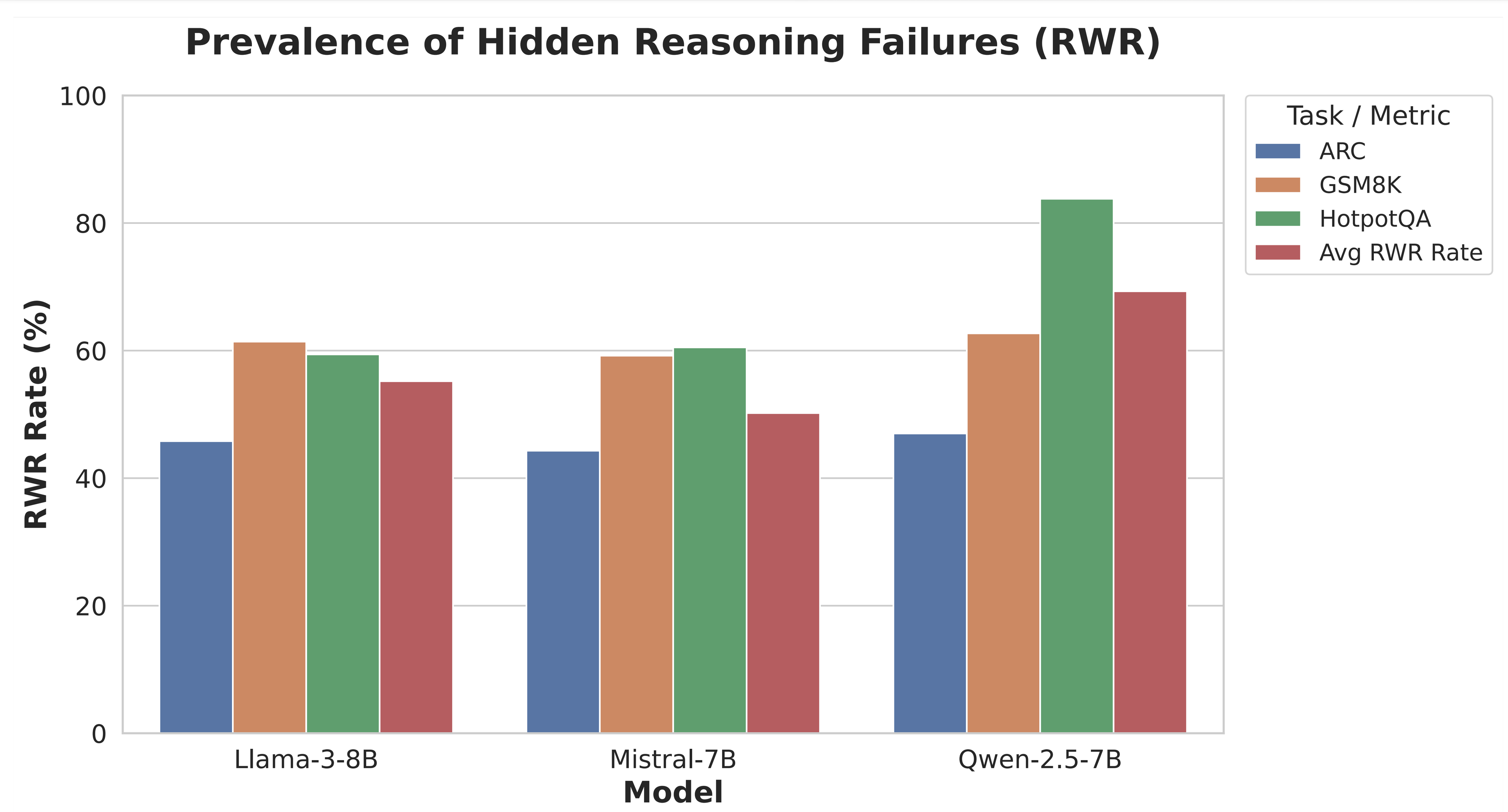
Small Models - Right for the Wrong Reasons
Process Verification for Trustworthy Agents

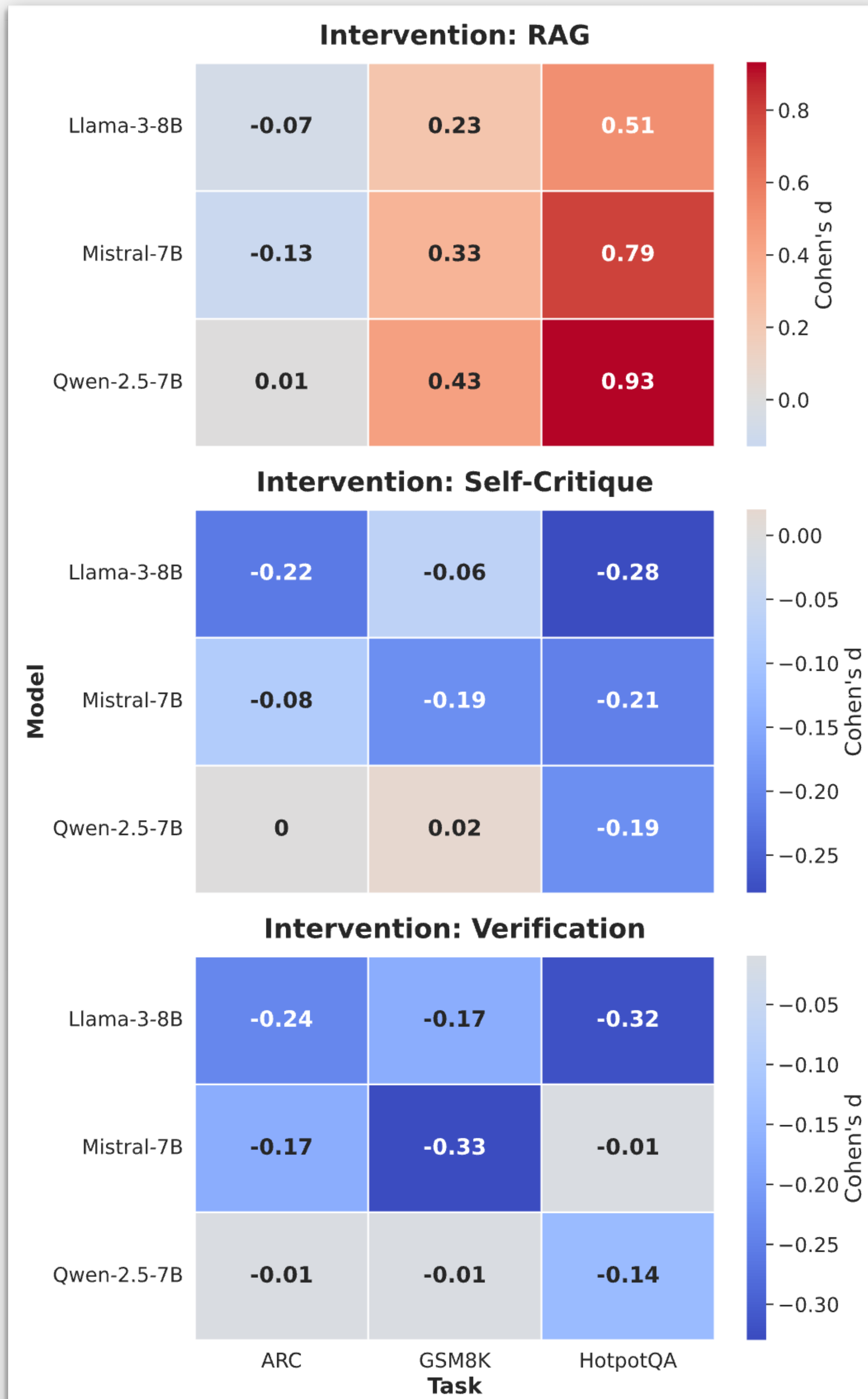
Laksh Advani - laksh.advani@colorado.edu

University of Colorado **Boulder**

Right for the Wrong Reasons (RWR)

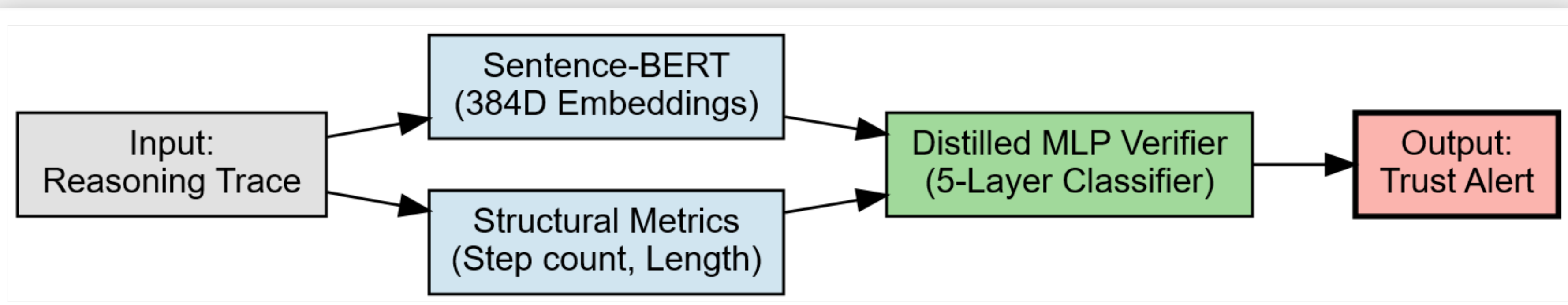
- **Core Problem:** Small language models (<10B parameters) frequently produce correct outputs using logically flawed internal reasoning.
- **RWR Prevalence: 50–69%** of correct answers contain reasoning failures that standard accuracy benchmarks fail to detect.
- **Deployment Risk:** These latent errors compound during autonomous agent tasks, leading to catastrophic system failure.
- **Case Study:**
  - **Task:** Calculate 15% of 80.
  - **Model Step:** Multiplying 80 by 0.2 (instead of 0.15).
  - **Result: 12** (Numerically correct somehow, logically flawed)





Data Analysis & Solution Refinement

- **Quantification:** Analyzed **~11k reasoning traces** using step-by-step RIS (Reasoning Integrity Score) scoring.
- **Benchmark Coverage:** Evaluated on **GSM8K** (Math), **HotpotQA** (Multi-hop), and **ARC** (Commonsense).
- **Interventions:** Tested RAG, Self-Critique, and Verification Prompts.
- **The Trust Alarm:** Created a distilled MLP classifier to detect flawed reasoning ($RIS < 0.8$) in pseudo real-time.
- **Detection Accuracy:** Achieved a **0.86 F1-score** and **0.88 precision** on identifying flawed traces.
- **Operational Efficiency: 100x faster** than LLM judging with **5–10ms latency**, making it suitable for live agent monitoring.



Conclusions

- **Accuracy as a Lagging Metric:** Output-based accuracy is a deceptive and insufficient proxy for reliability in models **<10B parameters**.
- **The "Trust Alarm" Utility:** The distilled verifier (**0.86 F1**) acts as a real-time safety layer, flagging flaws for review with a **100x speedup** over LLM judging.
- **RAG Mandate:** Retrieval-augmented generation is required for factual tasks to provide "external scaffolding" and mitigate internal logic drift.
- **Self-Critique Restriction:** Avoid meta-cognitive prompting in small models; it frequently triggers "pseudo-reflection," which increases error rates.

Architectural Standard

Process-based verification (not just output monitoring) is non-negotiable for deploying trustworthy autonomous agents.