Mehmet Ali Bayram

Email: malibayram20@gmail.com
Phone: +90 551 729 14 10
Location: Ataşehir, Istanbul, Turkey

Profile Summary

PhD candidate in Computer Engineering with a linguistic and multicultural background, fluent in Turkish, Kurdish, English, Arabic, and conversant in Persian and Zazaki. Advancing NLP research for low-resource languages, dialects, and specialized domains. Skilled in multilingual tokenizers, LLM fine-tuning, adaptive learning, and model merging, focusing on computational efficiency and semantic depth. Extensive experience in medical and Turkish-focused LLM research, open-source NLP projects, and context-sensitive modeling that addresses linguistic structures and cultural nuances. Integrates diverse linguistic knowledge and advanced AI frameworks to create models that excel in real-world communication and bridge global language communities and computational intelligence.

Education

PhD in Computer Engineering (2020 - 2025), Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü, Turkey

- Specializing in NLP, AI model architectures, tokenization, and low-resource domain-specific LLMs.
- Developed adaptive learning, model merging, and linguistic principles for AI-driven language understanding.

Visiting PhD Researcher (Feb 2023 - Jul 2023), Instituto Politécnico de Tomar (Erasmus), Portugal

- Collaborated internationally on multilingual NLP frameworks and algorithmic optimization.
- Integrated diverse linguistic datasets and advanced architectures for improved multilingual performance.

MS in Information Technology (2018 - 2020), Marmara University, Turkey (GPA: 4.00)

- Focused on data management, ML implementations, and IT methodologies.
- Completed research applying ML and data analytics, forming a basis for advanced NLP and AI work.

MSc in Information Technology (Jan 2020 - May 2020, Erasmus), Aristotle University of Thessaloniki, Greece

- Expanded international academic perspectives in IT and data science.
- Conducted comparative studies for cross-lingual, cross-cultural data processing insights.

BS in Computer Engineering (Oct 2020 - Aug 2022)

- Built foundational computer engineering principles, algorithms, and software design.
- Gained technical proficiency for NLP research, model deployment, scalability, and optimization.

Additional Degrees and Studies:

- Web Design & Coding (Associate, Anadolu University, 2018-2020): Front-end/back-end and UI/UX.
- Emergency & Disaster Management (Bachelor's, Istanbul University, 2014-2018): Crisis management and problem-solving, applicable to complex AI research challenges.
- Justice (Associate, Anadolu University, 2015-2017): Analytical reasoning and legal understanding, informing ethical and regulatory aspects of AI.

Key Skills

- **AI & NLP:** LLM training, fine-tuning, evaluation; Transformers; tokenization standards; LoRA adapters; SLerp merges; MMLU benchmarking.
- **Research & Development:** Domain adaptation, adaptive learning rates, data quality-driven optimization, Turkish NLP standards, cross-lingual strategies for global user bases.
- **Programming & Tools:** Python, TensorFlow/Keras, PyTorch, Hugging Face, GitHub, Docker, CI/CD, scalable infrastructure.
- **Domains:** Healthcare-focused LLMs, Legal Q&A modeling, Academic benchmarking, E-commerce and media text analysis.

Selected Academic Publications & Research

Healthcare-Focused Turkish LLM:

- "Healthcare-Focused Turkish LLM: Training on Real Patient-Doctor Q&A Data"
- Fine-tuned Llama 3 (8B) with 167k+ patient-doctor Q&A.
- Applied LoRA adapters, SLerp merges, and evaluated against GPT-3.5 and expert feedback.

Turkish MMLU Benchmark:

- "Setting Standards in Turkish NLP: TurkishMMLU for Large Language Model Evaluation"
- 6,200-question benchmark from 280k+ questions, covering 62 sections, 67 disciplines, 800+ topics.
- Established Turkish LLM performance standards and improved tokenization, semantics, and cross-domain adaptation.

Conference Paper on Adaptive Learning Rate:

- "Data Quality-Based Adaptive Learning Rate: A Case Study on Medical Text Classification"
- Dynamic learning rate tied to content quality, improving convergence in a 167k medical dataset.
- Enhanced accuracy and efficiency for specialized NLP applications.

Model Development & Contributions

Doctor-Llama & DoctorGemma Series:

- Developed Turkish medical LLMs from 167k+ patient-doctor Q&A data.
- Used LoRA adapters, SLerp merges for knowledge retention, reducing catastrophic forgetting.

Broader Contributions & Open Source:

- Integrated medical-domain insights into general LLM development.
- Experimented with architectures, embeddings, attention for scalable, reliable NLP models.
- Collaborated with domain experts to ensure ethically guided, domain-attuned AI assistants.

Datasets & Benchmarking

Turkish MMLU Suite:

- 6,200-question framework from 280k+ questions, spanning multiple domains and topics.
- Guides tokenization, LLM evaluation, cross-lingual transfer, and culturally aware solutions.

Medical Dataset (167k Patient-Doctor Q&A):

- Authentic clinical dialogues for specialized terminology and reasoning.
- Enables LLMs with medical comprehension bridging theoretical NLP and practical healthcare.

Legal, Consumer Review & Other Specialized Datasets:

- Broad domain coverage ensures contextual precision and semantic nuance.
- Adaptation across academic, e-commerce, and media data enables LLMs to excel in varied environments.

Open Source & Community Contributions

Tokenizer Projects:

- github.com/malibayram/tokenizer, malibayram/tokenizer_benchmark
- Linguistically informed tokenizers ensuring semantic richness, morphological integrity, and efficiency.

Model Architectures:

- github.com/malibayram/model-architectures
- Explored attention mechanisms, embeddings, architectures, performance profiling, and scalability.

Additional Tools & Optimizations:

- Apple Silicon optimization (mlx-examples) and training efficiency (unsloth) for improved resource usage, memory management, and training throughput.

Teaching & Mentorship

- Research Assistant at Yeditepe University: Supported AI, programming, and ML fundamentals.
- Udemy and YouTube Content Creator: delivering structured, up-to-date course contents for programming and AI.
- Mentorship in Flutter, Android, NLP bootcamps, and clubs: Guided learners in code efficiency, NLP techniques, and culturally informed AI approaches.

Encouraged creativity, critical thinking, and ethical technology use, producing graduates and practitioners equipped to meet global AI and NLP challenges.

Languages

- Turkish, Kurdish (Native), English, Arabic (Fluent), Persian, Zazaki (communicative proficiency).

Test Scores

- YÖKDİL (English): 87.50
- DGS (Quantitative): Ranked 338th among 300,000+ participants, reflecting strong quantitative reasoning and analytical skills.