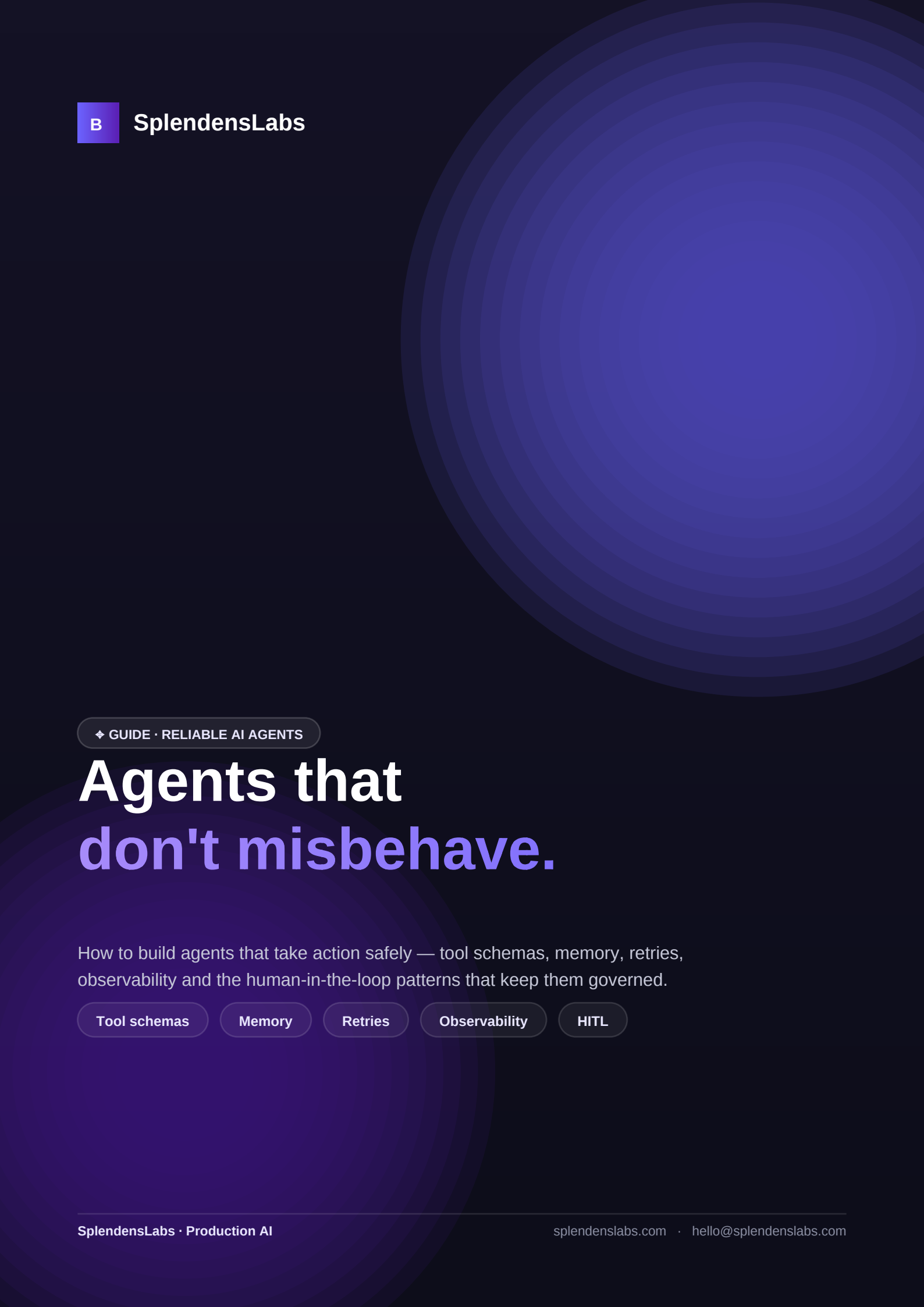




◆ GUIDE · RELIABLE AI AGENTS

Agents that don't misbehave.

How to build agents that take action safely — tool schemas, memory, retries, observability and the human-in-the-loop patterns that keep them governed.

Tool schemas

Memory

Retries

Observability

HITL

The anatomy of a reliable agent.

An agent is only as safe as its tools, its limits and its observability. Design all three before you grant autonomy.



Typed tools

Strict schemas and validation on every tool the agent can call.



Memory

Short-term scratchpad plus durable long-term stores.



Retries & recovery

Idempotent actions, retries and replay/resume on failure.



Budgets

Hard token, step and time budgets to bound any run.



Human-in-the-loop

Approval gates on consequential actions; easy escalation.



Observability

Trace every step; log decisions and tool calls.

Earn autonomy step by step.

1

Shadow mode

Agent proposes, humans act. Compare and log disagreements.

2

Co-pilot

Agent acts with human approval on every consequential call.

3

Bounded autonomy

Agent acts within policy; humans handle exceptions only.

4

Continuous evaluation

Keep evals running; catch drift before users do.