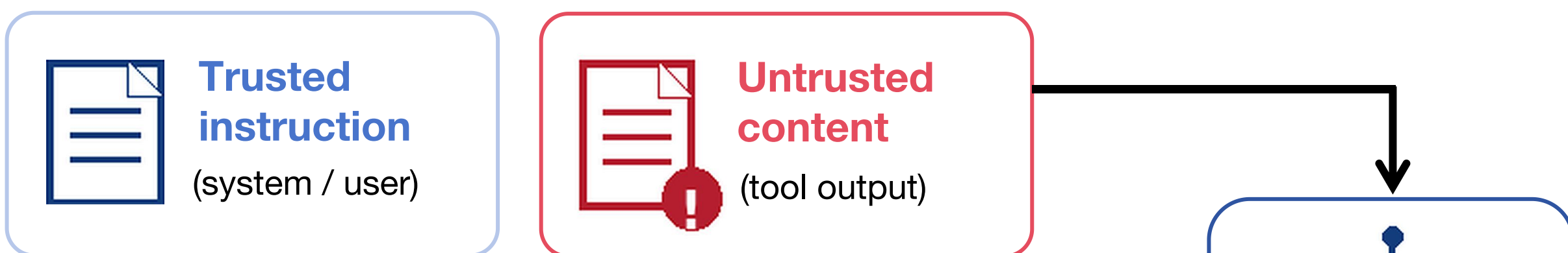


A. Threat channels

1) Tool output injection



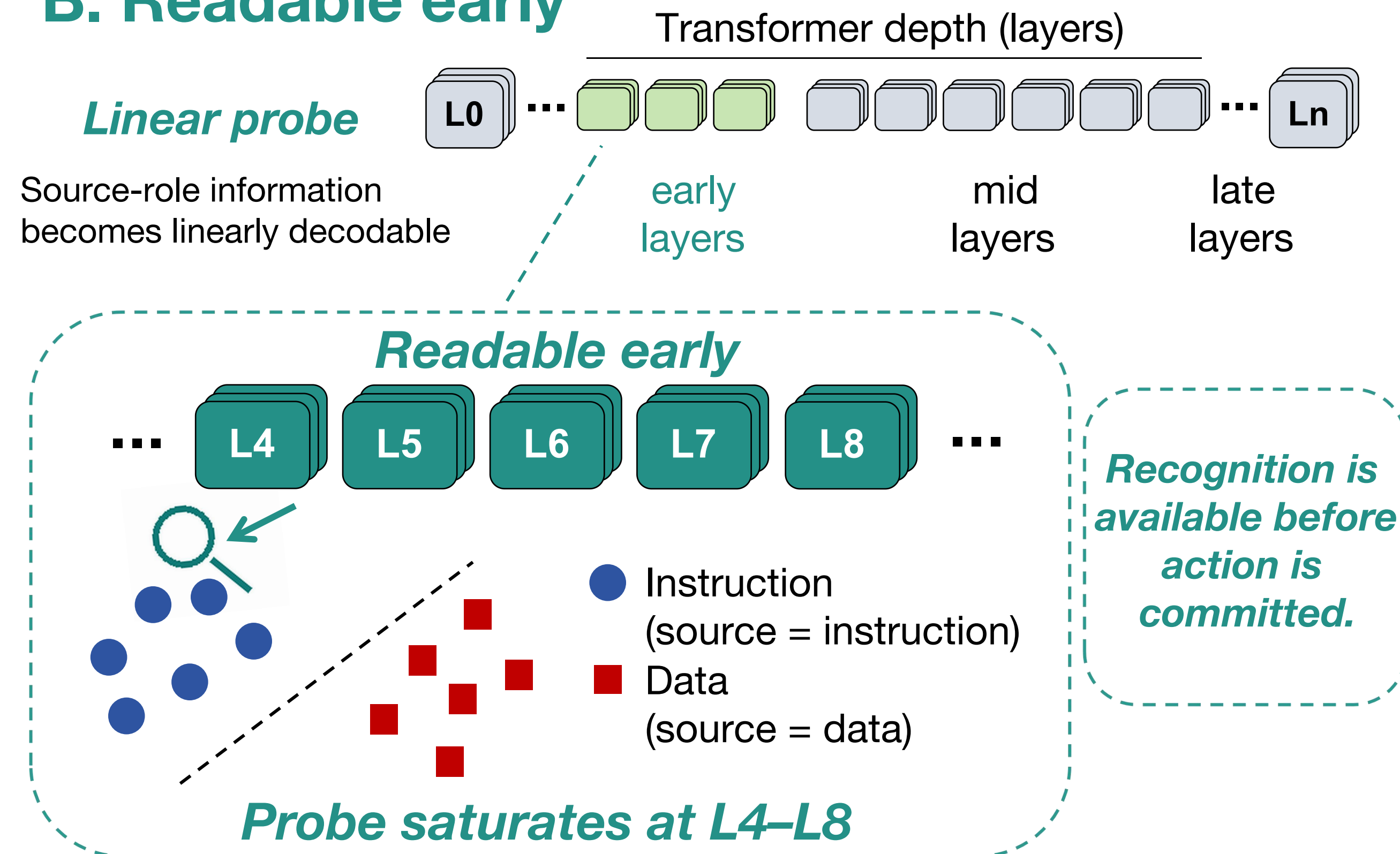
2) Slack / realistic multi-step trajectory injection



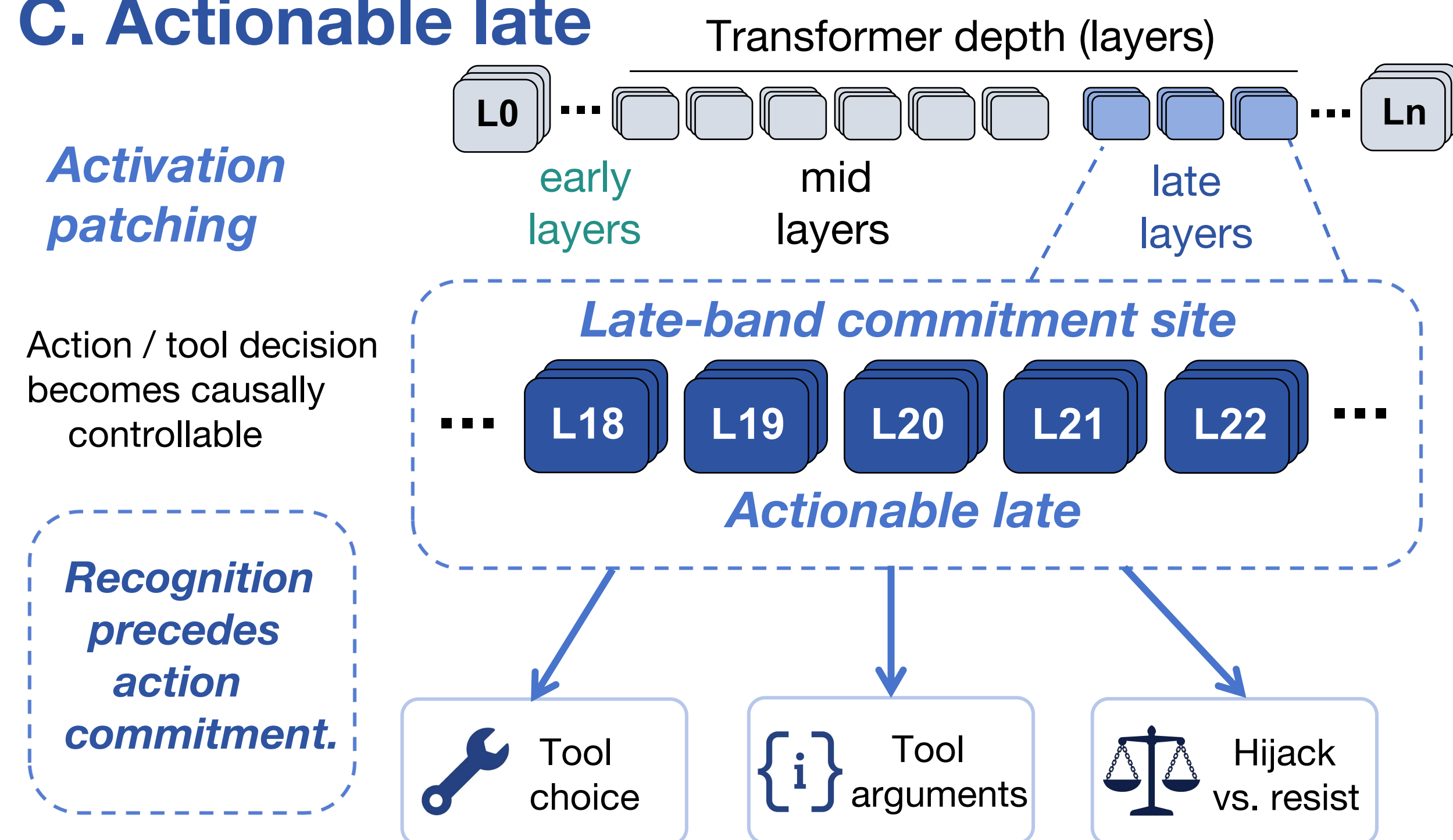
3) Memory poisoning



B. Readable early



C. Actionable late



D. Matched intervention

Direction d vs. channel (distribution)

	Controlled tool-output	AgentDojo-Slack	Memory poisoning	Obfuscated attacks
d_{tool}	✓	✗	✗	✗
d_{slack}	✗	✓	✗	✗
d_{memory}	✗	✗	✓	✗
d_{obf}	✗	✗	✗	✓

No universal safety subspace

Mitigation requires the right layer and the right channel/distribution-matched direction.

🗺️ **Knowing is early; acting is late; defending is channel- and distribution-specific.**