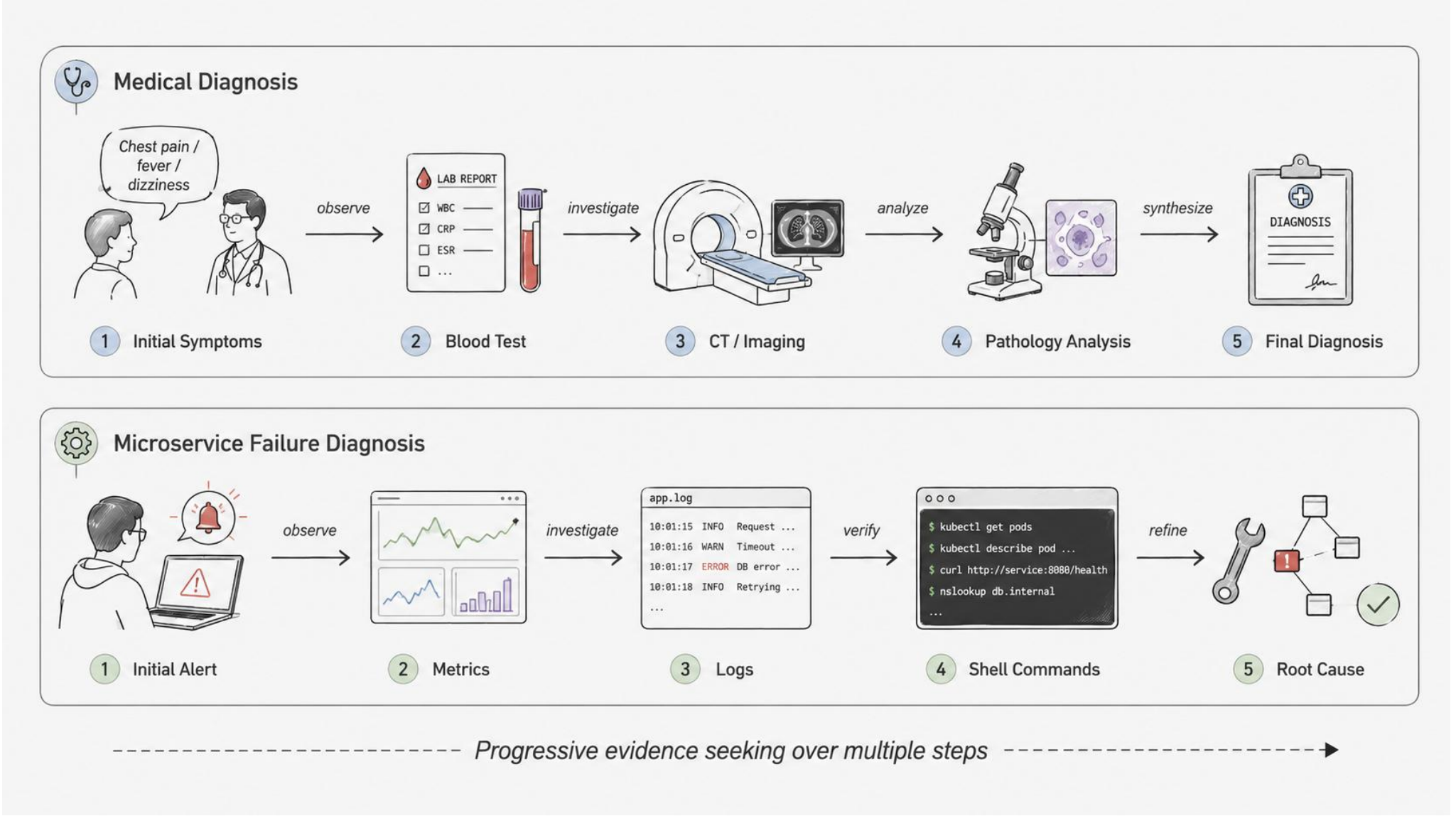
Graph of States

Solving Abductive Tasks with Large Language Models

ICML 2026 | Yu Luo et al.

Abductive Tasks are Complex Long-horizon Reasoning Tasks



General Reasoning Framework are Prone to Four Failure Modes

01. Evidence Fabrication

The model fabricates non-existent facts or spurious evidence within the reasoning chain to support its conclusion, thereby undermining the factual grounding required for abductive reasoning.

02. Context Drift

During long-horizon reasoning, the model gradually forgets the initial premises or critical contextual information, causing subsequent reasoning to drift away from the core problem.

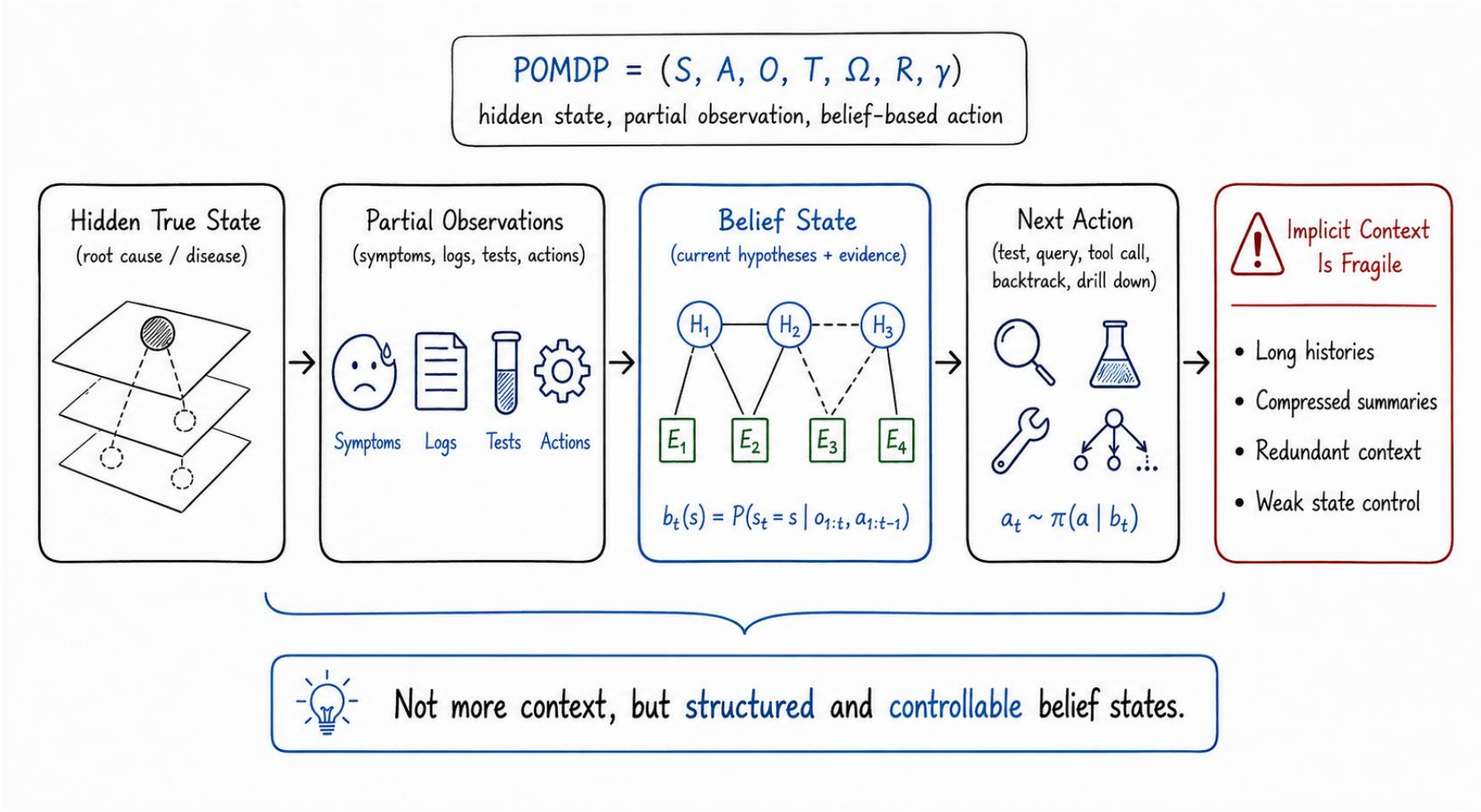
03. Failed Backtracking

When newly acquired evidence clearly contradicts the current reasoning path, the model lacks the ability to actively backtrack and revise its reasoning direction.

04. Early Stopping

The model prematurely settles on superficial, coarse-grained symptom descriptions, failing to drill down to the actionable root cause.

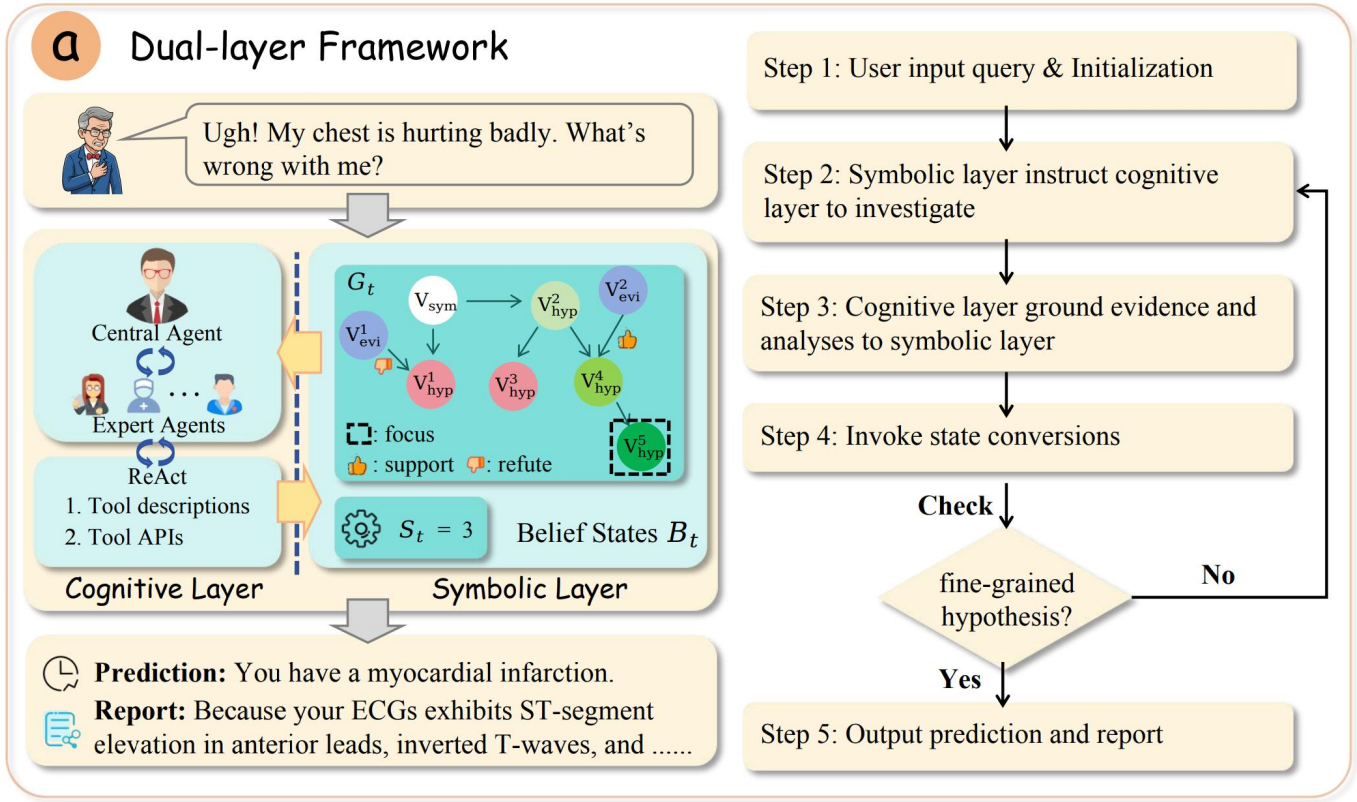
Belief States are Necessary



- Long-horizon LLM reasoning can be viewed as a POMDP;
- LLM agents must maintain belief states to decide what evidence to seek, when to revise hypotheses, and when to drill down.
- Unstructured context history is a fragile substitute for belief state modeling.

GoS: Overview

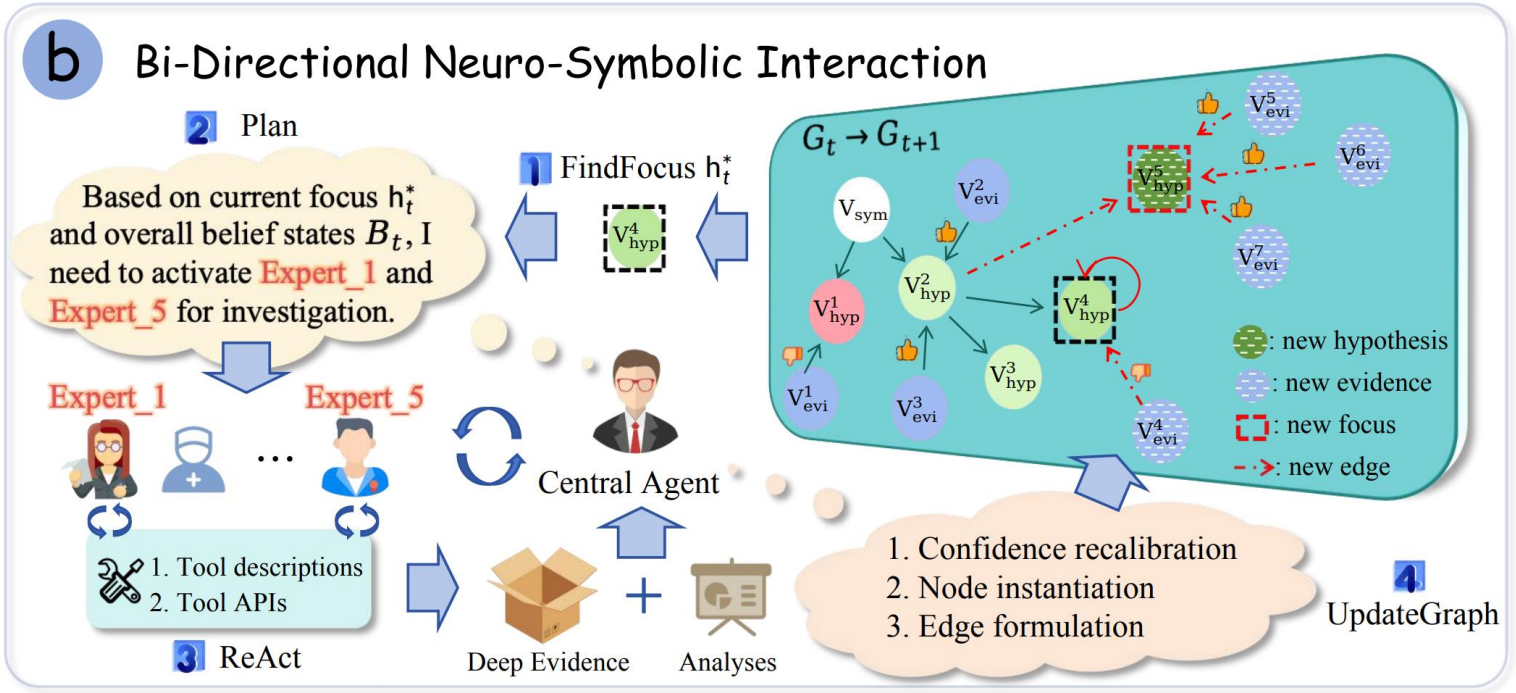
- GoS explicitly represents belief states for abductive reasoning.
- **Cognitive Layer:** executes domain-specific investigation through role-based agents.
- **Symbolic Layer:** maintains and controls reasoning states through a causal graph and a state machine.



Causal Graph → structured memory → mitigates 01. Evidence Fabrication & 02. Context Drift

State Machine → transition control → mitigates 03. Failed Backtracking & 04. Early Stopping

Bi-Directional Neuro-Symbolic Interaction



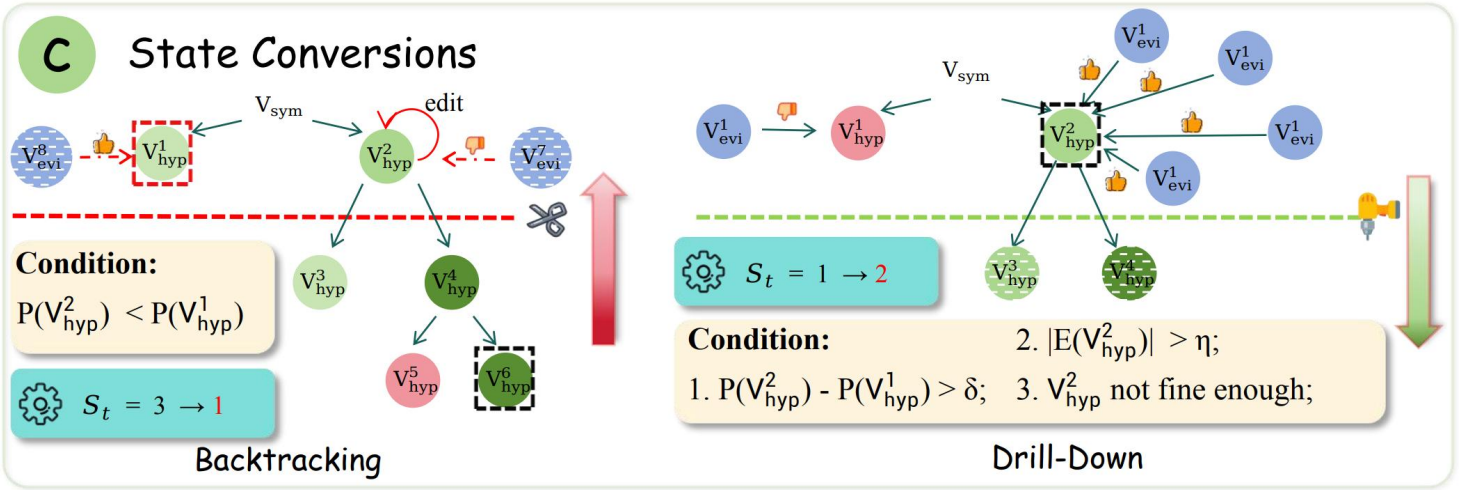
Symbolic → Cognitive

- Belief state identifies the reasoning focus.
- The focus is transformed into executable investigation plans.
- Expert agents perform targeted tool use.

Cognitive → Symbolic

- New observations are grounded into the graph.
- Hypothesis confidence is recalibrated.
- New evidence, hypotheses, and edges are instantiated.

State Conversions: Backtracking & Drill-Down



Backtracking

- When new evidence refutes an ancestor hypothesis,
- GoS returns to the earlier level and prunes invalid descendants.

Drill-Down

- When a hypothesis is sufficiently supported and still coarse-grained,
- GoS refines it into more specific sub-hypotheses.

Experiments: Performance Evaluation

Table 1. Performance of Medical Diagnosis (%)

Methods	LLM-as-a-Judge *Match*	LLM-as-a-Judge *Relevant*	Human-as-a-Judge *Match*	Human-as-a-Judge *Relevant*	$/case
GoS	**31.88**	**74.64**	**39.86**	**78.99**	0.12
Single/CoT	21.01	47.83	24.64	48.55	**0.03**
Single/ToT	18.84	47.10	19.57	45.65	0.08
Single/GoT	21.01	50.00	22.46	52.90	0.07
Single/FoT	21.74	58.70	21.01	61.59	0.32
Multi/CoT	21.01	49.28	23.19	50.72	0.07
Multi/ToT	20.29	50.72	23.91	53.62	0.17
Multi/GoT	21.74	52.17	23.91	55.07	0.15
Multi/FoT	23.19	63.04	26.09	65.94	0.73

Table 3. Performance of Failure Diagnosis in Distributed Systems (%)

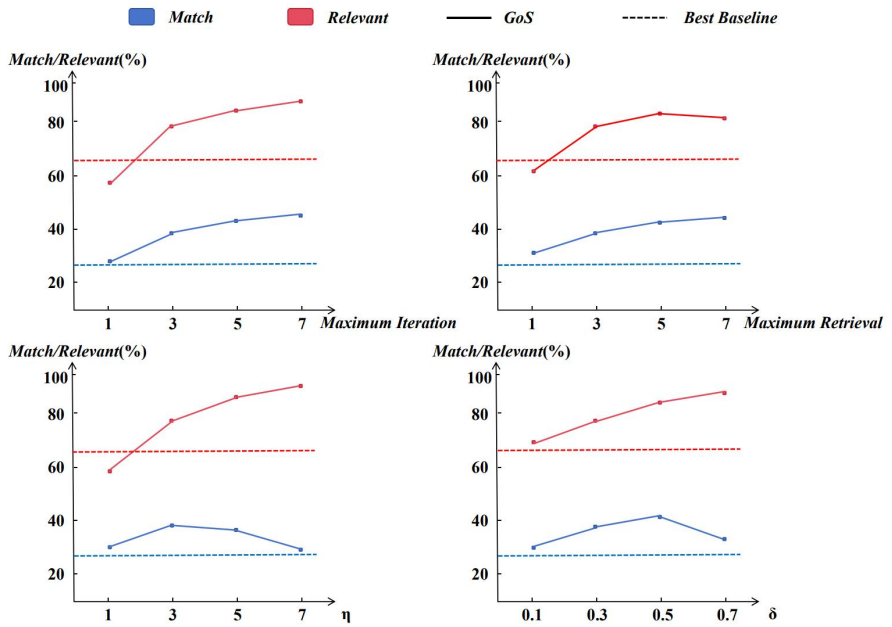
Methods	LLM-as-a-Judge *Match*	LLM-as-a-Judge *Relevant*	$/case
GoS	**70.67**	**88.00**	0.10
Single/CoT	26.67	81.33	**0.03**
Single/ToT	25.33	78.00	0.14
Single/GoT	27.33	80.00	0.11
Single/FoT	28.67	84.00	0.45
Multi/CoT	34.00	82.67	0.05
Multi/ToT	25.33	81.33	0.13
Multi/GoT	28.00	80.67	0.18
Multi/FoT	28.00	86.67	0.94

GoS achieves the best performance on both medical diagnosis and real-world failure diagnosis, with substantially higher fine-grained Match while maintaining competitive cost.

Experiments: Ablations & Sensitivity Analysis

Table 2. Ablation Study of Medical Diagnosis (%)

Methods	LLM-as-a-Judge *Match*	LLM-as-a-Judge *Relevant*	$/case
GoS	**31.88**	**74.64**	**0.12**
w/o reasoning focus	19.57	67.39	0.14
w/ structured state management	18.12	59.42	0.15
w/o causal graph	12.32	48.55	**0.12**
w/o state machine	12.32	50.00	0.17



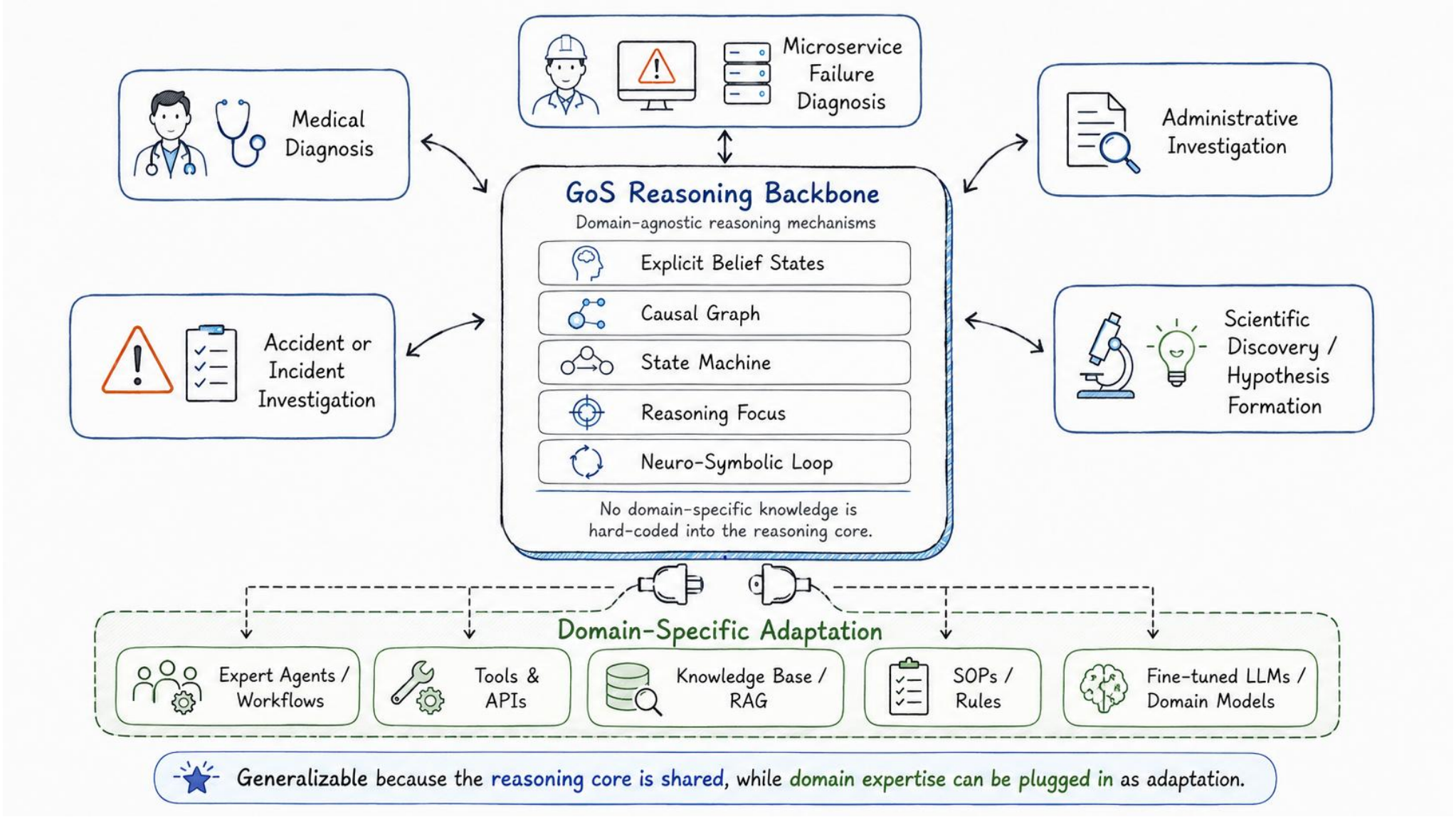
Key Findings

1. All core components matter.
Removing reasoning focus, causal graph, or state machine consistently degrades performance.

2. Explicit graph structure is essential.
Structured text management helps, but remains far below causal graph modeling.

3. Thresholds act as control knobs.
Evidence and confidence thresholds balance exploration depth and decision conservatism.

What Makes GoS generalizable across different Abductive Tasks?



Thanks !