Turkish AI Research and Contributions

📚 Academic Publications and Research

Peer-Reviewed Articles

Healthcare-Focused Turkish LLM

- **Title**: "Healthcare-Focused Turkish LLM: Training on Real Patient-Doctor Question-Answer Data for Enhanced Medical Insight"
- **Key Contributions**:
  - Utilization of 167,000+ real patient-doctor Q&A data
  - Customized fine-tuning on LLAMA 3 (8b) model
  - Optimization using LoRA and slerp merge techniques
  - Solutions for catastrophic forgetting problems
  - Performance analysis with GPT-3.5 and expert evaluations
- Link: https://docs.google.com/document/d/1l54QlwYtr3rQV0Oy-9zyJGmpcDrYGMaoZ0UdeVj2jMo/edit?usp=sharing

Turkish MMLU Benchmark

- **Title**: "Setting Standards in Turkish NLP: TurkishMMLU for Large Language Model Evaluation"
- **Key Contributions**:
  - 6,200 multiple choice questions
  - 62 different sections, 100 questions each
  - Selected from a pool of 280,000 questions
  - 67 disciplines and 800+ topics
  - Standard evaluation criteria for Turkish NLP
- Link: https://docs.google.com/document/d/1b28ZQmAjh0EWE2_4aSpkvtHC_GIrvfdQu2YHmoP8Cw0/edit?usp=sharing

Conference Paper

Adaptive Learning Rate Study

- **Title**: "Data Quality-Based Adaptive Learning Rate: A Case Study on Medical Text Classification"
- **Key Contributions**:
  - Data quality-based adaptive learning rate
  - Customized approach for medical text classification
  - Dynamic learning optimization based on expert level
  - Implementation on 167,000 samples
  - Performance and convergence improvements
- Link: https://docs.google.com/document/d/13zBC-LaQyjo8wdJl158NpIPs2LRrToZn/edit?usp=sharing&ouid=100950721933293531716&rtpof=true&sd=true

Healthcare-Focused Models and Merge Operations

Models are shared on both huggingface.com/alibayram and ollama.com/alibayram.

Doctor-Llama Series

- **Doctor-Llama-3-8b-slerp-cosmos**
    - Specialized in medical terminology and concepts
    - Enhanced through multiple model merging
    - Performance optimized with SLerp technique
    - Customized for patient-doctor dialogues
- **Doctor-Llama-3-8b-slerp**
    - Built on Llama 3 base model
    - Fine-tuned with Turkish medical literature
    - Performance optimization with SLerp merge techniques

DoctorGemma Series

- **DoctorGemma2-9b**
    - Medical version of the Gemma model
    - Multiple fine-tuning phases:
        1. General medical knowledge adaptation
        2. Turkish health terminology optimization
        3. Patient-doctor interaction fine-tuning
- **DoctorGemma2-9b-adapter**
    - Customized with LoRA adapters
    - Easily updatable due to modular structure
    - Adaptable to different medical sub-fields
- **Doctor-Gemma2-9b-it**
    - Optimized for specific use cases
    - Iterative fine-tuning approach
    - Continuous improvement with performance metrics

Model Merge and Optimization Techniques

- **SLerp (Spherical Linear Interpolation)**
    - Optimal combination of different checkpoints
    - Knowledge transfer between models
    - Performance/size balance optimization
- **LoRA Adapters**
    - Low-rank adaptation techniques

    - Efficient fine-tuning strategies
    - Modular model development approach

📊 Datasets and Evaluations

Academic Evaluation Datasets

Turkish MMLU (Multi-task Massive Language Understanding)

- **Scope**: Comprehensive evaluation set specific to Turkish education system
- **Content**:
    - Selected from 280,000+ question pool
    - 67 different disciplines
    - 800+ different topics
- **Use Cases**:
    - Model performance evaluation
    - Academic proficiency measurement
    - Comparative analysis of Turkish language models

MMLU Evaluation Sets

1. **turkish_ai_mmlu_model_answers**
    - Responses from different AI models
    - Comparative performance analysis
    - Model behavior examples
2. **turkish_ai_mmlu_leaderboard**
    - Model performance rankings
    - Comparative metrics
    - Development analysis over time
3. **turkish_ai_mmlu_section_results**
    - Detailed section-based analyses
    - Topic-based performance evaluations
    - Specific domain success rates

Professional Domain Datasets

Medical Dataset (doctorsite)

- **Scope**: 167,000+ real patient-doctor interactions
- **Features**:
    - Expert doctor responses
    - Various medical specialties
    - Patient questions and complaints
    - Diagnosis and treatment recommendations

- **Use Cases**:
  - Medical language model training
  - Healthcare consulting systems
  - Medical terminology analysis

Legal Dataset (legal_qa)

- **Content**:
  - Turkish law Q&A data
  - Legal terminology and concepts
  - Legal analysis examples
- **Use Cases**:
  - Legal text analysis
  - Legal consulting systems
  - Legal language modeling

User Feedback Datasets

E-commerce Reviews

1. **hepsiburada_reviews**
   - Product evaluations
   - Customer satisfaction analysis
   - Purchase experiences

Media and Entertainment Reviews

1. **beyazperde_reviews**
   - Movie and series evaluations
   - Viewer comments
   - Media content analysis
2. **kitapyurdu_reviews**
   - Book reviews
   - Reader evaluations
   - Literary content analysis

News and Content

1. **onedio_news**
   - News texts and headlines
   - Content categories
   - Current events

💻 Open Source Contributions

GitHub Project Contributions

- **mlx-examples**
    - Apple Silicon optimized ML examples
    - Performance improvements
    - Turkish documentation
- **unsloth**
    - LLM optimization tools
    - Training process acceleration
    - Memory optimization