ICML 2026 / Main Conference

MEMO

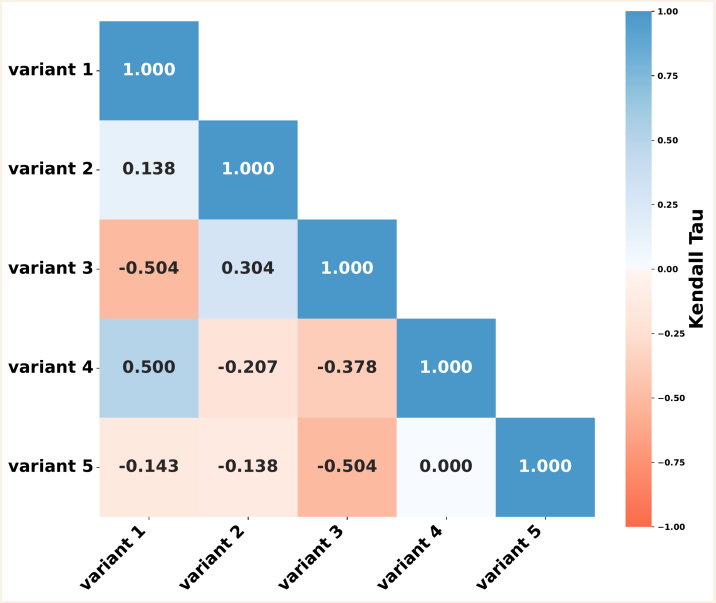
Memory-Augmented MOdel Context Optimization *for Robust Multi-Turn Multi-Agent LLM Games.*

Yunfei Xie* · Kevin Wang* · Bobby Cheng* · Jianzhu Yao · Zhizhou Sha · A. Duffy · Y. Xi · H. Mei · C. Tan · Chen Wei† · Pramod Viswanath† · Zhangyang Wang†

* *Denote co-first contribution*

LLMs Exhibit High Performance Variance in Games

Their ranking agreement varies across prompt variants



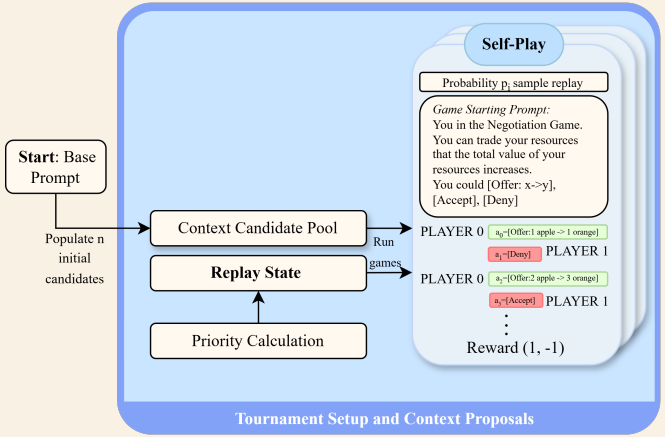
The same six-model tournament using the same prompt, was held under many prompt variants. The graph on the left depicts how often the ranking agrees across runs.

Why is the ranking noisy?

① **Context-dependent game behavior.**
Small wording changes affect how models interpret rules, legal actions, and strategic priors.

② **Multi-turn interaction amplifies noise.**
Each action becomes future context; one agent's deviation changes the other agent's response.

MEMO: a self-play loop over context



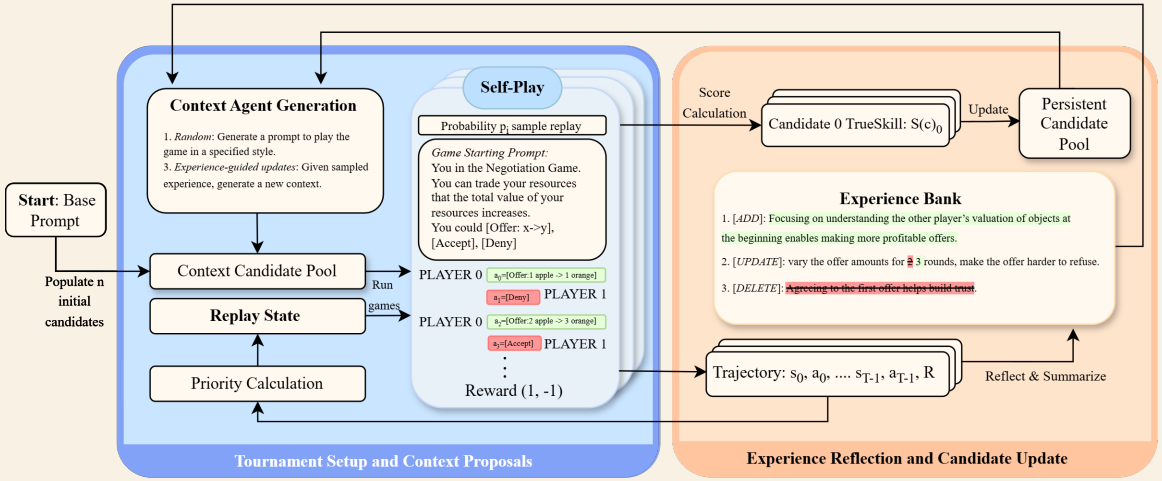
TWO COUPLED COMPONENTS

① Exploration

Tournament + Replay

Conservative TrueSkill selection. Replay over rare decisive states.

MEMO: a self-play loop over context



TWO COUPLED COMPONENTS

① Exploration

Tournament + Replay

Conservative TrueSkill selection. Replay over rare decisive states.

② Retention

Memory Bank

Distill self-play into structured insights. Inject as priors. Persist across generations.

RECIPE

Exploration generates and filters; retention accumulates lessons. The coupling works.

Tournament search + replay over the rare states

STRUCTURED SEARCH OVER CONTEXTS

Conservative TrueSkill selection.

SELECTION SCORE

$$S(c) = \mu_c - \kappa\,\sigma_c \text{, where } \kappa = 1$$

Each candidate's skill is a Bayesian Gaussian (μ, σ). LCB favours wins that are reliable

REVISIT THE DECISIVE, RARE STATES

Inverse-frequency replay.

PRIORITY & GATE

$$\text{priority}(\tau) = 1 \,/\, \text{count}(\tau)$$

$\alpha = 0.6$ · $\beta = 0.4$ · replay prob.

Stored as (prefix, env seed). Re-injected faithfully — surfaces legality boundaries, opponent quirks, breakdown points.

INSIGHT ✦ RECIPE

Exploration filters what memory keeps. Memory feeds what exploration mutates. The two together are what works.

Memory bank: lessons that persist across generations

STEP 1 · TRAJECTORY REFLECTION

Play → typed insights

Rule clarifications

e.g. legality, action format, transition effects

Strategic priors

e.g. when to bluff, when to concede, time pressure

Failure analyses

e.g. why this hand lost, what to do instead

LLM reflects on each completed trajectory.

STEP 2 · CRUD-STYLE MERGE

Add, edit, remove.

+ **ADD**

novel insight → keep it

~ **EDIT**

similar to existing → merge / generalize

– **REMOVE**

contradicts existing → drop both (no noise)

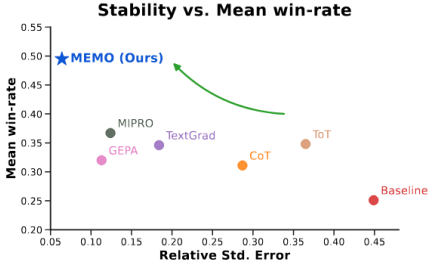
Bank persists. Sampled into 75% of next-gen candidates.

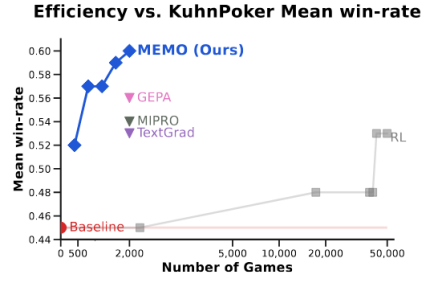
INSIGHT ✦ RETENTION

On long horizons, lessons MUST persist across episodes — that is what reduces variance and lifts performance.

Main Results

Summary Plot





Qwen2.5-7B-Instruct: aggregate across five games

Type	Method	SimpleNeg.	TwoDollar	KuhnPok.	Briscola	SimpleTak	Mean WR	Mean RSE	Games
STATIC	baseline	24.0%	17.1%	45.3%	2.8%	15.1%	**20.9%**	30.1%	2k
PROMPT	TextGrad	37.1%	29.3%	52.8%	7.1%	22.4%	**29.9%**	21.7%	2k
PROMPT	MIPRO	42.4%	47.5%	53.8%	2.2%	20.9%	**33.4%**	7.3%	2k
PROMPT	GEPA	34.4%	31.7%	55.8%	3.3%	19.3%	**28.8%**	14.8%	2k
RL	UnstableBaseline	41.1%	30.4%	52.7%	53.3%	47.3%	**45.0%**	43.3%	38k
RL	SPIRAL	45.7%	—	56.7%	—	32.7%	—	—	38k
OURS	**MEMO ✦**	**48.0%**	**48.4%**	**60.0%**	**31.5%**	**36.9%**	**45.0%**	**7.0%**	2k

Highlight

MEMO matches the RL mean win rate, but is far more stable and uses 19x fewer games.

Learned context transfers across games.

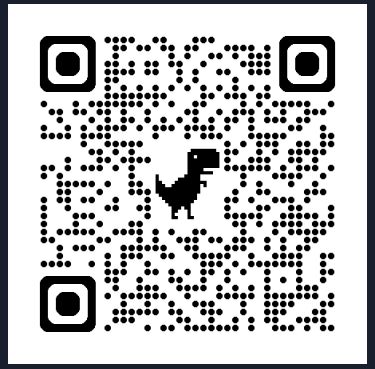
GPT-4o-mini · rows = source game (where MEMO trained) · columns = target game (zero-shot, no further tuning)

SOURCE ↓ / TARGET →	SimpleNeg.	TwoDollar	KuhnPok.	Briscola	SimpleTak
SimpleNeg.	**46.9%** *+15.6*	37.8% *+5.6*	48.9% *+9.8*	0.0% *-0.3*	37.7% *+16.3*
TwoDollar	31.1% *-0.2*	**48.7%** *+16.5*	53.3% *+14.2*	1.1% *+0.8*	47.8% *+26.4*
KuhnPok.	31.1% *-0.2*	34.4% *+2.2*	**57.2%** *+18.1*	22.2% *+21.9*	30.0% *+8.6*
Briscola	38.9% *+7.6*	27.8% *-4.4*	57.8% *+18.7*	**38.4%** *+38.1*	14.3% *-7.1*
SimpleTak	37.8% *+6.5*	35.6% *+3.4*	65.0% *+25.9*	0.0% *-0.3*	**30.7%** *+9.3*

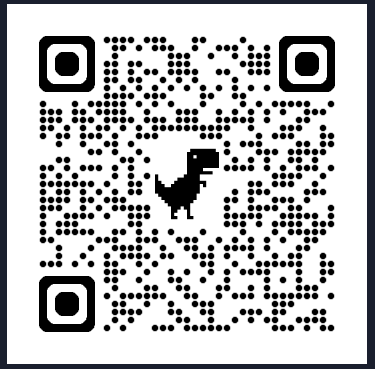
WHAT TRANSFERS

Learned context captures reusable protocol-level skills, but generalization is directional and task-alignment dependent.

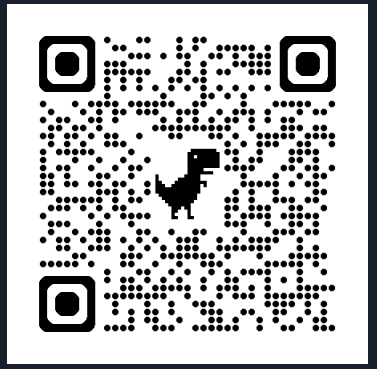
Thank you.



Website



Code



Paper

Code released alongside camera-ready