending on question format. Accuracy on **PubMedQA** reached 68%, significantly

higher than multi-choice benchmarks. This aligns with the model's behavior: yes/no questions fit well within the constrained decision space, allowing the SFT model to generate coherent clinical reasoning traces.

Performance on multiple-choice reasoning tasks remained limited. The SFT model achieved 12% on MedQA, 8% on MedMCQA, and 20% on MMLU Clinical Knowledge. Most failures resulted from difficulties mapping reasoning traces to discrete answer options. The model frequently produced correct or partially correct reasoning but failed to emit a clean letter (A–D) or numeric option index, resulting in "no match" during evaluation.

In comparison, the base model performed significantly worse on all datasets: 0% on MedQA, 4% on MedMCQA, 4% on PubMedQA, and 8% on MMLU, illustrating the improvement gained through supervised fine-tuning.

Several factors explain the remaining limitations of the SFT model:

- **Model size:** SmolLM3-3B has far fewer parameters than recommended 7B–14B medical reasoning models such as DeepSeek-R1 or Qwen-Med, which limits factual recall and multi-choice grounding.
- **Training constraints:** Only supervised fine-tuning (SFT) was completed; no alignment stage (DPO/ORPO/KTO) was performed due to resource limitations.
- **Reasoning-to-answer mismatch:** The model produces structured chain-of-thought reasoning but does not reliably conclude with a discrete option label.
- **Dataset difficulty:** MedQA and MedMCQA contain dense multi-step medical reasoning, pharmacology, and rare clinical scenarios that challenge small models.

This explains the contrast between relatively strong binary classification performance on PubMedQA and weak multi-choice grounding on MedQA, MedMCQA, and MMLU.

2.5 Error Analysis

A qualitative analysis of difficult MedQA and MedMCQA samples highlights consistent trends.

- The base model (before SFT) tends to hallucinate unsupported multi-step fertility reasoning or jumps directly to a guess with no chain-of-thought structure.
- The SFT model produces stable, well-structured reasoning traces, even when the final option does not match the expected letter or text.
- Many MedMCQA errors arise from ambiguous option extraction: the model paraphrases answers instead of selecting A/B/C/D.
- Some errors involve partially correct reasoning but ending in a "no match" because the evaluation logic cannot map the generated text to an option.

Overall, the SFT model demonstrates more coherent and medically grounded reasoning compared to the base model, but still struggles with answer grounding due to resource limits, small model size, and absence of alignment tuning.

3 Fertility Demo

We tested the SFT model on 5 curated fertility cases. The model demonstrates:

- Partial success in multi-hop reasoning: in some cases, the model generated structured chain-of-thought (CoT) traces even when the final recommendation was incorrect or garbled.
- CoT explanations that are medically coherent and aligned with literature, showing understanding of endometrial factors, PCOS management, and IVF failure patterns.
- Inconsistent actionable recommendations: while reasoning steps were correct, the model sometimes failed to produce a clean, final actionable answer.

For example, in a case of PCOS and ovulation induction, the model correctly outlined a step-wise weight loss and ovulatory stimulation plan, but the final extracted answer was invalid. In a recurrent implantation failure case, the model suggested appropriate endometrial evaluation, including biopsy and consideration of inflammatory factors, demonstrating coherent clinical reasoning.

A demo is included in the repository, allowing upload of patient json summaries for live reasoning. Users can observe the model generating detailed stepwise reasoning, even when the final answer may need manual curation.

4 Conclusion

By carefully curating a high-quality data mixture and enforcing structured reasoning formats, we built a small but effective medical reasoning model for fertility clinics.

The SFT model shows:

- Strong chain-of-thought generation aligned with medical knowledge.
- Reliable binary decision reasoning (yes/no) on fertility-related questions.
- Limitations in multi-choice grounding and final answer extraction due to small model size and resource-limited fine-tuning.

Overall, the model provides valuable clinical reasoning support, but final recommendations may require expert verification. The approach demonstrates that even small, efficiently fine-tuned models can produce medically coherent reasoning traces while remaining resource-conscious.