Mehmet Emre Tiryaki

mtiryak1@jhu.edu | linkedin.com/in/mehmet-emre-tiryaki | github.com/MehmetETiryaki

Education

Johns Hopkins University Aug. 2026 – Dec. 2026 (expected)
M.S.E. in Applied Mathematics & Statistics (combined B.S./M.S.E.) *Baltimore, MD*

- Coursework: Modern Convex Optimization, Probability Theory, Random Matrix Theory, Probabilistic Machine Learning, Matrix Analysis & Linear Algebra, Data Science Methods for Large-Scale Graphs.

Johns Hopkins University Aug. 2022 – May 2026
B.S. in Computer Engineering and Applied Mathematics & Statistics, Minor in Economics *Baltimore, MD*

- GPA 3.70 / 4.00; General Honors and Departmental Honors; Dean's List, 7 semesters.
- Coursework: Machine Learning, Natural Language Processing, Audio Signal Processing, Information Theory, Mathematical Statistics, Introduction to Optimization, Signals & Systems.

Research Experience

Researcher, AI Health-Coaching Team, JSALT 2026 Jun. 2026 – Present
JHU Center for Language and Speech Processing; Advisor: Dr. Mahnoosh Mehrabani *Baltimore, MD*

Uncertainty and safety for the health-coaching agent

- Core contributor to an LLM health-coaching system with simulated users and human oversight (pub. [1]) at JSALT, a six-week invited research workshop (Jelinek Summer Workshop on Speech and Language Technology).
- Worked on uncertainty and safety: when the agent should answer, clarify, retrieve, redirect to a professional, or escalate to a human, and the safety layer that screens every reply and routes rejected ones accordingly.
- Developed the evaluation set by injecting targeted edits into real coaching conversations, each paired with its unedited original as control: **3,006 pairs** across clarify (357), retrieve (249), redirect (1,200), and escalate (1,200).
- Showed why in-dialogue evaluation matters. Health questions the model handles well as standalone multiple-choice items it got right only **52%** of the time when injected into real conversations (249 questions, 25% without retrieval).
- On the clarification pairs (three failure types), the agent and 8 standard uncertainty methods stayed confident on unanswerable turns (AUROC 0.39–0.52), and token probability barely moved.
- Replaced whole-reply disagreement, which never agrees even when nothing is ambiguous, with a cross-model panel asked a targeted question about the turn: **AUROC 0.96** on withheld referents, 0.71 on two-candidate ambiguity.
- Presented this part at the workshop's closing session [video].

Dialogue-level uncertainty quantification for LLM agents

- Developing a dialogue-level uncertainty quantification method for multi-turn agents that predicts, mid-conversation, that the agent will fail its task, early enough to re-plan, ask the user, or hand off.
- Formulated it as the agent's risk of failing the task under the user's unknown goal, estimated by branching the conversation forward. Every uncertainty score then estimates that risk and maps to one of those actions.
- Found that per-turn scores and the published dialogue-level methods are **near chance** at predicting task failure (38 scores, best AUROC 0.65), because most failures are silent wrong actions the user never reacts to.
- Now studying which quantities must be estimated to recover the risk and where in the dialogue the information is lost, and testing a label-free method based on how much the next action depends on what the user has not said.
- Released a 21-method uncertainty toolkit with agent evaluations, **used by BBN Technologies** (ARPA-H).

AI Research Intern, Team EMMA (Multi-talker ASR), JSALT 2025 Jun. 2025 – Aug. 2025
JHU Center for Language and Speech Processing *Brno, Czechia*

Advisors: Prof. Sanjeev Khudanpur and Dr. Matthew Wiesner

- Selected as one of 7 undergraduates for the workshop and ran my own sub-project within the multi-talker speech recognition team: text-only speaker attribution, a drop-in stage for ASR systems without speaker labels.
- Developed an LLM system recovering who said what from the ASR transcript alone, with no acoustic features, which also probes how much conversational structure a text-LLM backbone captures on its own.
- Fine-tuned Qwen 2.5 to predict each utterance's speaker from the labeled context, decoded with beam search.
- Traced the remaining errors to mistakes propagating through the context window and to collapse onto one speaker in long calls, and removed both by conditioning the search on a fine-tuned DeBERTaV3 speaker-change classifier.
- Reached **18.7% cpWER on CALLHOME**: 3.4x fewer errors than the text-only baseline, 21% fewer than audio.

- Showed that the LLM alone does not solve the task, as its scores break down beyond two speakers and many labelings look equally plausible, and that structure (speaker-change evidence and search) is necessary.
- Cut inference cost with distillation, quantization, and FlashAttention. Presented at the closing session [video].

Undergraduate Researcher, CASPER Speech Dataset (Meta-funded) Feb. 2023 – Dec. 2025

JHU Center for Language and Speech Processing; Advisor: Dr. Leibny Paola Garcia *Baltimore, MD*

- Joined as a freshman and stayed three years, the only team member to do so, building CASPER, a 200-hour corpus of spontaneous phone calls between acquaintances for training speech language models (pub. [2]).
- Built the pipeline (Python, MySQL, SLURM) that turned raw recordings into the de-identified 208-speaker corpus: voice-activity segmentation, metadata management, LLM-assisted removal of personal information.
- Built diarization references from each participant's own channel and ran the paper's ASR and diarization baselines.
- Designed a graph algorithm for speaker-disjoint, metadata-balanced train/test splits via KL minimization.
- Released the corpus and splits on Hugging Face [dataset] and wrote major technical sections of the paper.

Publications

- [1] Sumba, X., Jain, V. D., Ong, A. Q. C., Gundecha, H., **Tiryaki, M. E.**, Chandupatla, S., Qiu, H., Kandula, H., Karakos, D., Mehrabani, M., Gligorić, K., Bojic, I., & Tänzer, M. "A Generative-AI Platform for Developing Health Coaching Agents with Simulated Users." *NeurIPS 2026 Workshop on Generative AI for Health (GenAI4Health)*, demonstration track, accepted. [paper]
- [2] Xiao, C., Liang, R., Zhang, X., **Tiryaki, M. E.**, Bae, V., Shankar, L., . . . & Garcia Perera, L. P. "CASPER: A Large Scale Spontaneous Speech Dataset." *Proc. IEEE ASRU 2025.* arXiv:2506.00267 [paper]

Presentations

- [P1] Solak, R. E.*, **Tiryaki, M. E.***, & Tanomkiattikun, S.* "Predicting and Gaining Insights on Olympic Games: A Temporal Hierarchical Bayesian Network Approach." Poster, *SIAM Washington–Baltimore (DMV) Area Conference*, 2025. (*equal contribution)
- [P2] Solak, R. E.*, & **Tiryaki, M. E.*** "Predicting and Gaining Insights on Olympic Games: A Temporal Hierarchical Bayesian Network Approach." Talk, *2nd Mid-Atlantic Research Exchange (MATRX)*, Johns Hopkins University, Mar. 2025. (*equal contribution)

Selected Projects

Single-Cell Analysis of Parkinson's Disease Microglia Apr. 2026 – Present

- In collaboration with the Dawson Lab, studying whether the brain's immune cells (microglia) shift state in Parkinson's disease, consistently across cohorts. Built the analysis from data acquisition to statistics.
- Integrated **6.8M nuclei** from 226 donors across 4 public atlases, 12 brain regions, and 2 assay chemistries in a Python pipeline running from quality control to automatically generated reports.
- Re-clustered 380K microglia with scVI and ran donor-level DESeq2, scCODA, pathway enrichment, and cross-cohort replication with FDR control. Cut peak memory from 200 GB to 4 GB with streaming merges.

Temporal Hierarchical Bayesian Model of Olympic Medals Feb. 2025

- Formed and led the JHU team in COMAP MCM 2025, the four-day international modeling contest: **Honorable Mention among 20,000+ teams** and the JHU AMS Mathematical Modeling Prize (pres. [P1, P2]).
- Built a temporal hierarchical Bayesian network in which country effects drive per-sport strength as geometric Brownian motion, and wrote a tie-aware truncated Plackett–Luce likelihood. Fit by stochastic variational inference.
- Predicted **72%** of 2024 country-sport medal counts within 10% and forecast 2028 from 2,000 posterior draws. A Kalman-filter anomaly detector flags coaching-driven shifts.

Graph Foundation Model for Single-Cell Multi-Omics Jan. 2026 – Jun. 2026

- Self-supervised model encoding each cell as a graph of genes, proteins, and metabolites (relational GAT), trained on unpaired RNA and ATAC cells with Sinkhorn optimal transport.

MoE–Mixer Interactions in Mamba vs. Transformer Decoders May 2025

- Tested whether the token mixer and the MoE routing interact. In a course project, pretrained 9 decoder LMs, and ReMoE was **best with attention but worst with Mamba** (cross-entropy 3.93 vs. 5.11).

Song Identification from 3-Second Audio Clips Dec. 2025

- Tested whether a learned fingerprint beats Shazam-style hashing (two-person course project): built the conformer encoder, worked on the hashing baseline, reaching about **96%** vs. 78% accuracy with a 9x smaller index [code].

Industry & Entrepreneurial Experience

Co-Founder and Technical Lead, Altai Health Apr. 2026 – Sep. 2026

AI post-surgical follow-up platform; JHU NSF I-Corps program (completed Sep. 2026) *Baltimore, MD*

- Co-founded a startup whose agent checks on post-surgical patients by text or phone and flags who needs attention.
- Built the product as the team's only engineer. The agent covers symptoms, wound care, medication, mood (PHQ-9), and social needs, then files a triage summary (Urgent / Nurse follow-up / Routine) to a provider dashboard.
- Designed it as a state machine that uses LLMs only for bounded steps, with a rule-based safety check the LLM cannot override that can stop the conversation and escalate to a human.
- Demo in 7 days. Interviewed JHU physicians and tested with 3 post-operative patients.

NLP Intern, Turkcell Jun. 2024 – Aug. 2024

LLM evaluation for customer-facing chatbots *Istanbul, Turkey*

- Benchmarked LLM-as-judge RAG evaluators (ARES, RAGAS); proposed a correlation test to skip redundant runs.
- Built LLM evaluation pipelines with Llama 3 for the company's chatbot systems.

Teaching

Lab Instructor, JSALT 2026 Summer School Jun. 2026

JHU Center for Language and Speech Processing *Baltimore, MD*

- Wrote a Jupyter-notebook lab, *Hands-on Uncertainty Analysis and Calibration for LLMs*: token-probability, sampling-based, and semantic-entropy measures, their evaluation (AUROC, ECE, Brier), and calibration methods.
- Ran the tutorial for researchers and students in the Uncertainty-Aware Conversational AI track [program].
- Built the accompanying competition: predict whether the answer is correct on a hidden test set, live leaderboard.

Course Assistant, Johns Hopkins University Aug. 2023 – Present

Head CA, Machine Learning; CA, Intro. to Optimization and Gateway Computing: Python *Baltimore, MD*

- Machine Learning (Head CA Aug. 2026 – present; CA Aug. 2025 – Dec. 2025), Introduction to Optimization (Aug. 2025 – present), Gateway Computing: Python (Aug. 2023 – May 2025), incl. 200+-student semesters.
- Recitations, office hours, writing and grading problem sets, exam questions (ML), lecture review sessions (Optimization). As Head CA, coordinate grading and recitations across the CA team.

Honors & Awards

- **Honorable Mention**, COMAP Mathematical Contest in Modeling (20,000+ teams worldwide) 2025
- Mathematical Modeling Prize, JHU Department of Applied Mathematics & Statistics 2025
- General Honors and Departmental Honors (Computer Engineering), Johns Hopkins University 2026
- Defne-Muhtar Kent Educational Foundation scholarship (selective, merit-based, four years) 2022 – 2026

Technical Skills

Languages: Python, C++, C, SQL, MATLAB, Bash, TypeScript
ML frameworks: PyTorch, Hugging Face Transformers, scikit-learn, Pyro, scvi-tools
Speech and LLM tooling: Whisper/WhisperX, pyannote.audio, Silero VAD, vLLM, LoRA, structured outputs
Systems: SLURM GPU clusters, Docker, AWS, Git, Linux, Weights & Biases, PostgreSQL, Next.js