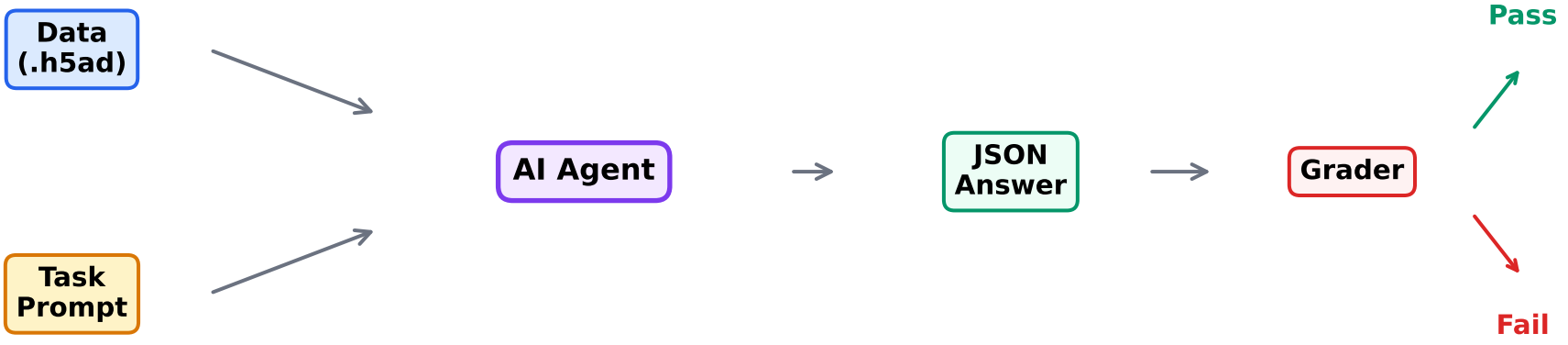

Each evaluation: data snapshot + natural language prompt + deterministic grader