Zihao Zheng

Production-Scale Trustworthy AI Architect | Interpretable & Verifiable AI Systems | Apache MXNet PMC Member

E-mail: zihaolucky@gmail.com GitHub: github.com/zihaolucky
LinkedIn: linkedin.com/in/zihaolucky Apache: zihaolucky@apache.org

PROFESSIONAL SUMMARY

AI architect with ∼10 years of practice whose defining discipline is building AI systems that are **explainable**, **verifiable**, and **reliable** at production scale. Rather than constructing opaque models and approximating explanations after the fact, his work designs interpretability into the architecture itself — explicit, auditable mechanisms that operators can inspect, diagnose, and trust.

This philosophy has produced three independent contributions: (1) cross-framework AI observability infrastructure adopted by AWS, PyTorch, and acknowledged by Google; (2) a patented NLU architecture whose routing decisions are traceable through attention weights by design; and (3) a patented ASR architecture with an explicit binary gate making every biasing decision auditable at the timestep level. Each was deployed at production scale (40M+ households) and subsequently recognized by unrelated organizations across industries and countries.

CORE CONTRIBUTIONS TO INTERPRETABLE & VERIFIABLE AI

1. Cross-Framework AI Observability Infrastructure

Standalone TensorBoard (`dmlc/tensorboard`), 2016–2017

- **Problem:** TensorBoard was the most widely adopted ML visualization tool, but it was tightly coupled to TensorFlow — unusable by practitioners working in MXNet, PyTorch, or any other framework. This framework lock-in meant a large portion of the ML community was cut off from essential training observability (loss curves, gradient distributions, activation histograms, embedding projections).
- **Solution:** Broke the framework lock-in by designing and delivering a complete standalone TensorBoard: pure-Python event-logging API, Bazel-buildable rendering stack, and PyPI distribution — so that any practitioner, in any framework, could visualize and inspect model behavior without switching ecosystems.
- **Adoption:**
  - AWS built its official `mxboard` library on this architecture (README credits author by name).
  - PyTorch community's `tensorboardX` referenced and extended the approach.
  - Rajat Monga, Google's TensorFlow engineering director, publicly acknowledged the contribution.
  - Cited in Amazon's SOCKEYE paper (ACL Anthology, 2018).
- **Interpretability significance:** Removed a barrier that had prevented non-TensorFlow practitioners from inspecting model internals, making model transparency accessible across the broader ML community.

2. Inherently Interpretable NLU Architecture

Patent CN112002313B (first inventor), filed 2019, granted 2023. Assignee: Alibaba Group.

- **Problem:** The standard industry approach was "first rules, then model": product teams manually crafted matching templates per domain, and the model served as fallback. As the platform grew to dozens of domains managed by multiple teams, templates contradicted one another — a new rule fixing one domain's bad cases would steal traffic from another. The rule system itself became the primary source of production failures, and cross-team coordination around it was unworkable. Diagnosing which rule caused a specific misroute required tracing through layers of conflicting logic with no structured accountability.
- **Solution:** Memory-Network domain classifier that fuses rule and template signals with learned semantic representations through memory mechanism. Instead of rules and model competing as separate layers, the network learns *when* to trust a rule match and when to override it based on contextual semantics — so product teams can continue adding rules without destabilizing the system. The mechanism makes the rule-model interaction inspectable: for any routing decision, operators can examine which rule signals the model relied on and which it overrode.
- **Production result:** Domain-classification error rate reduced by 35%; overall routing accuracy raised from ∼65% to 85%. Deployed to 40M+ households via Alibaba's Tmall Genie.
- **Independent recognition:** Examiner-entered prior-art citations from Ping An Life Insurance (Fortune Global 500, finance), Tianjin University (Project 985, energy/power-grid forecasting), and embedded AI manufacturers.

3. Inherently Verifiable ASR Architecture (Indicator Loss)

Patent CN113808593B (first inventor), filed 2020, granted 2025. Assignee: Alibaba Group.

- **Problem:** Google's Contextual LAS (CLAS) used implicit decoder signaling for personalized recognition — making it impossible to diagnose *why* a recognition failure occurred at any given timestep. The system was a black box with respect to its biasing decisions.
- **Solution:** Explicit, supervised, binary gate ("Indicator Loss") that directly teaches the model, at each timestep, whether to use personalized content. Every biasing decision produces a verifiable 0/1 audit trail — the architecture's own output, not an approximation. Engineers can inspect exactly where and why errors occur.
- **Production result:** Outperformed Google CLAS in head-to-head evaluation. Deployed to 40M+ households. Alibaba Cloud materials still describe the architecture in production use as of 2023.
- **Academic activity in the same direction:** Between 2023 and 2025, research groups in four countries published work at Interspeech exploring explicit step-level supervision for contextual biasing — the same methodological direction as the patent, filed approximately five years earlier.

PROFESSIONAL EXPERIENCE

Algorithm Expert, PDD Holdings (NASDAQ: PDD) *Dec. 2021 – Present*
Shanghai, China. Core algorithm expert on TEMU's discovery and matching systems.

Algorithm Expert — Core AI Architect, Tmall Genie, Alibaba Group *Jul. 2015 – May 2021*
Holding Ltd.
Hangzhou, China. First-wave AI architect for Alibaba's smart-speaker platform (40M+ households). Designed and deployed both the NLU domain-classification engine and the end-to-end ASR architecture.

OPEN-SOURCE GOVERNANCE

Project Management Committee (PMC) Member, Apache MXNet — The Apache Software Foundation *2022 – 2023*

Board-appointed. Among 51 PMC members out of ∼875 contributors (top 5.7%). Voting and project-governance responsibilities.

EDUCATION

B.Sc. in Information and Computing Science, South China Normal University *2011 – 2015*

School of Mathematics, Guangzhou, China