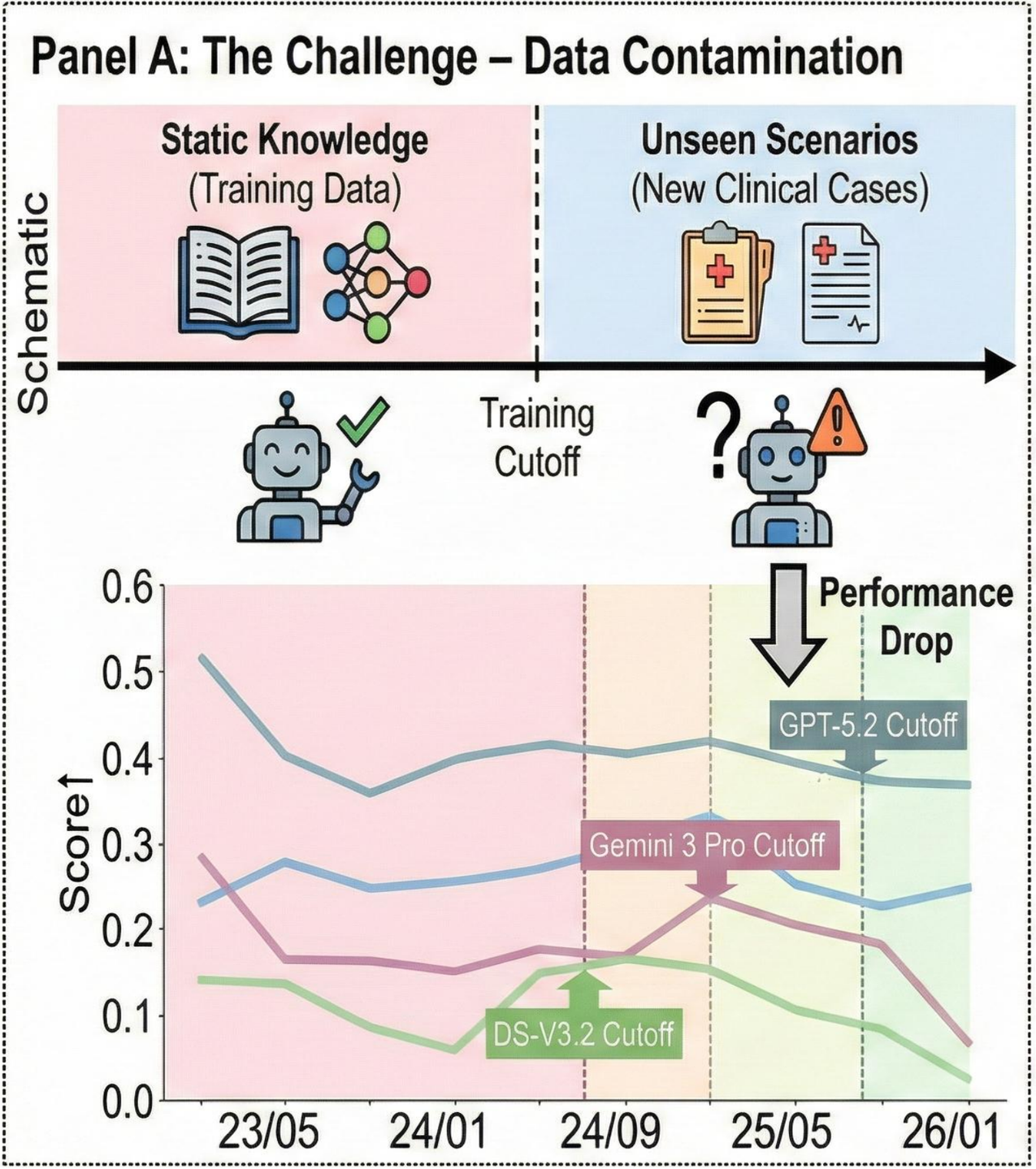
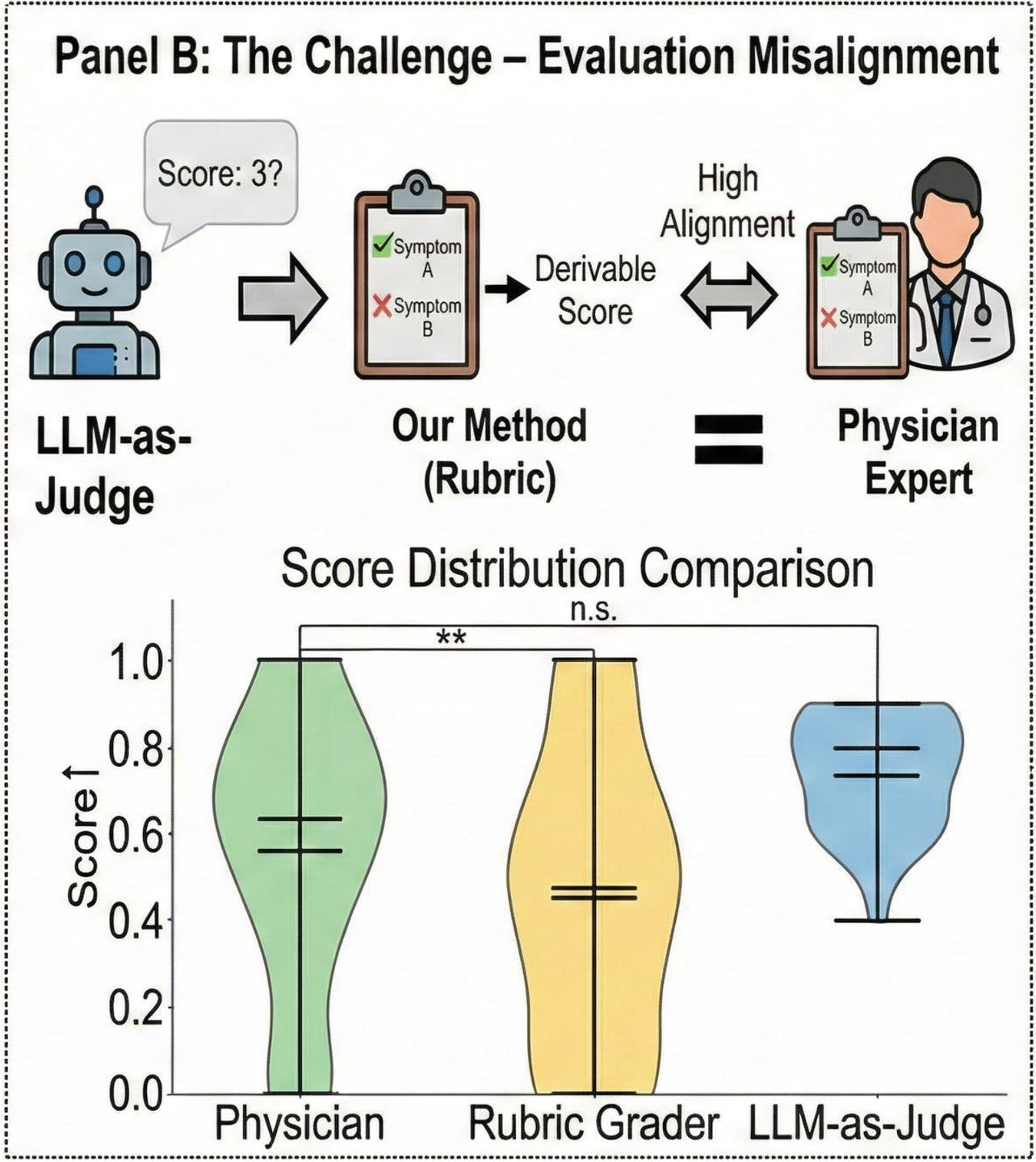

Panel A: The Challenge – Data Contamination
Schematic
Static Knowledge
(Training Data)
Unseen Scenarios
(New Clinical Cases)
Training
Cutoff
Performance
Drop
GPT-5.2 Cutoff
Gemini 3 Pro Cutoff
DS-V3.2 Cutoff
Score↑
0.0
0.1
0.2
0.3
0.4
0.5
0.6
23/05
24/01
24/09
25/05
26/01


Panel B: The Challenge – Evaluation Misalignment
Score: 3?
Symptom A
Symptom B
Derivable
Score
High
Alignment
LLM-as-
Judge
Our Method
(Rubric)
=
Physician
Expert
Score Distribution Comparison
n.s.
**
Score↑
0.0
0.2
0.4
0.6
0.8
1.0
Physician
Rubric Grader
LLM-as-Judge