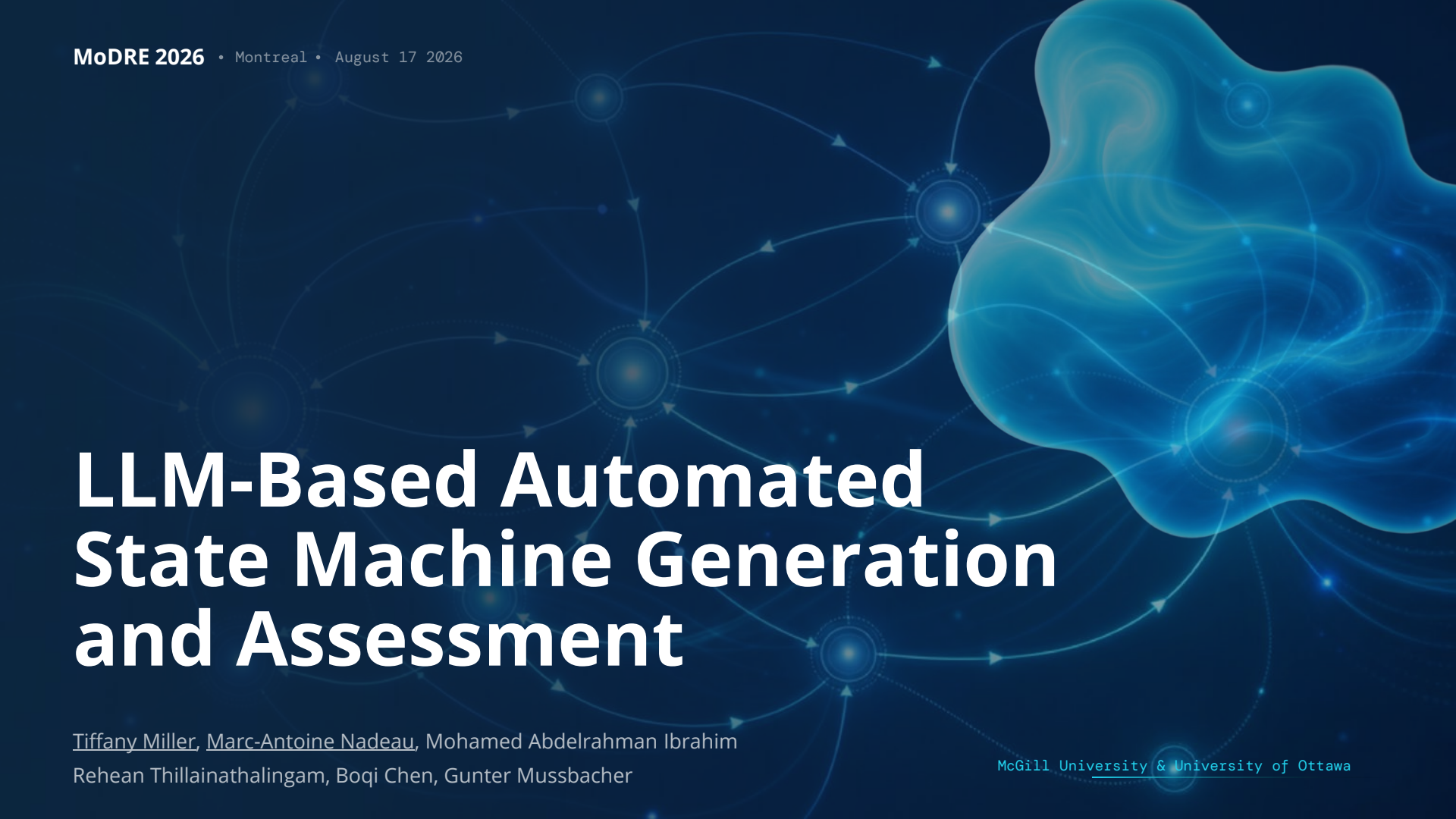
MoDRE 2026 • Montreal • August 17 2026



LLM-Based Automated State Machine Generation and Assessment

Tiffany Miller, Marc-Antoine Nadeau, Mohamed Abdelrahman Ibrahim
Rehean Thillainathalingam, Boqi Chen, Gunter Mussbacher

McGill University & University of Ottawa

Research Questions & Contributions

RQ1

Generation quality?

RQ3

Assessment trends across generators

RQ2

Automated assessment quality?

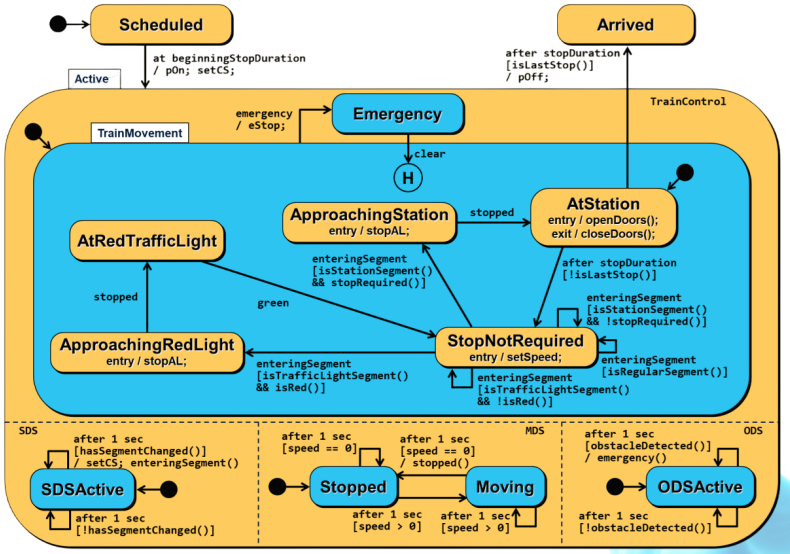
Contributions

- **Generation framework**
- **Assessment framework**
- **Benchmark evaluation**

The Problem

Manual UML state machine modeling and their assessment is time-consuming

Automated generation has been explored but automated assessment has barely been studied at all.



The Approach

01

Single-Stage Baseline

3-shot prompting, single LLM call

02

Two-Stage Refinement

3-shot or 6-shot prompting

Stage 1: Generate draft state machine

Stage 2: Review and fix omissions/errors

03

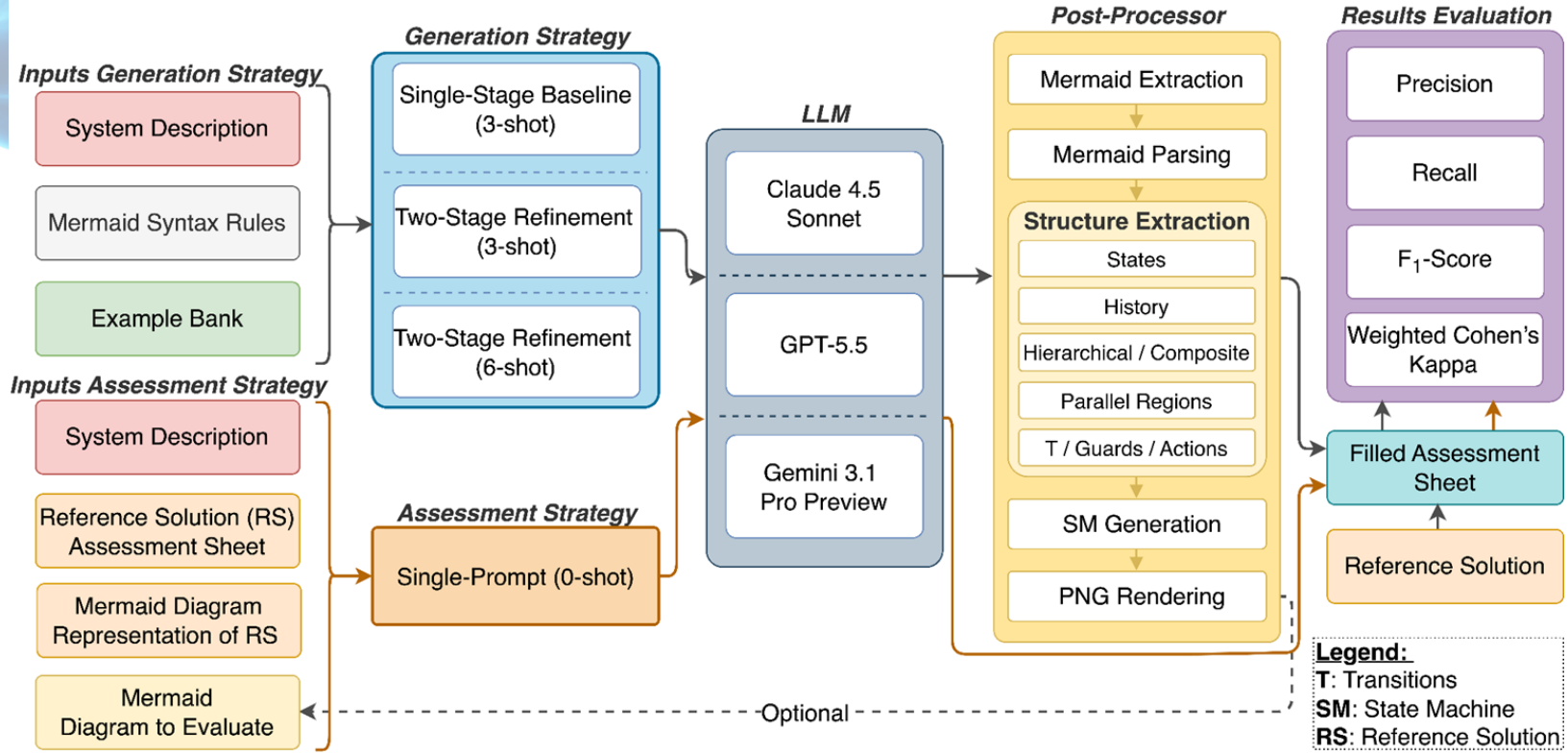
Assessment Framework

Input: System description + reference rubric + reference state machine

Output: Completed assessment sheet

The Approach

LLM pipeline takes plain-language requirements, generates a UML state machine, and evaluates its correctness.



Dataset

Summary of state machine model elements in ground-truth solutions:

Model Element	P	TA	SM	D	C	B	TM	W	S	Total
Action	3	12	0	7	8	10	6	36	21	103
Composite State	2	2	3	2	3	3	1	5	1	22
Guard	6	14	4	4	4	4	7	3	11	57
History State	1	1	1	1	1	1	1	1	1	9
Parallel Region	0	4	5	2	2	0	0	2	0	15
States	8	17	16	12	13	12	11	24	9	122
Transition	17	20	15	19	14	18	17	37	23	180
Total	**37**	**70**	**44**	**47**	**45**	**48**	**43**	**108**	**66**	**508**

P: Printer
TA: Train Automation System
SM: Spa Manager
D: Dishwasher
C: Chess Clock
B: Bread Maker
TM: Thermomix (Kitchen appliance)
W: W-UMPLE (Watch)
S: SSC7 (Self-Service Checkout)

Mermaid Model

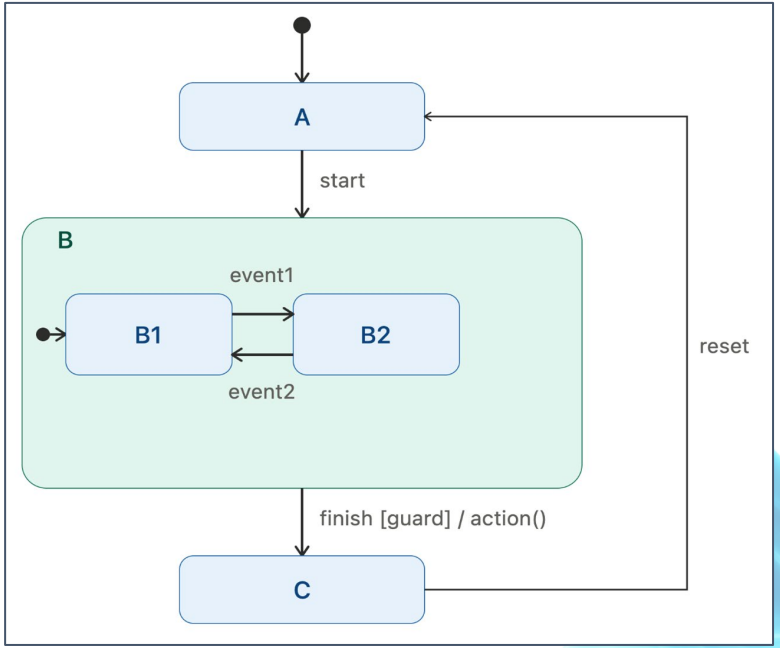
Custom Mermaid Syntax → Generated State Machine

```
1.  stateDiagram-v2
2.     state A
3.     state C

4.     [*] --> A

5.     state B {
6.         state B1
7.         state B2
8.         [*] --> B1
9.         B1 --> B2 : event1
10.        B2 --> B1 : event2
11.     }
12.
13.     A --> B : start
14.     B --> C : finish [guard] / action()
15.     C --> A : reset
```

Example Mermaid Code



State Machine Generated from Mermaid Code

Claude 4.5 Sonnet Results

Overall Precision, Single-Stage → Two-Stage (6-shot) Refinement

0.991 → 0.956

Overall Recall, Single-Stage → Two-Stage (6-shot) Refinement

0.674 → **0.846**

Overall F1-Score, Single-Stage → Two-Stage (6-shot) Refinement

0.775 → **0.898**

One extra refinement pass closes most of the gap — without giving up precision.

Across all 9 benchmark examples, the 2-stage (6-shot) approach produced **only 2 false positives** total.

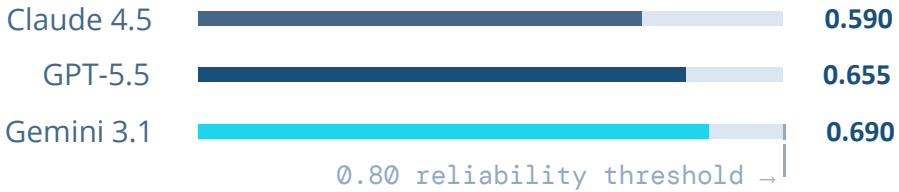
Assessment Isn't Solved Yet

Weighted Cohen's Kappa (K) to determine agreement

Best agreement with human raters

κ = 0.69

No tested LLM reaches the **0.80 threshold** needed to trust automated grading on its own



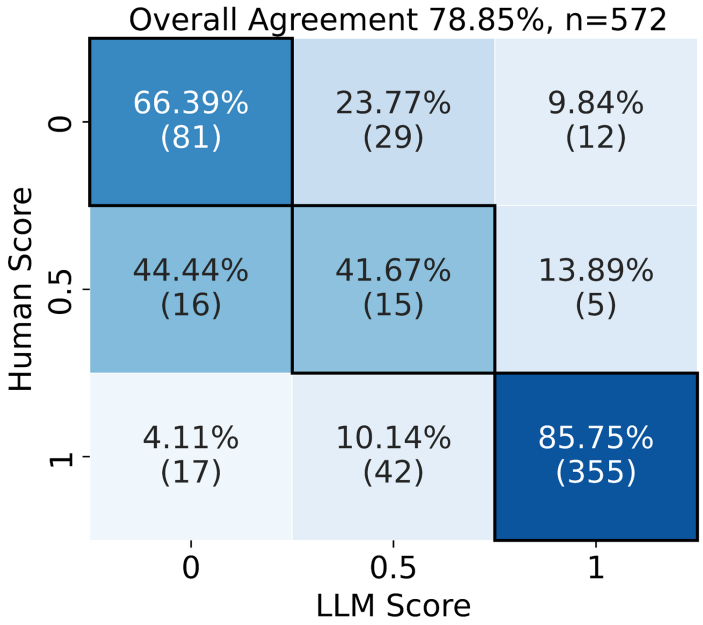
Trustworthy Assessment Scores

How often do human and LLM scores match?

Row-normalized across the three score options (0, 0.5, 1).

Rows = human scores; **Columns** = GPT-5.5 scores.

GPT-5.5 shows a **systematic conservative bias**, driven primarily by **collapsing partial credit** (0.5) to incorrect (0).



Confusion Matrix: GPT-5.5 vs. Human

The Recommendation

GENERATION

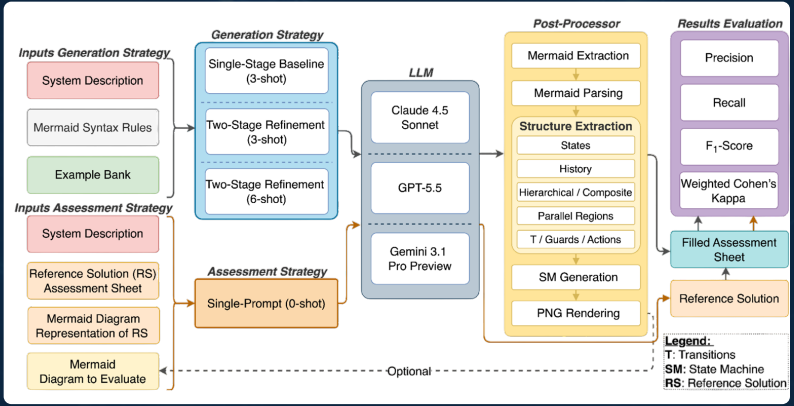
What can be trusted from the generated state machine

- **Precision is consistently very high** (> 0.91) — elements that are included almost always correct
- **Only 2 false positives** across all nine examples with two-stage prompting
- States, transitions, composite states, and regions are reliably generated
- Use two-stage prompting to recover missed elements, especially actions and guards

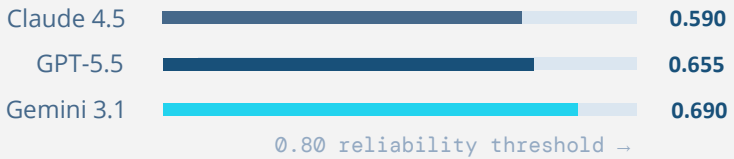
ASSESSMENT

What can be trusted from automated LLM assessment

- **Score 1** assignments are **trustworthy** — LLMs are strict with giving credit
- Regions and composite states reach strong automated agreement
- Actions, guards, transitions, and states require human review — **partial credit collapse** is most severe here
- The **bias is conservative**: LLMs underestimate quality rather than overestimate it



Assessment Results



Generation Results

Overall F1-Score, Single-Stage → Two-Stage Refinement

0.775 → 0.898

Experiment Data