

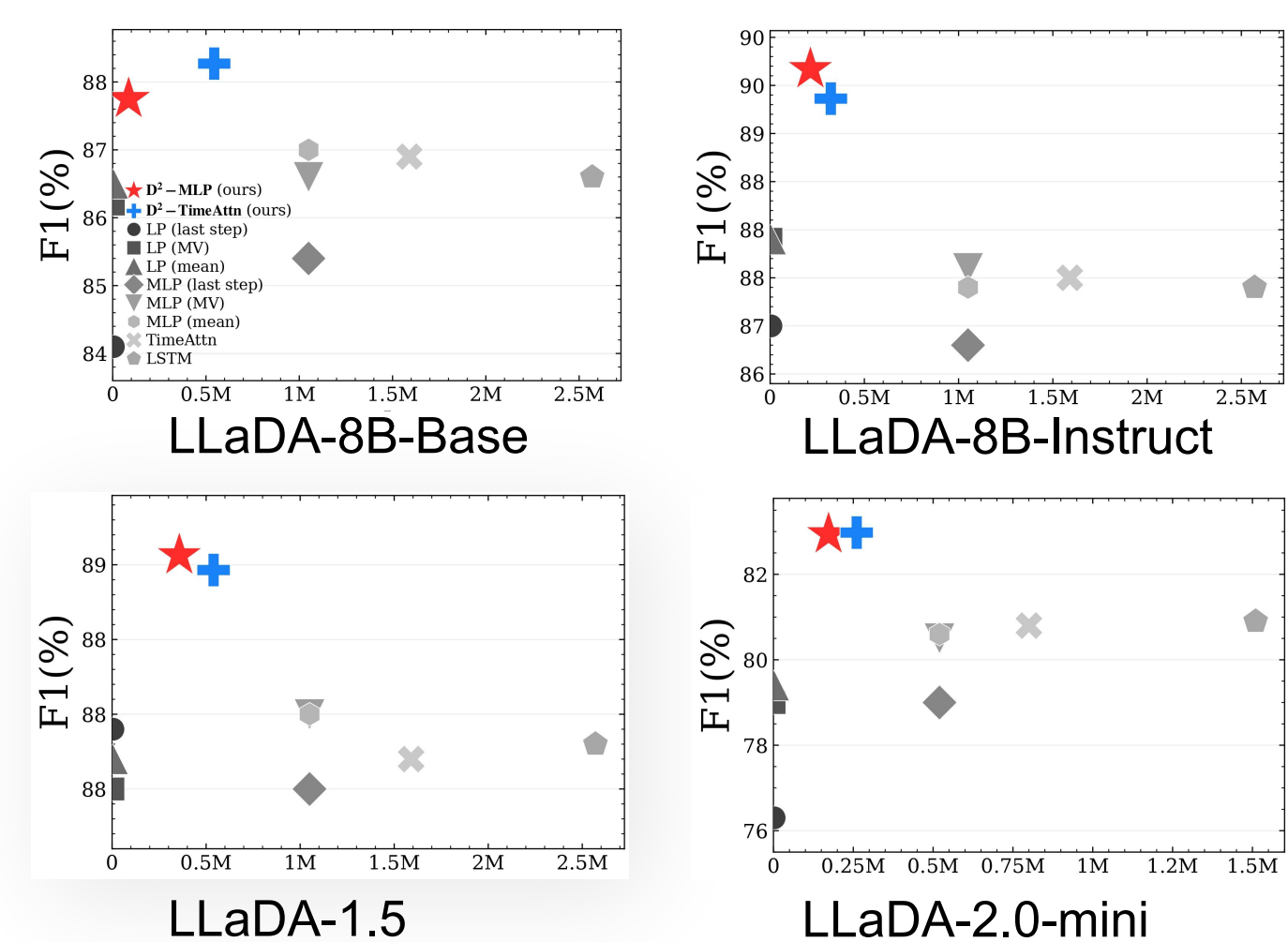
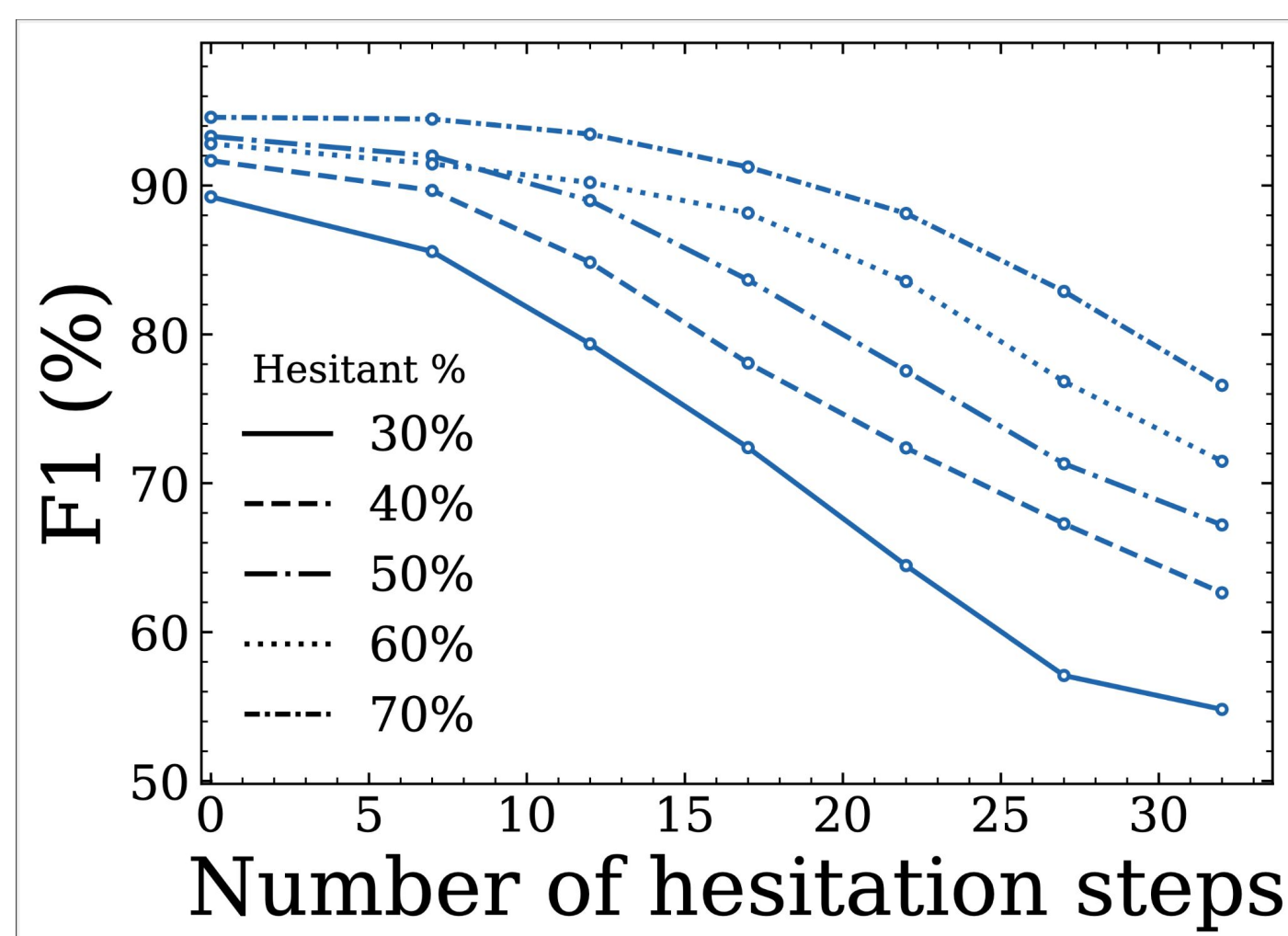
D²-Monitor predicts when safety probes will fail using "hesitation steps" and dynamically routes hard samples to a stronger classifier.

D²-Monitor: Dynamic Safety Monitoring for Diffusion LLMs via Hesitation-Aware Routing

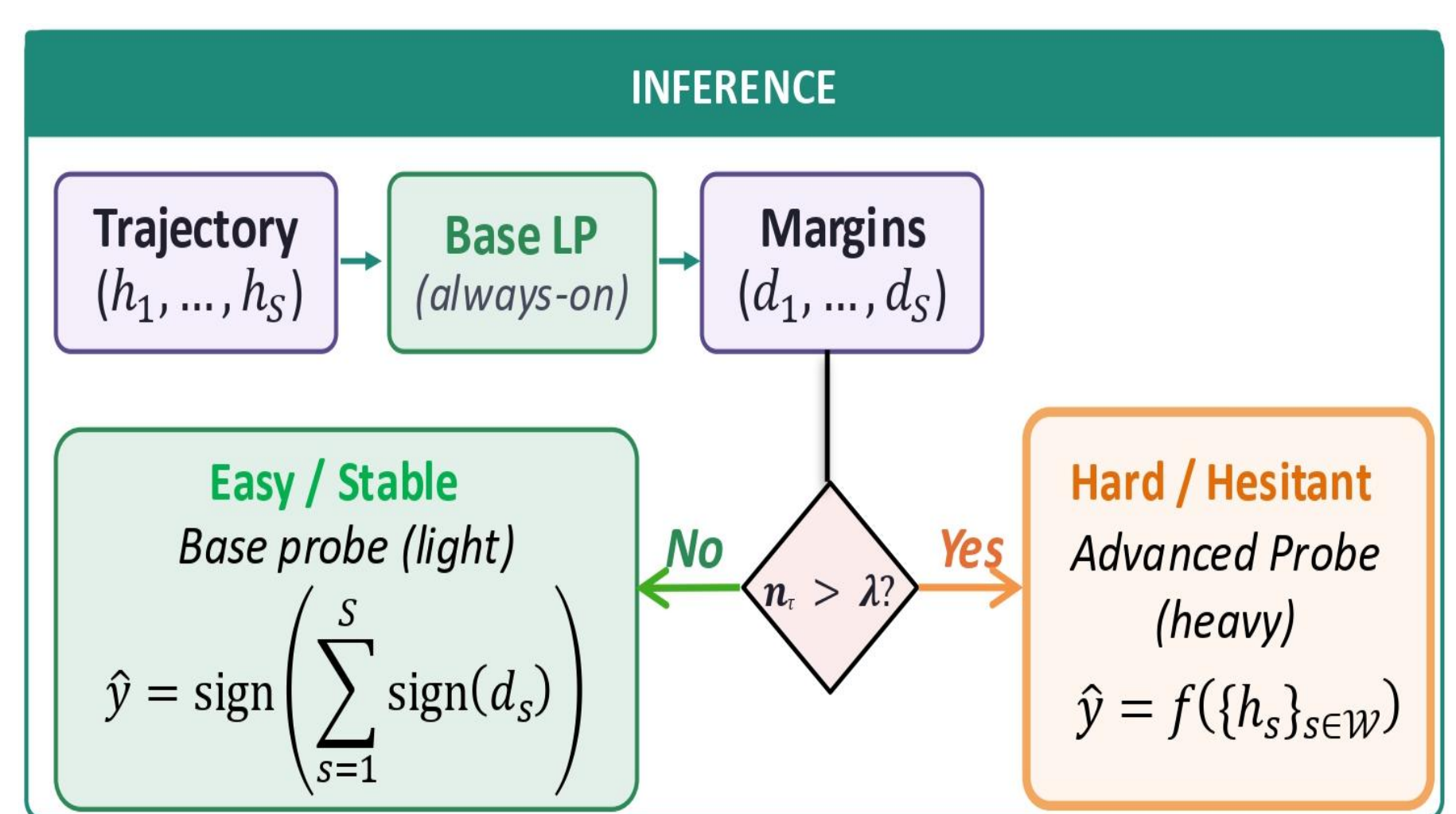
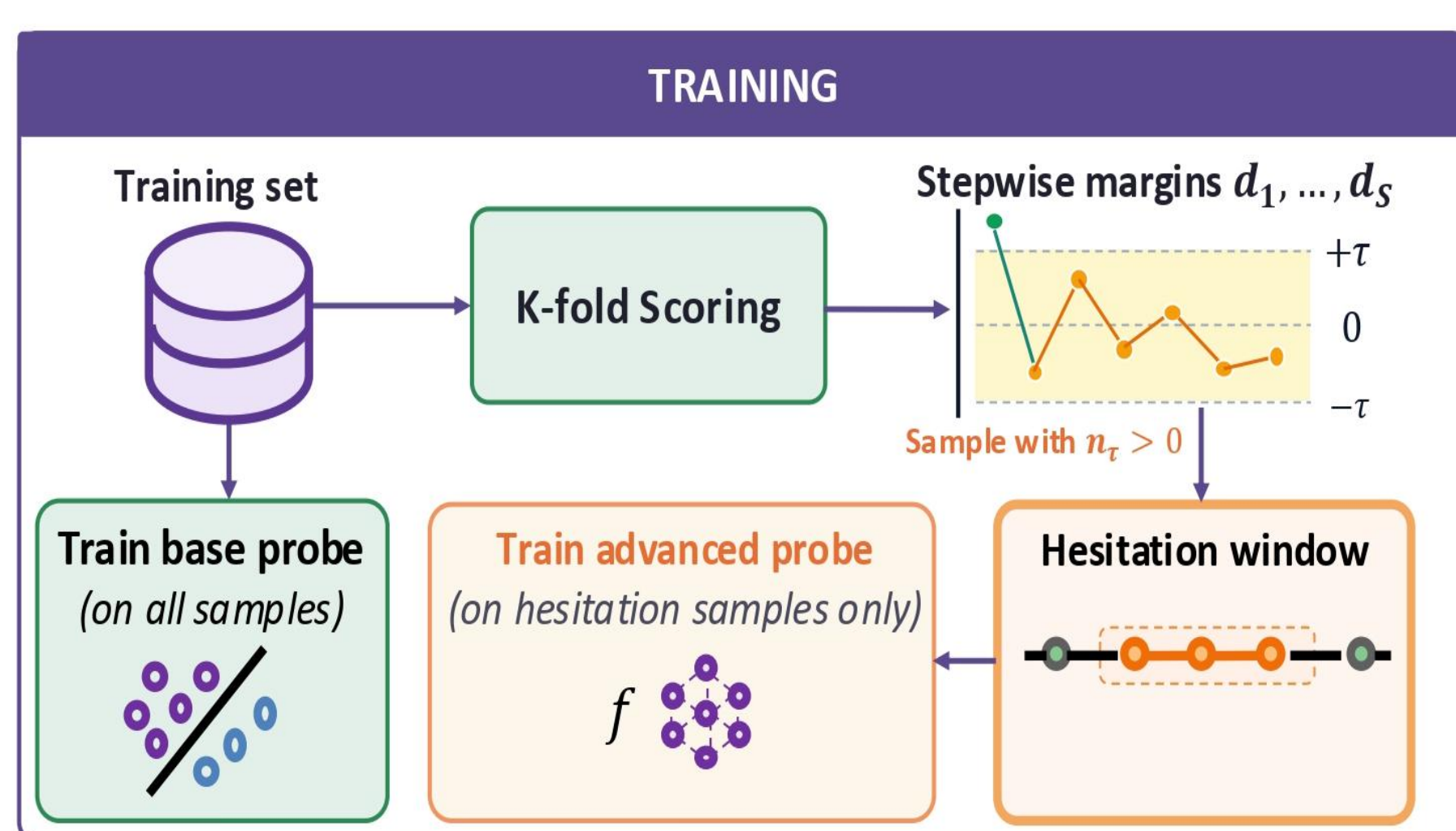
Background: Diffusion LLMs (D-LLMs) generate text through a multi-step denoising process, exposing intermediate hidden states unavailable in standard single-step monitoring. Yet safety monitoring for D-LLMs remains largely unexplored.

Finding: The number of hesitation steps predicts linear probe failure. As it increases, probe F1 drops monotonically.

Result: D²-Monitor achieves the best efficiency-effectiveness trade-off.



Methods



Limitation: Experiments are currently limited to D-LLMs with up to 16B parameters. Adaptive adversaries may suppress hesitation steps to evade the second-stage probe; complementary defenses are recommended.