Research Statement

Fatemeh Shahrokhshahi
ftm.shahrokhshahi@gmail.com | fatemeshahrokhshahi.github.io

Motivation

Large language models are now embedded in tools used daily by millions of people for writing, decision support, education, and healthcare. Most of these users are not AI experts. They cannot distinguish a confident but logically wrong answer from a correct one. This asymmetry is what drives my research: when an LLM fails silently on a basic logical inference, the consequences fall on people who had every reason to trust it.

I came to this problem from a practical concern rather than a theoretical one. As LLMs became mainstream tools, I became aware of a persistent gap between the confidence these systems project and the reliability they actually deliver, particularly in reasoning. This led me to ask three questions that have structured my work: where exactly do LLMs fail logically, why do they fail there, and can we know in advance when a failure is coming?

Research to Date

My first project investigated why LLMs fail on modal and conditional inference patterns involving operators like "must" and "might". I conducted root cause analysis for each failure type and developed LogiCue, a pattern-specific prompting framework grounded in modal logic and philosophy of language. Each intervention targets the precise mechanism behind a given failure rather than applying a generic reasoning nudge. The work was published in Procedia Computer Science (Elsevier), presented at ACLing 2025.

The follow-up work asked a harder question: can we predict which reasoning problems will cause failure before querying the model at all? I extracted structural features from problem texts and trained external classifiers to predict per-instance LLM failure across a large set of instances, models, and prompting strategies. The results showed that failures are structurally predictable to a meaningful degree, and that the pattern of predictability varies significantly across inference types. A finding I find particularly important: targeted prompting reduces failures substantially but makes the residual errors harder to predict, suggesting that input-side and output-side reliability mechanisms are complementary rather than substitutable. This work is published at KI 2026 (Springer LNAI), where I will present it in Bremen this August.

In parallel, I have started looking at the problem from a different angle: whether reasoning rules can be built into a model's architecture rather than only invoked through prompting. In work presented at the MeMo Workshop on Mechanistic Interpretability and Neuro-symbolic Approaches (University of Rome Tor Vergata), I explored storing inference rules explicitly within an associative memory architecture's

correlation matrix. This is early and exploratory, but it reflects a question I keep returning to: prompting can correct behavior at inference time, but I want to understand whether reasoning failures can be addressed further upstream, in how a system represents and applies rules in the first place.

Research Vision

My work so far has focused on single-step logical inference in controlled settings. The natural extension is to agentic systems that plan, reason across multiple steps, and act in the world. The failure modes I have studied do not disappear in agentic pipelines; they compound. A single invalid inference in a chain of reasoning can propagate through a multi-step plan, producing consequential errors that no individual output check would catch.

I want to pursue three directions in my PhD. First, extending failure prediction from single-step inference to multi-step agentic reasoning, building predictors that can flag structurally dangerous reasoning chains before execution. Second, developing evaluation frameworks that go beyond aggregate accuracy to characterize the structure of failure: which patterns fail together, why, and under what conditions. Third, exploring the relationship between prompting interventions and abstention mechanisms, the two-stage reliability framework my current work points toward.

These directions connect directly to AI safety and the rigorous evaluation of frontier models. Safe deployment of LLMs in consequential domains requires not only improving average performance but understanding the boundaries of reliable behavior. I believe that knowing when not to trust a model is as important for AI safety as any alignment technique applied after the fact.

Background

I am completing an MSc in Computer Engineering at Istanbul Aydin University, where I have satisfied all degree requirements, including the department's publication requirement, with three completed works to date (two published, one accepted). I have worked on all three projects described above independently within my Master's program, and I have hands-on experience designing and running large-scale LLM evaluations across multiple model families. I hold an IELTS Academic score of 7.0 and have served as a PC member for EITS 2026.

I am at an early stage of my research career and I know it. My knowledge has real limits that I work to expand through reading and honest engagement with work that challenges my assumptions. What I can offer is a clear problem motivation, demonstrated ability to produce research independently, and a genuine commitment to the question of how we build AI systems that people can actually trust.