HOKKAIDO UNIVERSITY

Neuro-Symbolic Instruction Tuning for Explainable Mahjong Agents via Two-Stage Dual-LoRA

ICPR 2026 / ORAL PRESENTATION

Lyon, France – 17 August 2026

Zhaohao Fang[1,*], Junhuai Xu[2], Jiawei Yu[2], Hanjie Li[3], Shuotian Chen[2], Jiyi Li[1], Masaharu Yoshioka[1]

[1]Hokkaido University [2]Tsinghua University [3]Xidian University

*Presenter

Why Explainable Mahjong AI?

AI success in imperfect-information games
Systems such as Suphx, NAGA, and MAKA achieve top-level results on platforms such as Tenhou.

Yet their decisions remain opaque
Most agents expose only discard probabilities or value estimates, not the strategic rationale behind them.

Explanations create educational value
Natural-language justifications can turn a strong Mahjong agent into a coach for human players.

NAGA provides numerical analysis, but no strategic explanation.

Two Challenges for LLM Agents

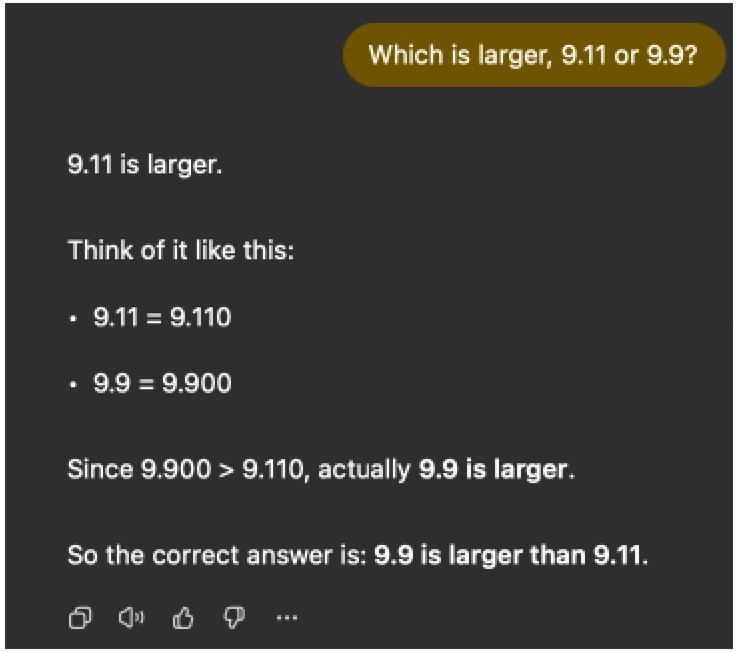
- **Potential of LLM approaches**: LLMs are well suited to natural-language explanations, but Mahjong exposes two critical limitations.

1. Logical hallucination

LLMs can fail at **exact calculation** and **strict rule application**, such as miscalculating Shanten or Ukeire, or misidentifying Genbutsu.

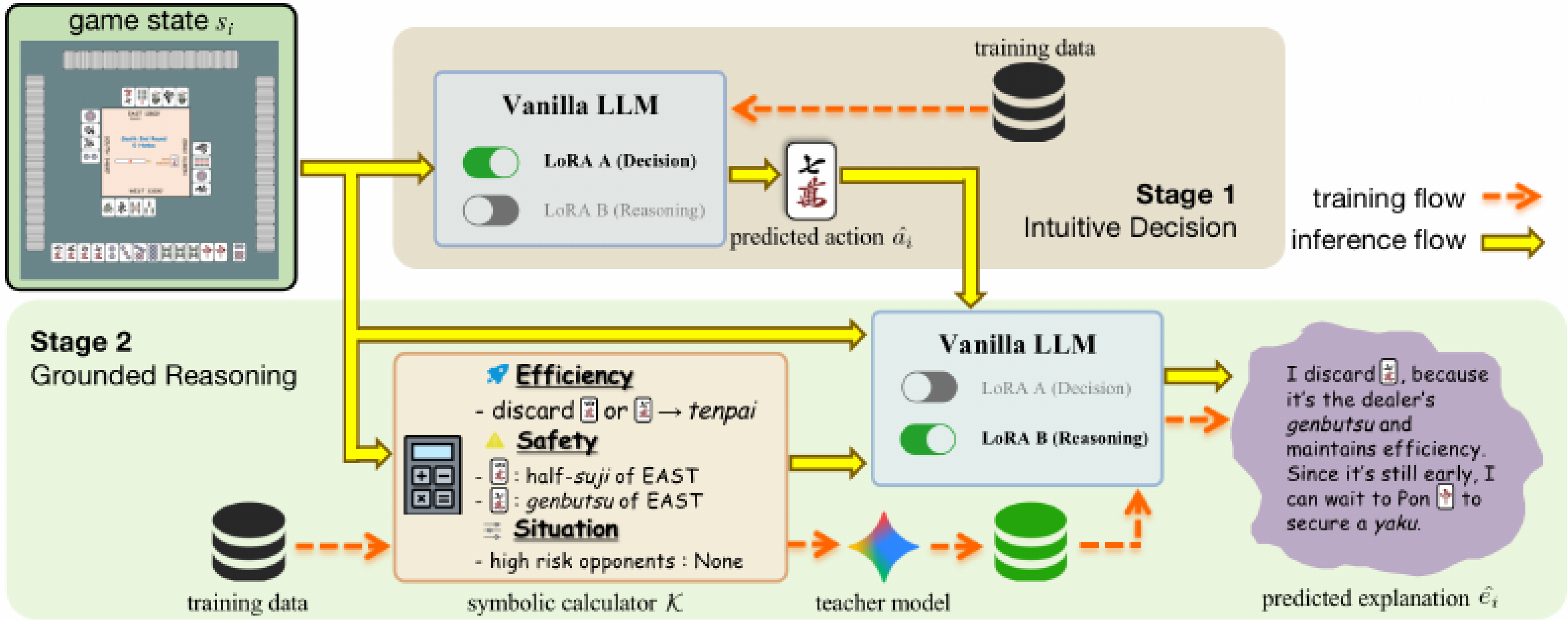
2. Catastrophic forgetting

Fine-tuning for explanations can degrade playing strength. Decision making and reasoning may interfere when optimized in a single adapter.



A fluent answer can still violate an exact numerical relation.

Method Overview



1. Intuitive decision

LoRA-A predicts the discard.

2. Symbolic calculation

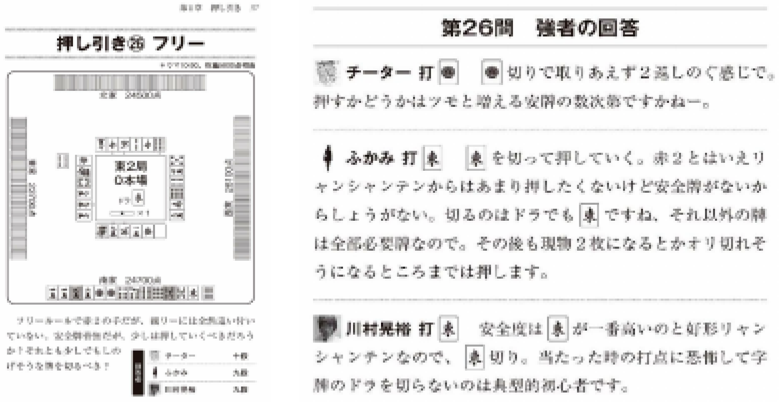
Compute objective game metrics.

3. Grounded explanation

LoRA-B explains the decision.

Task and Learning Objective

- **Task definition**
  - The Mahjong discard problem is modeled as a **partially observable Markov decision process (POMDP)**.
  - Input s: observable state (hand, discards, scores, melds, etc.); opponents' hands remain hidden.
  - Output: action a (discard) and explanation e (rationale).



Nanikiru: choose one discard from the visible board state.

Expert alignment objective

$$\mathcal{D} = \{(s_i, a_i, e_i)\}_{i=1}^{N} \qquad \max_{\theta} \sum_{(s,a,e)\in\mathcal{D}} \log P_\theta(a, e \mid s)$$

Align with both expert discards and their underlying thought processes.

Symbolic Calculator

A deterministic **symbolic calculator** grounds explanations in three dimensions: **efficiency**, **safety**, and **situation**.

Shanten $S(H)$

Minimum distance from hand H to a Tenpai hand:

$$S(H) = \min_{H' \in \mathcal{T}} \mathrm{dist}(H, H').$$

Visible Ukeire $U_{\mathrm{vis}}(H)$

Remaining observable tiles that reduce Shanten:

$$U_{\mathrm{vis}}(H) = \sum_{t \in \mathcal{T}_{\mathrm{imp}}} \max\big(0, 4 - N_{\mathrm{vis}}(t)\big).$$

Safety

Genbutsu is provably safe; Suji and tile exhaustion are heuristic risk indicators.

Situation

Threat estimates guide the balance between offensive efficiency and defense.

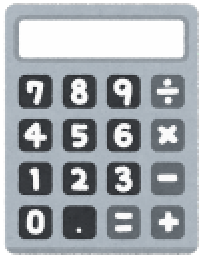
Neuro-Symbolic Distillation

- Game logs provide actions but no discard reasons; **Gemini-2.0-Flash** is used as the teacher model.
- Symbolic metrics $k = K(s)$ are injected into the prompt to prevent calculation errors and hallucinated safety claims.

Game state and expert action

EAST 2, 0 Honba, 4 tiles left
Hand: 4m 4m 5m 6m 7m 8m 9m 2p 8p 8p 9p
Candidate 9m: 4-Shanten, 29 Ukeire; dead tile; Genbutsu for P0/P2/P3.

Ground-truth discard: 9m



$K(s)$

$\longrightarrow$

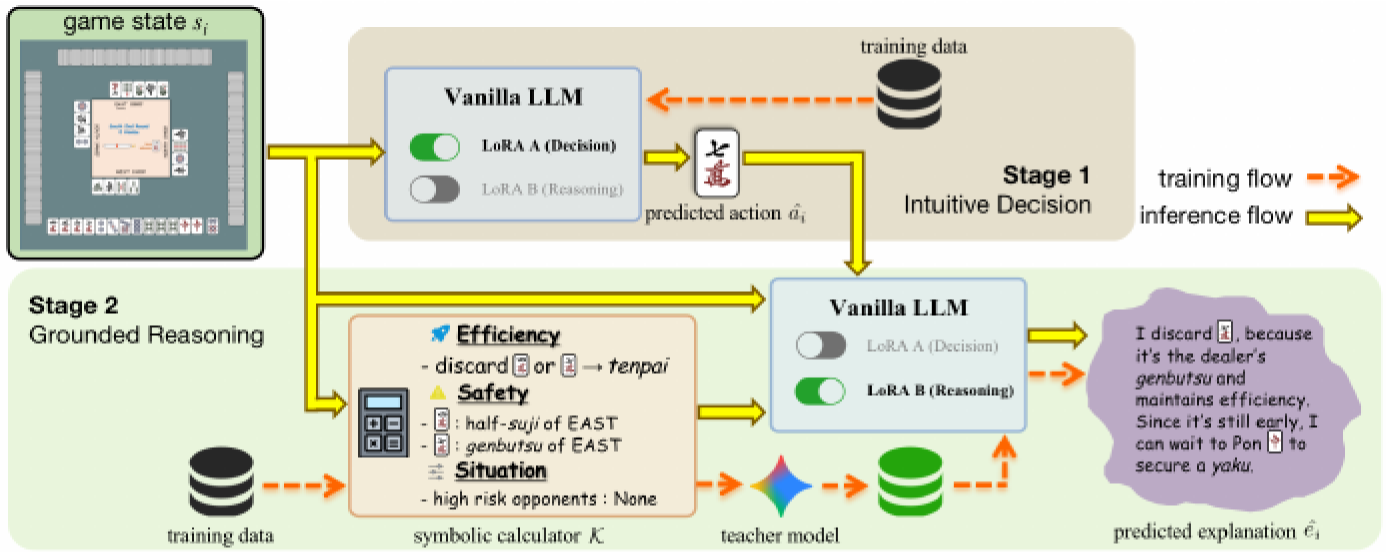
Teacher

Grounded explanation

In the late game, 9m is already exhausted and is safe against all opponents. Although other discards preserve more Ukeire, safety should be prioritized; therefore, discard 9m.

$$e \sim P_{\text{Teacher}}(e \mid s \oplus k \oplus a)$$

Two-Stage Dual-LoRA



Reusing one adapter across both stages causes interference and degrades discard accuracy.

1 Decision adapter (LoRA-A)

Predict expert discard a from state s, without free-form reasoning.

2 Reasoning adapter (LoRA-B)

Explain a using state s and symbolic knowledge k.

Inference: LoRA-A selects the discard; after $k = K(s)$ is computed, LoRA-B generates the explanation.

Data Source and Chronological Split

- **Source**: 2024 logs from the *Houou* (Phoenix) room on the online Mahjong platform Tenhou.
- High-level four-player South games with red Dora.

00:01 | 26 | 四鳳南喰赤－ | 牌譜 | 笹野理一(+55.8) 笹餅(−0.5) ミラーねじ(−21.4) ＠次男(−33.9)
00:01 | 14 | 三鳳南喰赤－ | 牌譜 | 雀極のへ(+50.3) R.PT(+13.8) 京都のリーチ超人(−64.1)
00:02 | 15 | 四鳳南喰赤－ | 牌譜 | はるか！ 春を呼ぶ(+59.2) 東方定助(+17.2) こてる(−20.6) ささり(−55.8)
00:04 | 18 | 三鳳南喰赤－ | 牌譜 | らくた(+57.3) mtk(−12.9) ヨッシー9zde(−44.4)
00:05 | 30 | 四鳳南喰赤－ | 牌譜 | makit123(+76.6) 宮保鶏肉(−2.9) bakura08(−31.8) 御影ゆずき(−41.9)
00:05 | 33 | 四鳳南喰赤－ | 牌譜 | 琴音文乃(+57.8) モンキートリック(+16.0) オフショアガール(−25.0) tanpin8(−48.8)
00:06 | 30 | 四鳳南喰赤－ | 牌譜 | suzume3(+40.7) 湘北⑭三井寿(+7.8) 馬座標胤舞(−16.8) てんくう(−31.7)
00:07 | 16 | 三鳳南喰赤－ | 牌譜 | 椿なの嘆きないで(+44.1) アサ3(+2.9) 鳴き麻雀(−47.0)
00:08 | 32 | 四鳳南喰赤－ | 牌譜 | 憤怒鳥(+42.3) Aria(+8.3) うずめ改弐(−15.3) Halfs(−35.3)
00:09 | 26 | 三鳳南喰赤－ | 牌譜 | 誤作動(+39.8) 春風送福(−5.7) 闇書遥香(−34.1)
00:10 | 37 | 四鳳南喰赤－ | 牌譜 | 侍(+56.4) やま悠YYYY(+8.5) タマネギ(−21.4) Pbotking(−43.5)
00:11 | 16 | 三鳳南喰赤－ | 牌譜 | 異世界双子☆瀧(+35.4) 理論力学(−7.7) 葉山陽和(−27.7)
00:11 | 23 | 四鳳南喰赤－ | 牌譜 | イニエスタ(+50.7) えいきち@5mf(+14.5) koala(−19.8) ビターキャラメル(−45.4)
00:12 | 33 | 四鳳南喰赤－ | 牌譜 | 菅田将暉(偽物)(+63.7) tori0612(+0.9) ロロん(−23.5) 小宮たずき(−41.1)
00:14 | 26 | 四鳳南喰赤－ | 牌譜 | だっ(+64.5) Gagnant(+12.3) ジルコニア(−29.2) gokuaku(−47.6)
00:16 | 35 | 四鳳南喰赤－ | 牌譜 | スカポン(+57.7) まとた(+16.9) 隣混状態麻雀(−28.5) oym20NOV(−46.1)
00:17 | 21 | 四鳳南喰赤－ | 牌譜 | 挑戦者！(+47.6) はるか！ 春を呼ぶ(+4.5) 多摩ソージ(−20.1) ミスターぴーナツ(−32.0)
00:18 | 32 | 四鳳南喰赤－ | 牌譜 | フクロウサン(+43.9) ナカにし(+7.6) BBです。(−13.4) ARISA73(−38.1)
00:19 | 11 | 三鳳南喰赤－ | 牌譜 | R.PT(+45.7) 御中元(−7.2) 細吹雪(−38.5)
00:20 | 28 | 四鳳南喰赤－ | 牌譜 | wengding(+42.1) 眼科医(+4.0) saichanz(−17.8) ぐてぃ(−28.3)
00:21 | 14 | 四鳳東喰赤速 | 牌譜 | みみず(+42.3) ヴム９て(+4.8) 易利箇伯(−17.2) 【トリュフ】(−29.9)
00:21 | 17 | 三鳳南喰赤－ | 牌譜 | 新人麻酔科医(+51.4) マッスルカズ(+15.2) つるっぱげ(−66.6)
00:22 | 24 | 四鳳南喰赤－ | 牌譜 | ホンイツ赤３(+62.5) renji83(+11.9) うけながしません(−18.9) あおまめ(−55.5)
00:23 | 32 | 四鳳南喰赤－ | 牌譜 | きいな(+51.0) ささり(+3.2) ヤマダさん(−17.4) 盛岡のリーチ超人(−36.8)
00:24 | 17 | 三鳳南喰赤－ | 牌譜 | ジャストぴぃ(+57.9) 椿なの嘆きないで(−11.5) ガーネット(−46.4)
00:25 | 20 | 四鳳南喰赤－ | 牌譜 | 青学寺リア勢(+70.8) 偽善者高橋みなみ(+2.8) 敗北者ねる(−25.7) ドラゴンひでポン(−47.9)
00:27 | 30 | 四鳳南喰赤－ | 牌譜 | タイ人＠バンコク(+49.2) 成多嘉智(+11.9) ブレスブ(−21.0) 麻雀太郎(−40.1)
00:28 | 10 | 三鳳南喰赤－ | 牌譜 | 傍若無人51(+82.5) 葉山陽和(−19.3) shorty(−63.2)
00:28 | 25 | 四鳳南喰赤－ | 牌譜 | 東方定助(+41.5) ＠次男(+10.7) こくらカバ蔵(−20.2) あっき～♪(−32.0)

Chronological split

2024-01-01
Stage 2 training (neuro-symbolic distillation)

2024-12-20 to 2024-12-29
Stage 1 training

2024-12-30
Stage 1 validation

2024-12-31
Stage 1 and Stage 2 testing

Selecting Expert Decisions

1

Select high-quality matches

Keep **Houou / four-player South with red Dora** games in which every player is 7-Dan or higher.

2

Choose a reliable viewpoint player

- Consider only the top-2 finishers.
- Sample in proportion to final score.
- If one candidate deals in by Ron, use the other candidate to avoid an obvious misplay.

3

Keep only active decisions

Remove post-Riichi discards: they are automatic and therefore provide no meaningful decision signal.

Balancing Six Game Contexts

- **Extreme data imbalance**: raw logs are dominated by early-offense states and contain little late-defense data.

Solution 1: 2-axis, 6-class bucketing

Strategic context

- Offense: no opponent Riichi
- Defense: at least one opponent Riichi

Progression by own discards *d*

- Early: $0 \leq d \leq 5$
- Mid: $6 \leq d \leq 11$
- Late: $d \geq 12$

$2 \times 3 = 6$ independent buckets.

Solution 2: random under-sampling

Align every bucket to the smallest class.

- Early-offense: 247,614 samples
- Early-defense: 7,339 samples
- Imbalance: approximately $33\times$

Final Stage 1 training set

$$7{,}339 \times 6 = 44{,}034$$

Dataset Summary

Stage 1: Decision

44,034
training samples

4,416
validation samples

Stage 2: Reasoning

5,544
training samples

Held-out Test

4,896
samples

Stage 1 uses an equal number of samples from each of the six context-progression buckets:

$$7{,}339 \times 6 = 44{,}034$$

Two-Stage Training Settings

Hyper-parameter	Stage 1: Decision	Stage 2: Reasoning
Base Model Precision	4-bit (NF4)	bfloat16
LoRA Method	QLoRA	Standard LoRA
LoRA Rank (r)	16	32
LoRA Alpha (α)	32	64
LoRA Dropout	0.05	0.05
Target Modules	All linear layers[†]	
Batch Size	32	16
Micro Batch Size	2	2
Gradient Accumulation	16	8
Learning Rate	3×10^{-4}	1×10^{-4}
LR Scheduler	Cosine	Cosine
Num Epochs	3	3
Max Sequence Length	1024	1024
Optimizer	AdamW	AdamW

[†] q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj.

Baselines and Research Questions

Baselines

Group A: Vanilla open-source LLMs (sub-10B)

- DeepSeek-LLM-7B-Chat
- Llama-3.1-8B-Instruct
- Qwen2.5-7B-Instruct
- Yi-1.5-6B-Chat

Group B: Commercial API-based LLMs

- DeepSeek-V3, Qwen-Plus
- Gemini-2.0-Flash, GLM-4.6
- Kimi K2

Research questions

RQ1: Decision efficacy
How accurate are the model's discard choices in Nanikiru problems?

RQ2: Reasoning quality
Are explanations coherent, factual, and tactically grounded?

RQ3: Mechanisms and ablation
How much do dual-stage training and neuro-symbolic injection contribute?

RQ1: Overall Decision Accuracy

Model	Top-1 Acc	Top-3 Acc
DeepSeek-LLM-7B	3.3%	–
Llama-3.1-8B	3.0%	–
Qwen2.5-7B	8.8%	–
Yi-1.5-6B-Chat	3.6%	–
Qwen-Plus	22.5%	–
DeepSeek-V3	32.1%	–
GLM-4.6	25.4%	–
Kimi K2	24.1%	–
Gemini-2.0-Flash	19.9%	–
NeSy-DeepSeek	65.4%	**88.7%**
NeSy-Llama	66.5%	88.4%
NeSy-Qwen	**67.0%**	87.0%
NeSy-Yi	66.2%	86.5%

Best commercial baseline

32.1% Top-1

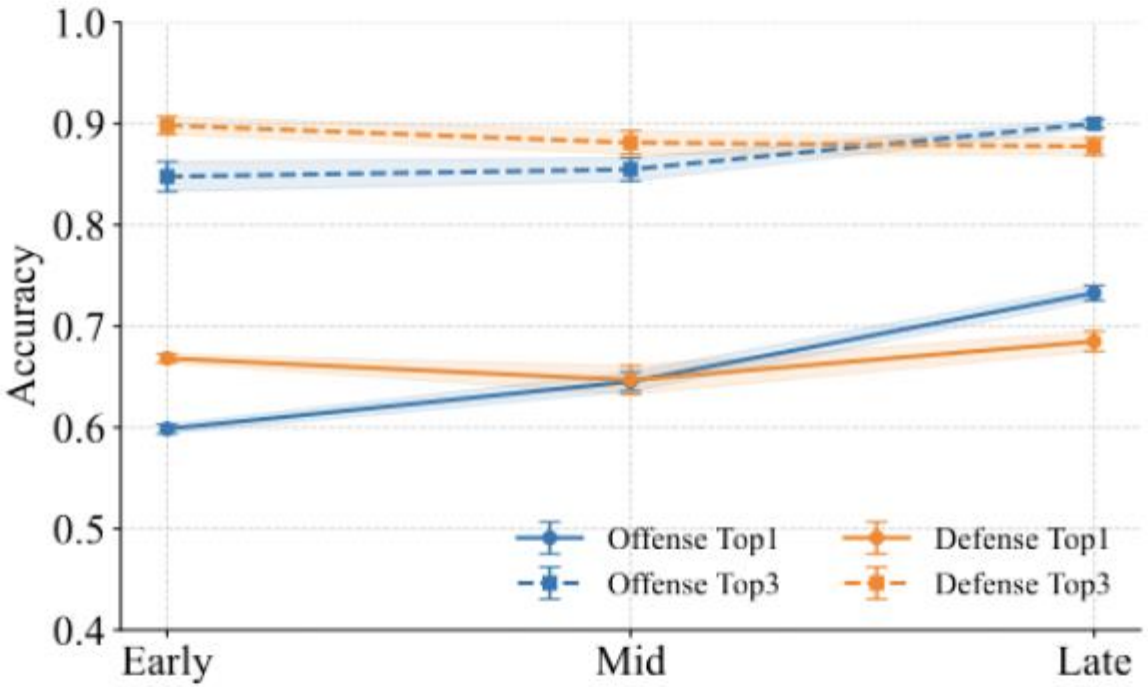
NeSy-Mahjong

67.0% Top-1

88.7% Top-3

All four backbones converge to a narrow, consistently strong accuracy range.

RQ1: Accuracy by Game Context



Accuracy remains stable across offense and defense; later decisions become more constrained and therefore more predictable.

RQ2: Evaluation Protocol

Stratified sample

240

test instances

$$40 \times 6$$

samples across progression $\times$ context buckets

GPT-4o as an independent judge

The judge receives three sources of evidence:

- the complete board state,
- the model-generated explanation,
- exact symbolic metrics as reference facts.

Three 1–10 scores

Logic	coherence of the argument
Precision	factual correctness
Strategy	tactical soundness

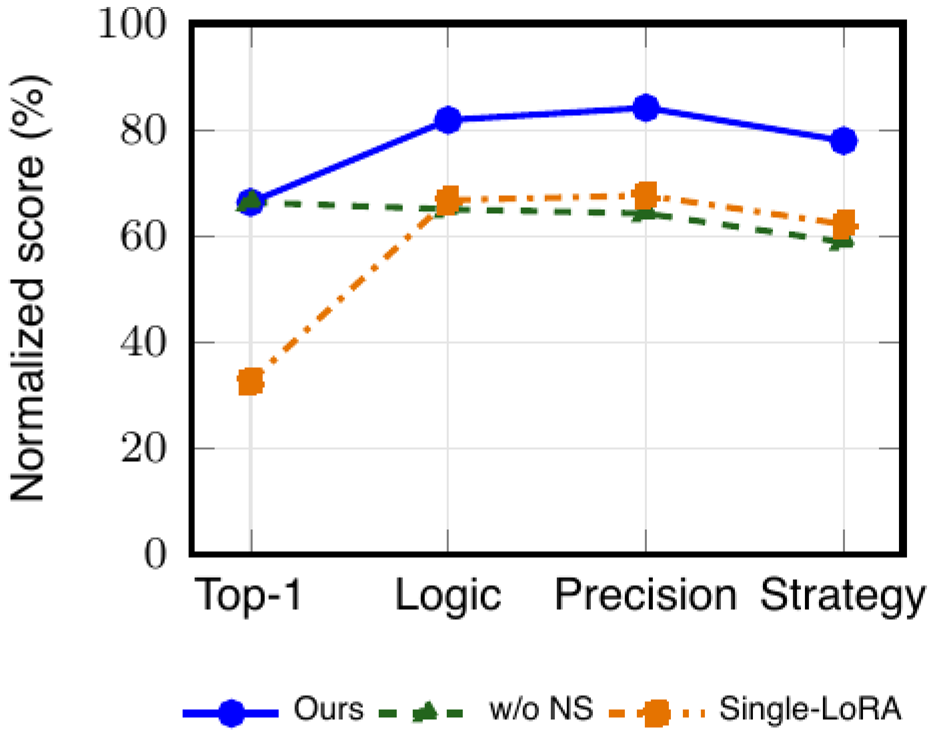
RQ2: Explanation Results

Model	Logic	Precision	Strategy
Llama-3.1-8B	2.26	1.38	1.40
DeepSeek-LLM-7B	2.09	1.46	1.44
Yi-1.5-6B-Chat	2.37	1.61	1.60
Qwen2.5-7B	2.93	2.14	2.21
Qwen-Plus	7.61	7.38	7.07
DeepSeek-V3	7.78	7.44	7.35
GLM-4.6	7.05	6.72	6.47
Kimi K2	7.32	7.19	6.82
Gemini-2.0-Flash	7.18	6.93	6.51
NeSy-DeepSeek	8.09	8.39	7.75
NeSy-Llama	**8.12**	**8.48**	**7.82**
NeSy-Qwen	8.08	8.43	7.81
NeSy-Yi	8.01	8.33	7.78

Takeaway: every neuro-symbolic model exceeds the commercial baselines on all three criteria.

RQ3: Component Ablations

Model Variant	Top-1	Logic	Precision	Strategy
NeSy-DeepSeek w/o NS	–	6.38	6.27	5.75
NeSy-Llama w/o NS	–	6.72	6.67	6.08
NeSy-Qwen w/o NS	–	6.59	6.57	6.05
NeSy-Yi w/o NS	–	6.30	6.16	5.62
NeSy-DeepSeek Single-LoRA	11.8%	6.79	6.64	6.10
NeSy-Llama Single-LoRA	43.4%	5.24	5.41	5.05
NeSy-Qwen Single-LoRA	17.7%	6.71	6.80	6.20
NeSy-Yi Single-LoRA	57.5%	7.93	8.20	7.53



Ablation findings

Removing symbolic grounding reduces reasoning quality. Reusing a single LoRA causes a much larger loss in decision accuracy, revealing decision–reasoning interference.

Decision Context

Game state

EAST 2, 0 Honba, 1 Riichi stick, 19 tiles left
Dealer: Player 1; Dora indicator: 8p

Your hand:
4m 5m 5p 6p 9p 2s 4s 6s 7s S S N N; draw 7m

Player 0 (Riichi) discards:
W E G Wh R N 9s 8p 6m 5s 6m 1s* 1m

Your discards:
1s 1m 8p 3p W 5m 9s 4m 3p 8m 1s 9s

What matters here?

- The game is already in a late phase.
- Player 0 has declared Riichi.
- The dealer must balance hand efficiency against immediate deal-in risk.

Which tile should be discarded?

Three Models, Three Behaviors

Vanilla Qwen2.5-7B: discard 5p

Claims that 5m–5p–5s forms a sequence and that 5p is safe.

Severe hallucination: this sequence is invalid and reveals a basic rule misunderstanding.

DeepSeek-V3: discard 9p

Treats 9p as isolated and prioritizes dealer offense.

Dangerous judgment: a non-Suji, likely live 9p is risky late in the game against Riichi.

NeSy-Llama: discard 2s

Detects the late-game Riichi threat; 2s is Suji to the Riichi player and widely visible.

Tactically grounded: uses symbolic threat and safety metrics to choose full defense, consistent with expert defensive principles.

Fluent language is not enough: the explanation must remain consistent with exact Mahjong rules and visible safety evidence.

Conclusion

1 **NeSy-Mahjong connects decisions with explanations.**

It combines intuitive neural decision making with explicit symbolic reasoning.

2 **Two mechanisms address two failure modes.**

The symbolic calculator reduces hallucination; Dual-LoRA preserves decision capability.

3 **Domain-specialized sub-10B models outperform strong commercial baselines.**

Explicit symbolic knowledge enables accurate decisions and grounded explanations.

Thank you.