CLNQ Gen-2B WebGL/TensorFlow.js Precomputed Embeddings integration

Live App: https://socr.umich.edu/GAIM/SOCR_CLNQ_2.html

Source: https://github.com/SOCR/GAIM/

Two-Part Solution

Part 1: Offline Embedding Precomputer (*CLNQ_2B_EmbeddingsPrecomputer.html*)

The first app is a dedicated tool for server-side embedding generation that:

- Processes 400K+ clinical terms efficiently with batching and progress tracking
- Exports embeddings as a JSON file for later use (clinical-embeddings-multicore-2025-06-26.json)
- Includes pause/resume functionality for long-running processes
- Provides memory management and optimization features
- Tests embeddings to verify quality before export

Part 2: Optimized Production App (*SOCR_CLNQ_2B.html*)

The second app is the main CLNQ app that:

- Loads pre-computed embeddings instead of computing them live
- Uses multiple fallback sources (local files, Supabase, CDN)
- Performs instant similarity computation using TensorFlow.js tensors
- Maintains all ML capabilities with dramatically improved performance

Implementation Steps

1. Run the Precomputer (Once)

On your powerful server:

1. Save the first app as "CLNQ_2B_EmbeddingsPrecomputer.html"
2. Place your HPO and biomedical files in an "assets" folder
3. Open in a browser and click "Start Precomputation"
4. Wait for completion (2-6 hours depending on hardware)
5. Download the generated "**clinical-embeddings-YYYY-MM-DD.json**" file

2. Deploy Embeddings

Upload the embeddings file to:

- Your web server: assets/clinical-embeddings.json

assets/clinical-embeddings-multicore-2025-06-26.json

- Supabase Storage: /embeddings/clinical-embeddings.json

- Any CDN or cloud storage

3. Update App Configuration

// In the optimized app, update these URLs:

const embeddingSources = [

'assets/clinical-embeddings-multicore-2025-06-26.json', // Local

'https://your-supabase-url.supabase.co/storage/v1/object/public/embeddings/clinical-embeddings.json', // Supabase

'https://your-cdn.com/clinical-embeddings.json' // CDN

];

Performance Improvements

Before (Gen-2)

- ❌ 405,909 embeddings computed at runtime
- ❌ Hours of initialization time
- ❌ Massive memory usage during computation
- ❌ Browser crashes on low-end devices

After (Gen-3 Optimized)

- ✅ Instant loading of pre-computed embeddings
- ✅ ~2-5 second initialization (just model loading)
- ✅ Memory efficient tensor operations
- ✅ Works on all devices including mobile

Expected File Sizes

- Embeddings JSON: ~200-500MB (depending on compression)

  clinical-embeddings-multicore-2025-06-26.json (642MB)

- Compressed (gzip): ~50-150MB when served with compression

  clinical-embeddings-multicore-2025-06-26.json.gz (261MB)

- Memory usage: ~100-300MB in browser (vs 2GB+ before)

Advanced Features

Multiple Embedding Sources

The app tries multiple sources in order:

1. Local assets folder
2. Supabase storage
3. CDN/cloud storage
4. Fallback mode if all fail

Validation & Fallback

- Validates embedding data structure
- Creates fallback embeddings if none available
- Graceful degradation with user notification

Memory Management

- Automatic tensor disposal
- Optimized batch processing
- GPU memory monitoring

Usage Instructions

1. Development: Use the precomputer tool on a powerful machine
2. Production: Deploy embeddings to your preferred storage
3. Scaling: Update embedding sources in the main app
4. Monitoring: Check browser console for loading status

The optimized **CLNQ Gen-2B** version maintains all the ML capabilities of Gen-2 but solves the annoying performance problems.