EXPERIENCE IS NOT KNOWLEDGE: WHERE THE DECISION KNOWLEDGE IN AN AGENT'S MEMORY COMES FROM

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents are commonly improved by writing a memory from their own trajectories and reading it back before the next episode. The gain is credited to experience, but the memory mixes what the writer already knew with what it induced from the trajectories. We ask where the useful decision knowledge in an agent's memory comes from. The testbed is Slay the Spire, a roguelike deckbuilder: every run is generated fresh and the same decisions recur, so a rule written from a few runs is tested hundreds of times in fresh ones. We hold the player model, the combat executor, the prompt and the form of the memory fixed, and vary only the source of the memory's content. The self memory is the player's own summary of its games. The others hold outcome statistics from randomized simulator games, a stronger model's prior knowledge alone, or, in the teacher memory, the player's trajectories read by a stronger model under a few human-written guidelines. The player's own memory does not raise its clear rate, and neither fewer games, a stronger writer nor writing one game at a time changes that. The memories written from statistics, from prior knowledge and by the teacher raise it, and the teacher memory raises it from 14 clears of 50 to 30. Replay at shared decision states shows that the player acts on its self-written memory, and that the memory poisons it. From the play counts in its own logs the writer concluded that the basic attack, which every deck starts with and so plays most, deserves an early upgrade, and the player then made in the exam an error it had almost never made in the games the memory was written from. The statistics and teacher memories reach their gains by different routes: one keeps adding cards to a large deck and the other stops. The teacher memory, written from one player's trajectories, is the best memory for the other two players, whose own memories again do not help them. The three players converge on what it states simply and stay apart on the rules they must compose. Experience is not knowledge. The player improved when its memory held correct, usable rules, whether a stronger reader extracted them from the player's own games or brought them from outside, and a wrong extraction from those same games became a new error.

1 INTRODUCTION

LLM agents are commonly improved by writing a memory from their own trajectories and reading it back before the next episode (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2024; 2025; Zhang et al., 2026; Ouyang et al., 2026). The gain is credited to experience, but the memory mixes what the writer already knew with what it induced from the trajectories, and a higher score after the update credits both.

We ask where the useful decision knowledge in an agent's memory comes from. A roguelike suits the question: every run is generated fresh, so what carries from one run to the next is a reusable rule, not a remembered path. In Slay the Spire, a roguelike deckbuilder, the same few kinds of decision, over cards, upgrades and routes, recur many times per run, so a rule written from a few runs is applied, and tested, hundreds of times in fresh ones. One LLM player makes the decisions between fights, and a fixed search-based executor plays every fight. We hold the player, the executor, the prompt and the form of the memory fixed, pseudonymize card and relic names so that the player

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

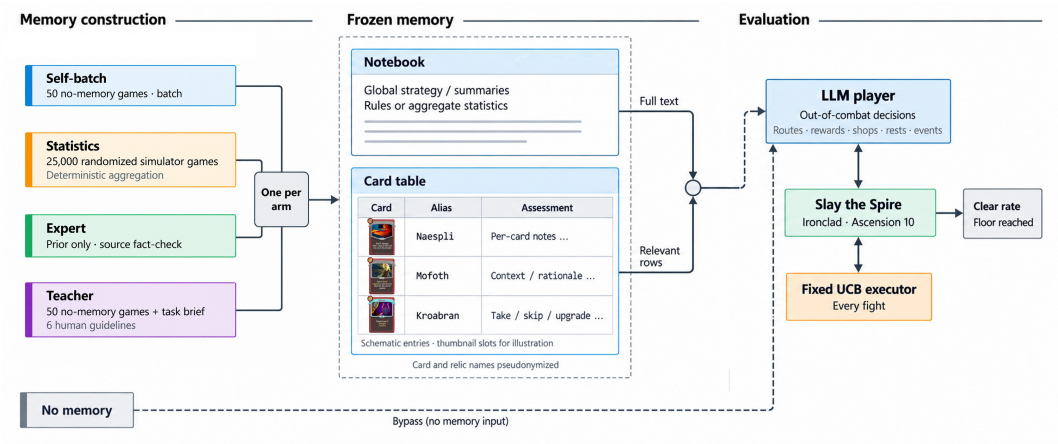


Figure 1: Overview. One memory, or none, is frozen and read by the player: the notebook in full, the card-table rows for the cards in view, all names pseudonymized. The player makes the out-of-combat decisions and a fixed UCB executor plays every fight. The table entries shown in the figure are schematic.

cannot lean on memorized knowledge of the game, and vary only the source of the memory's content (Figure 1, Table 1). Because the same player reads every memory in the same prompt, a difference between arms is a difference in what the memory says.

Five arms span the sources, one without memory (No memory). The *self* memory, Self-batch, is written by the player model in one pass from its own memoryless games. Statistics is a notebook of numbers computed by a program from 25,000 randomized simulator games. Expert is written by Claude Opus 5 from prior knowledge alone. Teacher is written by Claude Fable 5.1 from 50 memoryless trajectories under six human-written guidelines.

Four findings follow. First, the player's own memory does not raise the clear rate, and neither fewer games, a stronger writer nor writing one game at a time changes that. The three memories whose content reaches beyond the player's games clear more games, the teacher memory significantly, from 14 to 30 of 50 (Section 4.1). With the same writer and memory form, prior knowledge alone clears more games than the player's own games do. Second, a memory changes the player's actions whether or not it helps, and the memories that help do not share a route (Section 4.2). Replay at shared decision states shows the player acting on its self-written memory, changing its most common answer at over a third of the states, and one rule shows what it acted on. Strike, the basic attack, starts every deck in five copies and so tops the play count of any log. Both batch memories the player wrote took the count for merit and told it to upgrade Strike early, and it did so in 23 and 11 of 50 exam games, almost never in the games the writers read. Given the same games as the first, the stronger writer told it to remove Strike. A partial conclusion drawn from its own logs became a rule the player then followed: the memory poisoned a decision it was written to improve. The expert notebook moves take versus skip and the statistics notebook which card, and the statistics and teacher memories, which both raise the clear rate, reach it by different routes: one keeps adding cards to a large deck and the other stops. Third, the teacher memory shows what the trajectories are good for (Section 4.3). It sets corrections of failures seen in the trajectories beside engine facts and strategy the trajectories do not contain: the episodes locate the player's errors, and prior knowledge supplies the fix. It also carries errors that an independent check confirms. Fourth, the teacher memory, written from one player's trajectories, is the best of five arms for the other two players and more than doubles the clears of both (Section 4.4). The memories those players write from their own games again do not help them. Under it the three players' per-card choices nearly coincide: what the teacher wrote is knowledge of the game, not a diagnosis of the player it read. The three players converge on the rules the memory states simply and stay apart on the rules they must compose. Experience is not an asset in itself. The same episodes, read by the player, yielded a rule that misled it, and read against their outcomes by a writer that brought its own knowledge to them, yielded corrections that helped three players.

2 Related Work

Memory written from experience. One line of work turns an agent's own trajectories into reflections, insights or skills that it reads later (Shinn et al., 2023; Zhao et al., 2024; Wang et al., 2024). Others write workflows, guidelines or structured context (Wang et al., 2025; Fu et al., 2024; Suzgun et al., 2026; Zhang et al., 2026; Ouyang et al., 2026), and others store observations and retrieve from them (Park et al., 2023; Packer et al., 2023; Xu et al., 2025; Zhang et al., 2024b). Variants train the reflector on reward, reflect on card-game policies, or learn from failed trajectories (Yao et al., 2024; Zhang et al., 2024a; Song et al., 2024; Zhou et al., 2024). Agents follow a memory closely enough that errors in the store propagate (Xiong et al., 2025). Without external feedback, a model asked to correct its own reasoning does not improve it (Huang et al., 2024). Every such memory mixes what the writer already knew with what it induced from the trajectories. The closest control, in ExpeL, extracts insights from the few-shot examples instead of gathered experience and finds no gain (Zhao et al., 2024). We instead ask a writer without any trajectory for its own knowledge, in the same memory form.

Prior knowledge as context, and card-game agents. Other work gives the agent text it did not write itself. A game manual is read inside Monte-Carlo search (Branavan et al., 2012) and object interactions are extracted from Atari manuals (Wu et al., 2023a). A paper is traversed through a question graph (Wu et al., 2023b), and human-written skills initialize a strategy layer (Cheng et al., 2026). Our Expert arm belongs here, and our Teacher arm joins the two kinds of source. Bateni & Whitehead (2024) put an LLM agent in Slay the Spire (Mega Crit, 2019) and report where its play breaks down. Other systems build skills from GUI interaction or retrieve them through a state-action graph (Du et al., 2025; Tang et al., 2025). These works build the agent, and we fix the agent and vary the content of its memory.

3 Setup

3.1 Task and player

The environment is *Slay the Spire* (Mega Crit, 2019), the Ironclad character at Ascension 10, a raised difficulty setting, in an open-source simulator. A run climbs three acts of floors, and each floor is a fight, an elite (a harder fight that drops a relic), a rest site, a shop or an event. A won fight offers one of three cards, or none. A rest site heals the player or upgrades one card. Relics are passive items, and a boss relic is chosen after each of the first two act bosses. A run ends when the player dies or beats the act-3 boss, which clears it, and its score is the floor it reached.

We chose the game for four properties. Runs are procedurally generated, so a rule taken from one run must hold over a distribution of runs rather than a fixed path. Runs are independent, so a memory is the only channel between them. The outcome, the floor reached or a clear, is a dense signal that needs no annotation. A run holds hundreds of decisions of the same few kinds, so 20 memoryless games already contain 443 card-reward decisions, enough for the matched-state replay of Section 4.2. An environment that needs thousands of runs to show a signal would not allow this.

Three LLM players, DeepSeek Flash (DeepSeek-AI, 2026), Qwen3.8-Flash-Next (Qiu et al., 2026) and Ling-3.0-flash (inclusionAI, 2026), make every out-of-combat decision that has more than one sensible action (Appendix A.1). All three run the five arms of Section 3.3, and DeepSeek Flash also runs three ablations. Every fight is played by a fixed omniscient UCB search executor, identical in all arms, so an arm differs only in the memory the player reads. Card and relic names are replaced by pseudonyms under a fixed renaming, so the player cannot use memorized wiki knowledge, while the rules text of every item in view is still shown. Quotations from memories in this paper give the true names in place of the pseudonyms the player saw. A comparison with real names, in Appendix A.8, shows the prior the renaming removes. Appendix A gives the remaining settings.

3.2 Memory

A memory is a notebook of free text plus a card table with one row per card. One function renders it in every arm: the notebook whole, then the rows for the cards in view (Figure 1). The arms therefore share the form and differ in content and length. Appendix A.6 gives the rendering detail and the size of each memory.

Table 1: The five arms, run by all three players. DeepSeek Flash also ran the three self-memory ablations of Section 4.1, and Table 4 in Appendix A lists all eight with the writer-side detail and the rendered memory sizes. Exam games counts the games an arm was scored on, for each player.

Arm	What the writer read	Writer	Exam games
No memory	nothing, no memory is rendered	–	50
Self-batch	the player's own No memory games, 50 or 20	the player itself	50
Statistics	25,000 randomized simulator games	a program	50
Expert	prior knowledge, 0 games	Claude Opus 5	50
Teacher	50 No memory games, a brief and 6 guidelines	Claude Fable 5.1	50

3.3 Five arms

Table 1 lists the five arms, which all three players ran. Self-batch is the player's own memory, written by the player model in one pass from a bundle of its own No memory games. For DeepSeek Flash these are its 50 No memory games, the ones Teacher also read, and for the other two players their first 20 No memory games. The writer reads the games under the pseudonyms the player saw, so that the notes could carry nothing but what the games contained. Statistics is a notebook whose every sentence is a number computed by a deterministic program from 25,000 simulator games. Expert is prior knowledge with no trajectory, corrected against the card data and the simulator source. Teacher read the 50 No memory games of DeepSeek Flash, a task brief and six human-written guidelines. The brief and the guidelines are the whole of the human input (Appendix A.7). These three memories were written in true names and renamed for the player, because their purpose was to bring to the cards what the player's games do not contain. Qwen3.8-Flash-Next and Ling-3.0-flash read the Statistics, Expert and Teacher files of DeepSeek Flash, byte-identical, so none of the three was written from their games. Appendix A.10 gives their serving and writer settings.

DeepSeek Flash also ran three ablations of the self memory. In the schedule ablation, Self-seq, the player rewrites its notebook after each of 20 learning games and is scored on 20 exam games. In the material ablation, Self-batch-20, the player writes its one-pass memory from those 20 games instead of 50. In the writer ablation, Self-batch-20 (Opus), Claude Opus 5 writes from the same 20-game bundle, byte for byte. All three writers read the pseudonyms the player saw.

3.4 Evaluation

The primary endpoint is the clear rate, the fraction of games that beat the act-3 boss. The secondary endpoint is the score, the floor the run reached, with a cleared game scored 51. Each arm is scored on 50 exam games, and the schedule ablation on 20 (Appendix A). Every comparison in the paper is unpaired. The clear rate carries Wilson 95% intervals and Fisher's exact two-sided test, and the score the two-sided Mann–Whitney U test. Holm correction adjusts the comparisons against No memory, one family per player, which for DeepSeek Flash includes the three ablations (Appendix B).

4 Results

We report outcomes first, by content source (Section 4.1). We then look inside the memories: replay at shared states shows how each memory changes decisions, and an upgrade audit traces one written rule from the writer's input to the player's behaviour (Section 4.2). Section 4.3 reports the teacher memory and reads its content. A number in Sections 4.1 to 4.3 is for DeepSeek Flash, which also ran the three ablations, unless a player is named. Section 4.4 turns to the other two players and reports what holds for all three and what varies.

4.1 Outcomes by memory source

The player's own memory sits at the No memory rate, and the three memories whose content reaches beyond the player's own games sit above it (Figure 2). Teacher clears 30 of 50 games against 14 for No memory and separates from it after Holm correction over the seven memory arms of DeepSeek Flash at $p = 0.016$. Statistics and Expert separate only before correction (Appendix B),

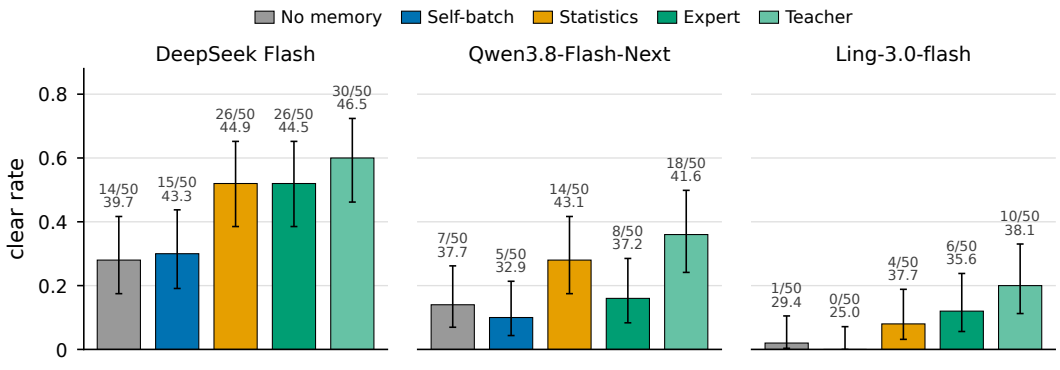


Figure 2: Clear rate per arm for the three players, with Wilson 95% intervals, and the clears over the arm's exam games and the mean score printed above each bar.

and the score orders the arms the same way (Table 6). Self-batch reaches act 3 about as often as Statistics and Expert do but clears the lowest share of those games of any arm, losing most of the rest at the final boss. Its notebook names the act-3 bosses only as places where runs ended and holds no plan for beating them. The order follows where the content comes from.

The self memory fails and the teacher memory helps whichever player reads them (Figure 2). Teacher has the highest clear rate for each, and No memory and Self-batch the two lowest, within two clears of each other. The middle of the order differs by player (Section 4.4). At the two ends the source of the memory sets the order, not its reader.

Ablations of the self memory. Fewer games, a stronger writer or a different schedule leave the self memory at the No memory rate (Table 6). Claude Opus 5, writing from no trajectory, produced Expert, which clears more games in the same form than Self-batch-20 (Opus). The two differ in what the writer was allowed to bring, its own knowledge or only the pseudonymized games. The stronger writer helps through what it knows, not through what it reads. Fixing the games instead of the writer gives the same answer: Self-batch and Teacher were written from the same 50 No memory games and clear 15 against 30 of 50. The clear rate moves with what the writer may bring to the material, not with how much of it there is, who writes from it or on what schedule.

4.2 From memory content to decision changes

Common-state replay. The three memories that raise the clear rate all change the player's card choices, each through a different channel. Expert changes whether the player takes a card at all. Statistics changes which of the offered cards it takes and makes that pick consistent. Teacher does both. We replayed 148 card-reward nodes from played games 10 times each under the five arms (Figure 3, Appendix C). Self-batch, the player's own memory, moves the channel of Statistics without its gain. It changes the player's most common answer at 36% of nodes, makes it more consistent and takes cards at least as often as No memory. A player can act on a memory without gaining from it. Followed about as closely as Statistics, the self memory fails in what it says, not in being read. Statistics and Teacher reach their gains by different routes. With 30 cards or more in the deck, Statistics still takes a card at 92% of nodes and Teacher at 69%, below No memory. One keeps adding to a large deck and the other stops. Expert, with the clear rate of Statistics, stops sooner still. A clear rate does not fix the policy behind it.

Each memory settles a state that the player alone leaves undecided, and they settle it differently. Two act-2 reward screens of a 30-card deck show this (Table 12 in Appendix C). Offered one upgraded attack, No memory takes it in half of its replays, Statistics in every one, and Expert, Teacher and Self-batch in almost none. Offered three cards, No memory spreads its replies over every option, skipping included. Statistics and Self-batch take its most frequent pick, Anger, every time, Expert always skips, and Teacher never takes Anger and splits between skipping and Cleave, an attack that hits every enemy. Almost every Statistics reply cites its table, which rates both cards a strong take. A row of that table averages a card over every deck it was offered to, so its verdict does not weaken

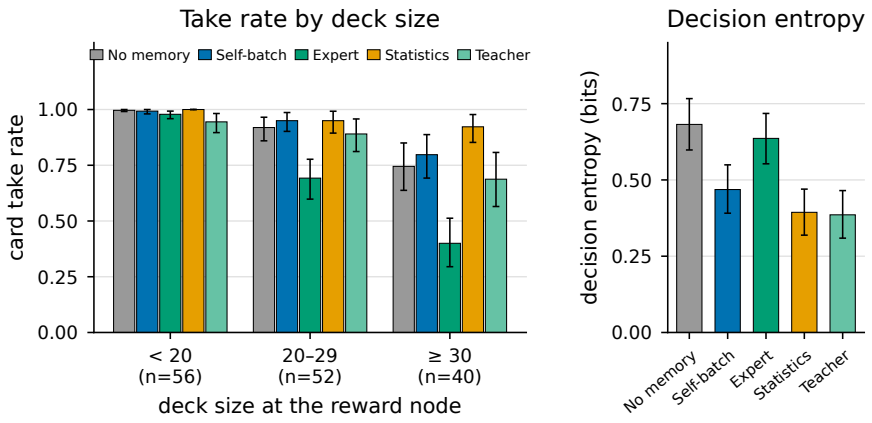


Figure 3: Left: card-take rate by deck-size stratum in the common-state replay, with 95% intervals from resampling nodes. The strata hold 56, 52 and 40 of the 148 nodes. Right: entropy, in bits, of the player's choice over the 10 replays of a node, averaged over nodes, with the same intervals.

as the deck grows. A memory that rates cards apart from the deck resolves such states toward taking, and a memory that weighs the deck resolves them toward skipping.

What the memory changes when a card is offered. The exam games show the same channels card by card. A memory first changes whether a card is taken at all. In the exam games DeepSeek Flash takes a card at fewer offers under Expert and Teacher than under No memory, and the gap grows with the act. In act 3 it takes a card at 80% of the offers under No memory, 68% under Teacher and 52% under Expert. The skipping under Expert matches how its rules are written. Expert makes the act the condition of a rule far more often than Self-batch-20 and Self-batch-20 (Opus). Its take rate also falls further from act 1 to act 3 than theirs (Table 14 in Appendix C). A memory then changes which card is taken, and each memory moves its own cards (Figure 4). Under Teacher, Disarm, Armaments and Flame Barrier go from mostly skipped to mostly taken, and Anger, Headbutt and Combust fall to no takes at all. Expert skips several of the cards Teacher takes, and the two disagree openly on Bludgeon and Uppercut, which Teacher takes at most offers and Expert at few. Two memories can both raise the clear rate while agreeing on little beyond skipping more.

From text to behaviour. The memory names the cards it moves. Teacher names all six of the cards above and sorts cards into tiers of its own, from A-tier, to take on sight, down to Skip. Its card table writes of Disarm "A-tier: taking 2 Strength off the enemy for the whole fight beats raw block against Champ [an act-2 boss] and the act 2 elites". Its notebook lists Anger and Headbutt under "Skip". A conditional rule in the text comes back as a near-absolute rule in behaviour. The Expert rows say to take Disarm "when Block is already sufficient", Flame Barrier "when the act ahead is full of multi-hit enemies" and Uppercut "when the deck has no other source of Vulnerable". The player takes the three cards at 0 of 29, 1 of 21 and 3 of 20 offers, against 4 of 36, 6 of 25 and 13 of 16 without memory. The hedge is read as a skip.

From an aggregate count to an upgrade rule. A memory induced from the agent's own logs can write a misreading of those logs into a rule that the agent then follows (Table 15 in Appendix C). Strike, the basic attack, starts every deck in five copies and so tops the per-card play count the self writers received, which has no field for copies. As writer, DeepSeek Flash took the count for merit. Self-batch-20 opens its upgrade order with "your most-played cheap attack (Strike, Twin Strike, Anger)", and the Self-batch row for Strike advises "Smith [upgrade] one early". The player follows the rule. Under Self-batch it upgraded Strike in 11 of 204 picks, against 1 of 201 in the games its writer read ($p = 0.0057$), and under Self-batch-20 more often still. The misreading belongs to the writer, not to the count. Given the same bundle byte for byte, the Self-batch-20 (Opus) writer notes the same count but writes that Strikes "become what removal screens are for" as the deck grows. Under its memory the player never upgrades Strike. Read against their outcome, the games behind Self-batch point the other way: the Teacher writer found that winners ended with fewer Strikes and

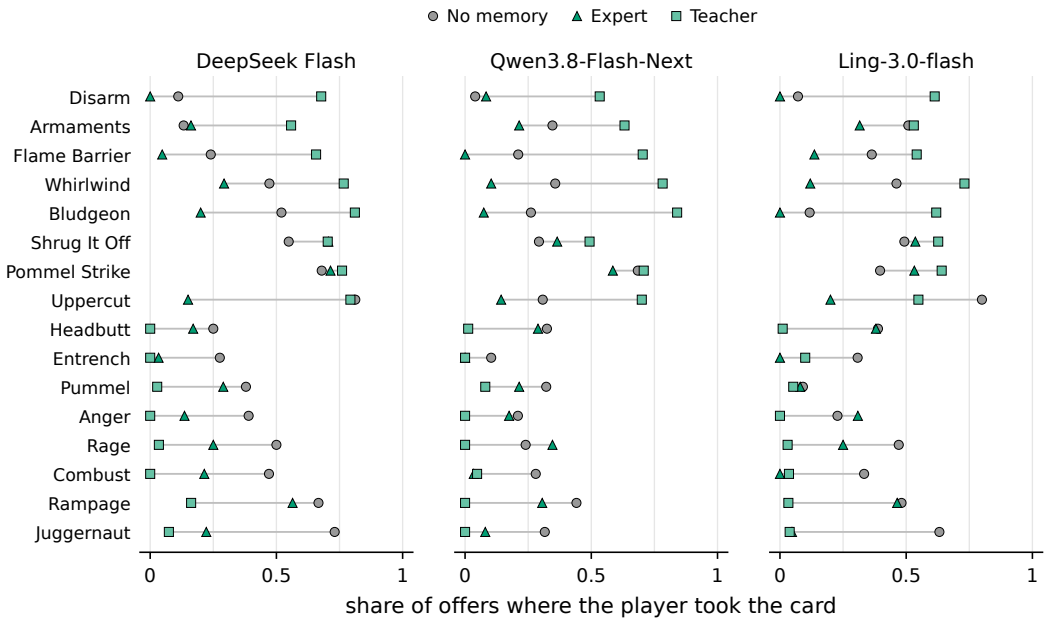


Figure 4: Per-card take rate at the card-reward screen for three players reading the same memories, over the 50 exam games of each arm. Every cell rests on at least 10 offers and most on 25 or more. The 16 rows are the 14 cards whose DeepSeek Flash take rate moves most between No memory and Teacher or Expert, plus two cards DeepSeek Flash takes by default, sorted by its Teacher minus No memory shift. Circles mark No memory, triangles Expert and squares Teacher, and a grey line spans a card's three rates.

Defends than losers. A memory can take what its logs record most for what matters most, and its reader acts on the mistake.

4.3 What the teacher memory contains

With DeepSeek Flash, Teacher is the one arm that separates from No memory after correction, though not from Expert (Section 4.1). This section reads what the memory contains.

How it was written. The writer was a Claude Fable 5.1 session with tools. It read the 50 No memory games of DeepSeek Flash in true names, a task brief and six human-written guidelines, and returned a 52-item notebook and a 125-row card table. The human wrote nothing else and did not edit the output (Appendix A.7). The teacher's prior knowledge therefore has two sources, the writer's own and the six guidelines, which run to 125 words in all. Two of the notebook's headline rules restate guidelines, Runic Dome first among boss relics and never upgrading Strike or Defend. We do not separate the two sources of prior knowledge.

What it says. The notebook holds the two kinds of content of Section 1: rules induced from the games, which correct what the trajectories show the player doing, and prior knowledge, restated with a reason attached. The corrections point at particular runs. "Never hold two relics that each remove a resource (Rest, gold, potions, rewards)" is written against a run that took Coffee Dripper and then Ectoplasm, two relics whose drawbacks remove healing at rest sites and all gold income, and so entered act 3 unable to heal or to shop. The rule for upgrades at a rest site, "Smith the biggest gain among cards actually played in the last two fights, never a card unplayed in both", answers upgrades of cards the executor never played. Other rules hand the player a number that separated winners from losers, "Spend gold: losers died holding 136 on average", or a count read straight off the games, "Coffee Dripper (0 wins in 5 takes)". The self memory, written from the same games, lists where runs died but never sets the runs that won against those that lost. The prior knowledge comes back with its reasons written out. Runic Dome is a strong boss relic that human players take, and its drawback, not seeing what the enemy will do next, is mild for a player whose fights are

played by the search, so the notebook ranks it first among boss relics. Cards with a random effect are mostly strong cards, and the notebook tells the player to judge them by their average effect. The advice never to upgrade Strike and Defend stays, the reverse of the self memory's rule.

The notebook departs from the prior where the games contradict it. The brief's example of act-1 play named three attack cards as ones not to take on sight, and after reading the games the notebook puts two of them on its take-on-sight list. The guideline to rest more and fight fewer elites is reported as contradicted by the 50 games, in which winners fought more elites than losers, and is rewritten as a gate on health and on the distance to the next rest site. The act is a condition throughout: "Act 3: HP spent here never comes back" heads the act-3 routing rules, the deck targets are written per act, and the last section is a preparation list per boss. The reading shows what the memory is made of, corrections of observed play and prior knowledge the trajectories do not contain. It does not show how much each contributes to the clear rate.

Errors in the memory. The memory is not correct line by line. The teacher reported 11 problems against its own memory, and an independent check confirms 8 (Appendix E). One note confuses two attack cards. One rule turns a winners' average into a ceiling, "winners ended with at most 4 Strikes and Defends". The block, exhaust and damage thresholds in the notebook appear in neither the games' summaries nor the human input, and three card descriptions are imprecise. A memory can raise the clear rate while carrying errors that can be pointed at line by line.

4.4 The other two players

Most of what Sections 4.1 to 4.3 report for DeepSeek Flash holds for the other two players, and where it does not, the difference lies in what each player makes of the same rules. The batch memory a player writes from its own games raises no player's clear rate, and for Qwen3.8-Flash-Next and Ling-3.0-flash it sits below No memory on both endpoints (Figure 2). Teacher clears 18 of 50 games with Qwen3.8-Flash-Next and 10 with Ling-3.0-flash, the most of any arm for both, and after Holm correction its gain is significant for Ling-3.0-flash only (Table 7 in Appendix A.10). Written from one player's games, the teacher memory carries to players it never read.

Same memory, same cards. Under Teacher the memory, not the player, sets the per-card preference. Over the cards of Figure 4, the mean pairwise correlation of the three players' take rates rises from 0.49 without memory to 0.80 under Expert and 0.97 under Teacher. Over all cards offered often enough to every player, the spread left between players under Teacher is close to the sampling noise of three players that share one rate per card (Appendix D). A per-card verdict is the simplest rule a memory can state. The convergence stops where the rules stop being simple.

Simple rules converge, compound rules do not. The three players clear different numbers of games under one memory, and the gap is visible in how they follow its rules (Table 2). We classed the 21 checkable rules of the Teacher notebook by what they ask of the reader. A simple rule names the answer or gates it on one number printed on the screen: never take a Skip-tier card, rest below half health, never upgrade Strike or Defend. A compound rule combines several conditions or pursues a goal across screens: fight an elite only when health, a damage source and a rest ahead all allow it, or keep taking A-tier cards once the deck passes 30. On the simple rules the players converge, and the gap between the most and the least compliant player all but closes. On the compound rules it does not close at all, and the two classes differ in how the memory moves them at $p<0.001$. Past 30 cards DeepSeek Flash still takes an offered A-tier card at 72% of offers and Ling-3.0-flash at 35%, below its own 65% without memory. The replies of Ling-3.0-flash cite the notebook's act-2 target of 30 cards in act 3 as a reason to skip. Where a rule must be composed, the same text pulls readers apart.

Where the clears go. DeepSeek Flash still clears the most games under Teacher (Appendix B.3), and the other two lose most of the difference at the act-2 and act-3 bosses. Their decks deal an eighth to a sixth less damage per turn than DeepSeek Flash's to the same act-2 bosses, a gap already there without the memory. Qwen3.8-Flash-Next loses nearly all its missing clears there, while Ling-3.0-flash also dies more often on the way and misreads one simple rule. In 8 of the 52 shop visits where it declined an affordable card removal, it tells itself that removal works only on curses, which the other two almost never do. The memory equalises the decisions it can state simply. What remains is the reader, the deck it composes from the same verdicts and what it makes of a rule.

Table 2: How the three players read the memories. Left, mean compliance with the Teacher rules under Teacher, No memory in parentheses. Simple: the 13 rules that name the answer or gate it on one on-screen number. Compound: the 8 that combine conditions or pursue a goal across screens (all 21 in Table 16, Appendix D). A-tier 30+: take rate of offered A-tier cards at 30 or more cards in the deck. Right, the Expert skip rule under No memory, Expert and Teacher (N/E/T). Act-1 take: card rewards taken in act 1, %. Deck size: % of Expert skips whose reply cites deck size. Non-basic: cards other than Strike, Defend, Bash and curses on entering act 2.

	Teacher rules, T (N)			Expert skip rule		
Player	Simple	Compound	A-tier 30+	Act-1 take N/E	Deck size in E skips	Non-basic N/E/T
DeepSeek Flash	.90 (.69)	.74 (.59)	.72 (.58)	98/91	20	12.6/10.2/10.6
Qwen3.8-Flash-Next	.90 (.67)	.62 (.49)	.47 (.35)	93/77	73	10.9/ 8.6/10.0
Ling-3.0-flash	.88 (.56)	.58 (.51)	.35 (.65)	99/81	66	11.3/ 9.2/ 9.6

Conditional rules and the reader. Qwen3.8-Flash-Next and Ling-3.0-flash read the Statistics table as well as DeepSeek Flash does, all three taking its top-rated offer at about 72% of the offers it rates, but they apply the Expert skip rule outside its stated condition (Table 2). Every Expert notebook rule carries a condition, where Teacher also gives flat lists of cards to take and to skip. The skip rule holds in act 2 or 3 with 20 or more cards, and DeepSeek Flash, the strongest player without memory, makes most of its skips there. The other two skip in act 1 too, cite deck size in most skips and enter act 2 with the thinnest decks in the table. Half of Qwen3.8-Flash-Next's losses under Expert end at the act-2 boss. Qwen3.8-Flash-Next's own Self-batch notebook gives the same advice without the condition and leaves the same thin decks, consistent with its clears below No memory. In clears, no player-by-memory interaction separates from 50-game noise (p=0.61). The pattern is consistent with the stronger player applying the condition and the weaker ones the verdict. A memory of conditional rules asks more of its reader than one of flat verdicts.

5 Discussion and Conclusion

Implications for agent memory. The clear rate followed what the memory correctly extracted, not whether or how many of the player's own games were read. The player's games do not say which of their counts carry decision value, and both batch memories it wrote took frequency for merit. A writer with its own prior knowledge and six human guidelines read the same games and wrote corrections of the player's errors beside rules the games do not contain. Statistics reaches the clear rate of Expert for DeepSeek Flash, but it exists only because the simulator scores every game for free, a signal most settings with a memory lack. The practical reading is to invest in what the writer knows before it reads, and to use the agent's own episodes to locate failures, not to derive strategy from them.

What transfers. Written from one player's games, the teacher memory was the best arm for two players whose games it never read. Under Teacher the three players' per-card take rates nearly coincide (Figure 4). The useful content of the teacher memory is therefore knowledge about the game and the executor (Section 4.3), not a diagnosis of the player it read. The three players converge on its verdicts and one-condition rules and stay apart on the rules they must compose (Section 4.4).

Limitations. The study covers one game, one character and one ascension level. Each arm has one memory, so writer variance is not sampled, and with 50 games per arm the design resolves differences in clear rate of about 20 points, not smaller ones.

Conclusion. Memories written from a player's own games, by the player or by a stronger model kept from its own knowledge, raised no player's clear rate. The same games, read against their outcomes by a writer with knowledge of its own, gave the memory that raised the clear rate of all three players. Experience written into memory is not an asset in itself. Statistics is experience too, but 25,000 scored games, 500 to 1,250 times what the self writers read, reduced by a program: outcome feedback at a density an agent's own episodes do not reach. The asset is what a memory correctly extracts from it, and a wrong extraction becomes an error the agent did not have before.

Ethics statement

The study runs an open-source reimplementation of a single-player video game. It involves no human subjects and no personal data. The only human-written inputs are the six guidelines given to the teacher writer, reproduced in Appendix A.7, and the task brief, summarized there.

Reproducibility statement

We release as supplementary material the simulator harness, the player and writer code, every memory notebook and card table exactly as the player read them, the per-game result files of all 370 exam games of the main study and the 500 of the other two players, the replay samples, the teacher brief and guidelines, and the scripts that compute every table and figure from those files. Per-arm run lists are in Appendix A.

AI use statement

We used generative AI tools to implement standard components of the experimental pipeline, to assist with data analysis, and to draft and edit parts of the paper. Every other use of a language model in this work is part of the object of study and is described in Section 3. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims and artifacts produced with the aid of generative AI.

References

Bahar Bateni and Jim Whitehead. Language-driven play: Large language models as game-playing agents in slay the spire. In *Proceedings of the 19th International Conference on the Foundations of Digital Games (FDG)*, pp. 1–10. ACM, 2024. doi: 10.1145/3649921.3650013.

S. R. K. Branavan, David Silver, and Regina Barzilay. Learning to win by reading manuals in a monte-carlo framework. *Journal of Artificial Intelligence Research*, 43:661–704, 2012.

Xiangchen Cheng, Yunwei Jiang, Jianwen Sun, Zizhen Li, Chuanhao Li, Xiangcheng Cao, Yihao Liu, Fanrui Zhang, Li Jin, and Kaipeng Zhang. AgenticSTS: A bounded-memory testbed for long-horizon LLM agents, 2026.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026.

Jiawei Du, Jinlong Wu, Yuzheng Chen, Yucheng Hu, Bing Li, and Joey Tianyi Zhou. Rethinking agent design: From top-down workflows to bottom-up skill evolution, 2025.

Yao Fu, Dong-Ki Kim, Jaekyeom Kim, Sungryull Sohn, Lajanugen Logeswaran, Kyunghoon Bae, and Honglak Lee. AutoGuide: Automated generation and selection of context-aware guidelines for large language model agents. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. arXiv:2403.08978.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.01798.

inclusionAI. Ling-3.0-flash, 2026. URL `https://huggingface.co/inclusionAI/Ling-3.0-flash`. Model card.

Mega Crit. Slay the spire. Video game, Mega Crit Games, 2019. Released 23 January 2019.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. ReasoningBank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.25140.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems, 2023.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.

Zihan Qiu, Zekun Wang, Xiao Li, Yanpeng Li, Yang Xu, Yixuan Wang, Huaqing Zhang, Rui Men, Bochao Mao, Chengruidong Zhang, Fan Zhou, Hao Luo, Haofeng Huang, Haoran Lian, Haoyan Huang, Hongqing Chen, Jianwei Zhang, Jing Xu, Junjie Wang, Langshi Chen, Liangyu Wang, Linlang Jiang, Man Yuan, Minmin Sun, Peng Jin, Siqi Zhang, Siyu Wang, Xingzhang Ren, Yakai Wang, Yi Zhang, Yiming Dong, Yizhong Cao, Yubo Ma, Yunfei Mao, Bo Zheng, and Dayiheng Liu. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability, 2026.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. arXiv:2303.11366.

Yifan Song, Da Yin, Xiang Yue, Jie Huang, Sujian Li, and Bill Yuchen Lin. Trial and error: Exploration-based trajectory optimization of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 7584–7600, 2024. arXiv:2403.02502.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Volume 1: Long Papers*, pp. 7080–7106, 2026. arXiv:2504.07952.

Chenwei Tang, Lin Long, Xinyu Liu, Jingyu Xing, Zizhou Wang, Joey Tianyi Zhou, Jiawei Du, Liangli Zhen, and Jiancheng Lv. SAG-Agent: Enabling long-horizon reasoning in strategy games via dynamic knowledge graphs, 2025. Version v3, revised March 2026.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research (TMLR)*, 2024. arXiv:2305.16291.

Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 63897–63911, 2025. arXiv:2409.07429.

Yue Wu, Yewen Fan, Paul Pu Liang, Amos Azaria, Yuanzhi Li, and Tom M. Mitchell. Read and reap the rewards: Learning to play atari with the help of instruction manuals. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023a. arXiv:2302.04449.

Yue Wu, Shrimai Prabhumoye, So Yeon Min, Yonatan Bisk, Ruslan Salakhutdinov, Amos Azaria, Tom Mitchell, and Yuanzhi Li. SPRING: Studying papers and reasoning to play games. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023b. arXiv:2305.15486.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior, 2025.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.12110.

Weiran Yao, Shelby Heinecke, Juan Carlos Niebles, Zhiwei Liu, Yihao Feng, Le Xue, Rithesh Murthy, Zeyuan Chen, Jianguo Zhang, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. Retroformer: Retrospective large language agents with policy gradient optimization. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2308.02151.

Qizheng Zhang, Changran Hu, Shubhangi Upasani, Boyuan Ma, Fenglu Hong, Vamsidhar Kamanuru, Jay Rainton, Chen Wu, Mengmeng Ji, Hanchen Li, Urmish Thakker, James Zou, and Kunle Olukotun. Agentic context engineering: Evolving contexts for self-improving language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.04618.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. Agent-pro: Learning to evolve via policy-level reflection and optimization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pp. 5348–5375, 2024a. arXiv:2402.17574.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024b.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024. arXiv:2308.10144.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning, acting, and planning in language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62138–62160, 2024. arXiv:2310.04406.

A Protocol details

A.1 What a decision looks like

The player makes every out-of-combat decision that has more than one sensible action. These are the map route, card rewards including the skip, shops, rest sites and which card to upgrade there, events, boss relics, and every card-selection screen. Combat is never the player's. The three screens below come from one game of the No memory arm, seed EXM0053, which the player cleared. Each box is the end of the user prompt for one decision: the screen type, a state summary, the screen header and the numbered options, in the pseudonyms the player saw. The part above the screen is 6,364 to 6,829 characters long. It holds the plan the player wrote at an earlier decision, its last eight actions and the report of the last fight, which stays until the player next picks a route. A REFERENCE block then gives the rules text of every card, relic and potion in view. We elide all of it as `[...]`, together with repeated state lines in the second and third box, which also stop at the options. Long lines are wrapped. A memory arm adds an EXPERIENCE block at the top of the prompt and a note on it to the system prompt. Under each box are the CHOICE line of the reply and its rationale, with the true name of the chosen card in brackets. The reply also carries a PLAN line, which the next prompt shows back as the player's plan.

Card reward. The screen follows a won fight on floor 22, in act 2. Three cards are offered, and option 3 skips them.

```
[...]
CURRENT SCREEN: REWARDS
Floor 22 | Act 2 | Ascension 10 | Act boss: Champ | Rooms until the act 2 boss: 10
HP 70/80 | Gold 44
Potions (3 slots): [0] Elixir Potion
Relics: Rokfial, Saitrouth, Ruboath, Skiskiol, Skixplou, Kruzosh, Staidroum, Tritrur
Deck (28 cards): Tigros (curse, cannot be removed), Zoplun, Stagrem, Kliopoush, Kladael
    x3, Thoaxloa, Kroabran x2, Baigath, Kloastir, Voulamoal, Voulamoal+, Greition,
    Stiokles x2, Ramkuth+, Toskaen, Maezae x5, Nioser, Roapol x2, Kuthfoar

Rewards: Select Reward to Take.
OPTIONS:
  [0] take card Lokthiosh
  [1] take card Thaisouk
  [2] take card Tremkraix
  [3] skip rewards

Reply with `CHOICE: <index>` on the first line.
```

`CHOICE: 1`, which takes Thaisouk [Metallicize]. "Thaisouk adds passive end-of-turn block, which is exactly what the deck lacks for the Champ's heavy multi-hits, and it costs no card slots in hand during critical turns."

Table 3: Settings identical in every DeepSeek Flash exam game.

Setting	Value
Player model	DeepSeek Flash
Character, difficulty	Ironclad, Ascension 10. The run ends at the act-3 boss
Out-of-combat decisions	every one with more than one sensible action, made by the player
Combat	fixed omniscient UCB search, 50,000 simulations per search, three times that in boss fights, each search followed by up to 15 actions of the line it found
Temperature	0.3
Player max output tokens	16,000, fallback 1,500
Player reasoning effort	not set, so the provider default applies
Max LLM calls per game	300
Card and relic pseudonyms	on, renaming seed 20260912
Simulator binary	the vendored `sts_lightspeed` build, md5 `da52f59b1fdfd412c34de78ed689387a`

Shop. The player entered the shop on floor 21 with 316 gold. The shop stays open after a purchase, and the player made five purchases here, one per decision. The box shows the first.

```
[...]
CURRENT SCREEN: SHOP
Floor 21 | Act 2 | Ascension 10 | Act boss: Champ | Rooms until the act 2 boss: 11
HP 80/80 | Gold 316
[...]

Shop Room: Select Item to Purchase.
[...]
OPTIONS:
  [0] Buy card Nioser 51g
  [1] Buy card Voulamoal 49g
  [2] Buy card Kroabran 33g
  [3] Buy card Stagrem 80g
  [4] Buy card Budouk 76g
  [5] Buy card Dernoal 91g
  [6] Buy card Shoamzoul 177g
  [7] Buy relic Pugrai 304g
  [8] Buy relic Droklosh 261g
  [9] Buy relic Moplour 149g
  [10] Remove Card 75g (curses marked "cannot be removed" are not eligible: Tigros is your
      only curse and cannot be removed)
  [11] proceed
```

`CHOICE: 1`, which buys Voulamoal [Perfected Strike] for 49 gold. "Buying the second Voulamoal both adds a strong Maezae-family body and boosts our upgraded Voulamoal's damage (6+3 per family card), which is our main Champ-killer."

Event. The Augmenter event on floor 27, in act 2, offers a card, a transformation of two cards in the deck, or a relic.

```
[...]
CURRENT SCREEN: EVENT
Floor 27 | Act 2 | Ascension 10 | Act boss: Champ | Rooms until the act 2 boss: 5
HP 70/80 | Gold 44
Potions (5 slots): [0] Elixir Potion
Relics: Rokfial, Saitrouth, Ruboath, Skiskiol, Skixplou, Kruzosh, Staidroum, Tritrur,
    Stoanaesh, Traikraix (0/6 turns)
[...]

Event - Augmenter: Select Action.
OPTIONS:
  [0] [Test Kaidriosh] Obtain a Kaidriosh.
  [1] [Become Test Subject] Choose and Transform 2 cards in your deck.
  [2] [Ingest Mutagens] Obtain Piomex.
```

`CHOICE: 0`, which obtains Kaidriosh [J.A.X.]. "Kaidriosh is a 0-cost Strength source that synergizes with Thoaxloa+, Voulamoal, Tritrur, and can even help trigger Skiskiol by dropping us below half HP."

A.2 Settings held constant in every arm

The simulator is `sts_lightspeed`, an open-source C++ reimplementation of the game, vendored at upstream commit `7476a81` and driven by our Python harness. The harness, not the player, resolves every screen that has a single sensible action. These are gold, a relic among a fight's

rewards, opening a chest, a forced single map step and a lone proceed. A potion is taken too when a slot is free and the player does not hold Sozu, a relic that blocks potions.

A.3 HARNESS DESIGN

Each choice below keeps the memory the only thing an arm changes, or keeps every analysis computable from the logs.

Decisions and fights. The player never writes a command. The harness parses the simulator's legal actions into numbered options and the player answers with an index, so every action it takes is legal. The screens listed above are resolved by the harness and logged as harness actions, so every logged player decision is a real choice. Fights never reach the player. The executor searches a copy of the true battle state, draw order and random generators included, and reads no memory, so a memory reaches the outcome only through the player's out-of-combat decisions. The search is seeded from the game seed and the floor, so a game is a deterministic function of its seed and the player's choices.

One variable and the plan. At a given game state the prompts of two arms differ only in the EXPERIENCE block, built by the rendering function of Appendix A.6, and the system-prompt sentence that introduces it. The only text the player carries from one decision to the next is the PLAN line that ends each reply, capped at 600 characters. The plan, like the rest of the loop's state, is reset at the start of every game, so the memory is the only channel between games. Printed below the memory, the plan could pass for experience, so an end marker closes the memory and the plan's heading says it "records your intent, not anything you learned".

Freezing. The exam scheduler plays each block in its own directory from a frozen copy of one memory, whose sha256 the manifest records with the player model and the settings of Table 3. A directory holding results from another memory, player or setting is refused. Each decision record also carries the sha256 of its user prompt, system prompt and memory block. An exam calls no writer and its memory is frozen before its first game, so no game of a block can reach the writer of the memory the block scores.

Valid games and the score. A game counts only if it ends in a victory or a defeat. A game that stops on a screen with no parsable option, a simulator crash, a client failure or the call cap of Table 3 is not a game. It has no score and enters no mean and no memory update. One function scores every game: the highest floor reached for a defeat and 51 for a victory, whatever floor the log shows. An unparsable reply is asked once more, then replaced by a flagged random legal option. In the 370 DeepSeek Flash exam games 19 replies were asked again, and none fell back to a random option.

What a game leaves on disk. Each decision record holds the screen, state and options as the player saw them and who chose what. A player decision adds the plan shown and carried forward, the card-table rows shown, the reply, the reasoning and the token usage. Each fight adds the executor's report, and an atomically written summary holds the outcome, score and settings of the game. The card-take and upgrade audits read these records directly. The common-state replay of Section 4.2 sends the logged commands back through the simulator, rebuilds each prompt, checks it against the logged hash and then swaps only the memory block.

What a writer reads. The writers' material is built from the same logs. The game summary turns one game into a floor-by-floor event log of offers, picks, shop, rest and event choices and fight results. It ends with the final deck and relics, the executor's play count per card and where each owned card came from. The sequential writer reads one summary after each game, with the rules text of the cards it names. The batch writers read all their summaries at once. Both also receive the merged ledger, the harness's per-card counts summed over every game read so far, such as offers, takes, plays and clears with the card in the deck. The ledger has no field for the act or for copies of a card, and the player never sees it. The Teacher writer read its games in true names, mapped back through the inverse of the renaming map.

Table 4: Protocol summary per arm, the full version of Table 1. Material names the source and the number of games behind it. The 20 games behind Self-seq, Self-batch-20 and Self-batch-20 (Opus) are the learning games of Self-seq. The first was played without memory and each later one with the notebook written from the games before it. Self-batch-20 thus differs from Self-batch in how many games it reads and in how they were played. Eff. is the writer's reasoning effort and n.r. means no artifact records it. Names is what the writer read; the few true names a pseudonym writer wrote were masked at rendering (Appendix A.9). FC is a source-code fact-check of the memory. Human is the human's part in the memory beyond starting the run that wrote it. A dash means none, and approval is approving the finished notebook without editing it. Review and sign-off are the same for the Expert draft. Build script is writing the program that computes the Statistics notebook, and the brief and 6 guidelines are the two inputs of Appendix A.7. No human edited any memory. Chars is the player-facing notebook length measured by the exam's own rendering function, Rows the card-table rows. Exam games counts the games the arm was scored on.

Arm	Material	Writer	Eff.	Names	FC	Human	Chars	Rows	Exam games
No memory	no memory	–	–	–	–	–	0	0	50
Self-batch	own 50 None games, one bundle	DeepSeek Flash (the player)	max	pseudonym	no	–	5137	126	50
Self-seq	own 20 games, one at a time	DeepSeek Flash (the player)	default	pseudonym	no	–	7344	125	20
Self-batch-20	one bundle of those 20 games	DeepSeek Flash	max	pseudonym	no	approval	6469	125	50
Self-batch-20 (Opus)	the same bundle, byte-identical	Claude Opus 5	max	pseudonym	no	approval	7894	125	50
Statistics	25,000 randomized simulator games	program	n/a	true	n/a	build script	2474	147	50
Expert	prior knowledge, 0 games	Claude Opus 5	n.r.	true	yes	review, sign-off	6473	133	50
Teacher	50 None games + brief + 6 guidelines	Claude Fable 5.1 + 9 sub-agents	n.r.	true	partial	brief, 6 guidelines	6484	125	50

A.4 Memory provenance per arm

The Self-batch-20 and Self-batch-20 (Opus) writers read one bundle of 437,188 characters of user prompt and 6,098 of system prompt. Self-batch-20 wrote from it in one notebook call and 16 card-table calls, and Self-batch-20 (Opus) in a sandboxed command-line session with no tools. The Self-batch writer read a bundle of its 50 No memory games, built by the same pipeline, with 892,862 characters of user prompt and the same system prompt. It also wrote in one notebook call and 16 card-table calls. The Statistics program computes its numbers from 10,000 uniform-random simulator games and three rounds of 5,000 epsilon-greedy ones. It uses the 168,437 of their 346,805 card-reward offers at which the pick was random, runs with seed 0 and makes no model call. The Expert fact-check corrected 17 card-table rows and 5 numbered rules against the card, relic and potion data and the simulator source. The Teacher check found every item name of its memory in the true-name reference. The Teacher writer was not a single model. A Claude Fable 5.1 main session authored and revised the memory with 9 sub-agents, 7 Claude Opus 5 trajectory readers, 1 Claude Sonnet statistics agent and 1 Claude Opus 5 card-table agent.

A.5 Exam blocks and outcomes

Six arms draw their exam seeds from one fixed EXM family. Teacher draws its seeds from a disjoint TCH family. Self-batch draws its seeds from the TCH family. The 50 No memory games that the Self-batch writer read are EXM games. Self-batch is therefore the one arm without the teacher memory that was measured on the TCH family, and it cleared 15 of 50 there, one more than No memory on the EXM family.

A seed fixes the map and the act bosses of a run. What follows depends on the choices made, so two runs on one seed part at the first choice on which they differ, and their outcomes are nearly independent: on the 48 EXM seeds that No memory, Self-batch-20, Self-batch-20 (Opus) and Expert share, the phi coefficient between each arm's win indicator and that of No memory is 0.09, 0.19 and 0.18. Pairing by seed would remove almost no variance, so every test in the paper is unpaired. The Teacher exams of all three players ran on the same 50 TCH seeds, from a family the Teacher writer never read. The replay nodes of Section 4.2 come from 20 No memory games, 18 of them among those the Self-batch and Teacher writers read (Appendix C), so the replay shows what a memory

Table 5: Every exam block, grouped by arm and in play order within each arm. Mean score follows the scoring rule of Section 3.4 and is recomputed from the per-game summary files. Dates are local time.

Arm	Exam block	n	Clears	Mean score	Date
No memory	EXM0001–0010	10	2	38.10	2026-09-14
No memory	EXM0011–0020	8	5	45.25	2026-09-14
No memory	EXM0021–0030, 0051–0072	32	7	38.78	2026-09-22
Self-seq	EXM0001–0010	10	4	40.00	2026-09-14
Self-seq	EXM0011–0020	10	2	36.90	2026-09-14
Self-batch-20	EXM0001–0020	20	6	43.60	2026-09-15
Self-batch-20	EXM0021–0030	10	2	35.20	2026-09-21
Self-batch-20	EXM0051–0070	20	8	43.25	2026-09-22
Self-batch-20 (Opus)	EXM0001–0020	20	10	44.25	2026-09-15
Self-batch-20 (Opus)	EXM0021–0030	10	1	37.10	2026-09-21
Self-batch-20 (Opus)	EXM0051–0070	20	4	40.85	2026-09-22
Statistics	EXM0001–0020	20	12	44.45	2026-09-14
Statistics	EXM0021–0030, 0051–0070	30	14	45.13	2026-09-24
Expert	EXM0001–0010	10	7	48.10	2026-09-14
Expert	EXM0011–0020	10	7	47.30	2026-09-14
Expert	EXM0021–0030	10	3	39.20	2026-09-21
Expert	EXM0051–0070	20	9	43.90	2026-09-22
Teacher	TCH20260922E0001–0050	50	30	46.48	2026-09-22
Self-batch	TCH20260922E0001–0050	50	15	43.32	2026-09-25

Table 6: Outcomes of the eight DeepSeek Flash arms. The three ablations of the self memory follow Self-batch. Each arm is scored on its full set of valid exam games and every comparison is unpaired. Clear rate carries the Wilson 95% interval. A cleared game scores 51. Act columns count games clearing that act. Fisher is the two-sided exact test on the clear rate against None. The Mann–Whitney U test on the score is in Table 10.

Arm	n	clears	clear rate	Wilson 95%	mean score	act 1 / 2 / 3	Fisher p
None	50	14	0.28	0.17–0.42	39.68	49 / 29 / 14	–
Self-batch	50	15	0.30	0.19–0.44	43.32	47 / 35 / 15	1.00
Self-seq	20	6	0.30	0.15–0.52	38.45	19 / 10 / 6	1.00
Self-batch-20	50	16	0.32	0.21–0.46	41.78	48 / 31 / 16	0.83
Self-batch-20 (Opus)	50	15	0.30	0.19–0.44	41.46	49 / 29 / 15	1.00
Statistics	50	26	0.52	0.39–0.65	44.86	49 / 37 / 26	0.024
Expert	50	26	0.52	0.39–0.65	44.48	50 / 34 / 26	0.024
Teacher	50	30	0.60	0.46–0.72	46.48	50 / 41 / 30	0.0023

does at states its writer saw. The No memory directories hold 52 games. Two of them, on EXM0011 and EXM0012, stopped on a screen with no parsable option, so they are not games under the rule of Appendix A.3, and the EXM0011–0020 block of No memory holds 8.

A.6 Memory rendering and sizes

The notebook is free text and sits above every decision of a run. The card table holds one row per card, written as `card | what you make of it`. The harness adds a never-played tag and an "as of run N" stamp. One function renders memory for every arm: the notebook whole, then the table rows for the cards on screen or in the deck. No arm shows the player harness statistics. The form is the same in every arm, and the length is not.

Rendering changes a notebook's length only by renaming, except for Self-batch-20 (Opus), where the never-claim strip of Appendix A.9 removed one sentence of 103 characters.

A.7 The human input to the teacher

The human gave the Teacher writer two things, a task brief and six guidelines, and wrote nothing else. The brief states the task: raise the clear rate of this player on unseen seeds, using the supplied 50 trajectories and any prior knowledge the writer has. It then gives a diagnostic checklist, with act-by-act examples of the kind of failure to look for. It sets the output format: a diagnosis file, a

notebook under 6,500 player-facing characters, a 125-row card table and a self-check file. The brief asked the writer to weigh each guideline against the trajectories and the actual executor, and to write a suitable condition of application rather than repeat a preference as an established result.

The six guidelines are the prior knowledge the human put in. They are rendered here in English, and the original text is released with the code.

1. Runic Dome has high priority, because the UCB executor is not affected by hidden enemy intent, so the relic has no downside.
2. Take more rest sites and fight fewer elites. Elites usually pay little and cost a lot of health. In act 2 take more unknown rooms and fewer monster rooms.
3. In act 3 fight as little as possible. Keeping health is the first priority.
4. Cards that can be taken almost without thinking: Shrug It Off, an upgraded Pommel Strike, Offering.
5. Almost never upgrade Strike or Defend. Upgrade the cards that gain most from an upgrade, such as Armaments.
6. Do not be too afraid of cards with a random effect, such as Havoc and True Grit. For the UCB executor they are not a problem.

Some notebook lines cite the checklist part of the brief, and one overrules it. Line 44 of the brief names Carnage, Hemokinesis and Bludgeon as act-1 examples and says not to take them on sight. The notebook puts Hemokinesis and Bludgeon on its take-on-sight list.

The human did not edit the memory. The three trims that brought the draft to its budget were made by the teacher model itself.

A.8 PSEUDONYMIZATION

Every card and relic name is replaced by a seeded pseudonym drawn from a generated word list. One renaming map serves every run. It is a function of a fixed seed and the sorted list of names, and a name added later is appended so that no earlier pseudonym moves. Names are matched against the simulator's own display strings, longest name first, so a name that contains another name is renamed as a whole. Plural and possessive forms carry the pseudonym, and a card or relic name inside an enemy or event name is renamed like any other. Enemy intent lines are left alone, because enemy moves share names with cards and would read as the player's own cards. The REFERENCE block supplies the rules text of every card, relic and potion in view, so the player reads what an item does but never its public name.

The same map renames whatever a writer returns, so a true name that a writer adds from its own knowledge never reaches the player. It also renames a memory written in true names before the player reads it. A card thus has one pseudonym in every log, memory and prompt, so one memory file can serve three players.

What the names carry. Before the main study we ran the player without memory on 10 paired seeds in three arms of 10 games each: real names without the REFERENCE block, real names with it, and pseudonyms with it. The pilot used an earlier version of the prompt. Its arm without the block ran with an output budget of 32,000 tokens and its two arms with the block with about 3,000, against 16,000 in the main study. The three arms clear the same number of games, 3 of 10 each, and reach similar mean scores of 41.2, 37.9 and 42.7. The card picks differ (Figure 5). Under real names the player reaches for cards that human players rate highly, such as Immolate and Flame Barrier, and avoids Anger, Rage and Berserk. Under pseudonyms those preferences vanish. Across the two real-name arms the player takes Anger once in 52 offers, against 12 times in 25 under pseudonyms ($p < 0.001$), and Rage in none of 11 offers, against all 5 under pseudonyms. The names carry a prior that moves card choices, and the renaming removes it, while the outcome under the executor did not move in this pilot. One game of the arm without the block, on seed CAL0008, was rerun after a simulator crash once the block existed, and it carries the block. Without that game the arm clears 2 of 9 games at a mean score of 40.1, and the two real-name arms take Anger once in 49 offers.

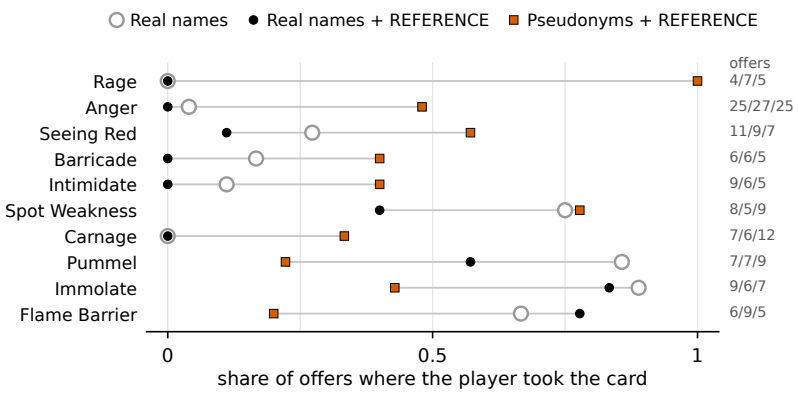


Figure 5: Per-card take rate at the card-reward screen in the name pilot, 10 games per arm on the same 10 seeds, no memory. The rows are the 10 cards whose take rate moves most between real names and pseudonyms, both with the REFERENCE block. They are drawn from the cards offered at least 5 times in each of those two arms and sorted by that shift. Offers without the block, with it and under pseudonyms are printed at the right. A dot inside a ring marks equal rates in the two real-name arms. Names are the true card names. The pseudonym arm saw only pseudonyms.

A.9 THE POST-STRIP STEP

The post-strip step runs four operations in order on an incoming notebook: the pseudonym mask, the ledger's veto on a claim that a card was never played when the ledger says it was, the lexical never-claim strip, and the prompt-example strip. Each card-table row gets the same treatment plus a 70-word cap. The writer prompt asks for about 6,500 characters. The notebook is then capped at 8,000 characters by dropping pasted non-notebook blocks, then whole entries oldest first, and only then trimming rule lines down to a floor. The Teacher memory did not go through that path. The renaming was applied offline, the alias text was checked to reproduce the true text, and the frozen memory file was written directly. Its notebook and table are the only external memory in the study that the post-strip never touched. Run afterwards on the Teacher source files, the step strips no sentence and no row, and its output equals the frozen memory apart from one final newline.

A.10 THE OTHER TWO PLAYERS

Table 8 lists how Qwen3.8-Flash-Next and Ling-3.0-flash were served and run. Both were served by vLLM on local GPUs. Every one of the 500 games ran with the settings of Table 3 except the player model. No player request carries a presence penalty, and no game fell back to a random action.

Table 7 gives the clears by act, the mean scores and the tests against No memory for the five arms of Figure 2. Appendix A.5 states which seeds each arm ran on and why the tests are unpaired.

A player request that spends its whole 16,000-token budget on reasoning is asked once more with reasoning off and a 1,500-token budget, with the tail of its reasoning appended. The row "Requests at limit" of Table 8 counts the requests that ended at the token limit. They are under 5% of Qwen3.8-Flash-Next's requests in every arm and at most one per arm for Ling-3.0-flash.

The Qwen3.8-Flash-Next writer first ran with no penalty, as the other writers did. It spent the whole 64,000-token budget on reasoning in its notebook call and in 12 of its 16 card-table calls, so most of that memory was the text of reasoning-off retries. That memory was examined on the same 50 seeds before the fault was found and cleared 3 of 50 games. A second attempt with a budget of 110,000 tokens spent it all on each of the first two card-table calls and was stopped. A probe that re-sent the first card-table request found a loop: the writer drafted all eight entries within its first 8,000 reasoning tokens and then repeated the same checks until the probe's 30,000-token limit. With a presence penalty of 1.5 and nothing else changed, the same request ended its reasoning at less than half that length and returned a valid table.

The Qwen3.8-Flash-Next Self-batch memory of Table 7 was written with that penalty, and none of its 17 writer calls reached the budget. The Ling-3.0-flash writer and the DeepSeek Flash writers of Self-batch and Self-batch-20 ran without a penalty and never reached it either.

Table 7: Five arms, three players. DeepSeek Flash rows repeat Table 6. Clears are counted over the arm's valid exam games, and act 1 / 2 / 3 counts the games that clear each act. Fisher is the two-sided exact test on the clear rate and Mann–Whitney (MWU) the two-sided U test on the score, both unpaired and against the same player's None arm. For Qwen3.8-Flash-Next and Ling-3.0-flash, None, Statistics and Expert ran on the 50 seeds of the DeepSeek Flash None arm, Self-batch on EXM0023–0072 and Teacher on the TCH family (Table 8). The DeepSeek Flash seeds of every block are listed in Table 5, and DeepSeek Flash Self-batch and Teacher also ran on the TCH family.

Player	Arm	clears	act 1 / 2 / 3	mean score	Fisher p	MWU p
DeepSeek Flash	None	14/50	49 / 29 / 14	39.68	–	–
	Self-batch	15/50	47 / 35 / 15	43.32	1.00	0.19
	Statistics	26/50	49 / 37 / 26	44.86	0.024	0.0081
	Expert	26/50	50 / 34 / 26	44.48	0.024	0.012
	Teacher	30/50	50 / 41 / 30	46.48	0.0023	0.00051
Qwen3.8-Flash-Next	None	7/50	49 / 20 / 7	37.68	–	–
	Self-batch	5/50	44 / 14 / 5	32.90	0.76	0.035
	Statistics	14/50	49 / 33 / 14	43.14	0.14	0.0047
	Expert	8/50	47 / 19 / 8	37.16	1.00	0.98
	Teacher	18/50	47 / 30 / 18	41.64	0.020	0.013
Ling-3.0-flash	None	1/50	44 / 9 / 1	29.36	–	–
	Self-batch	0/50	40 / 2 / 0	25.04	1.00	0.0076
	Statistics	4/50	50 / 22 / 4	37.66	0.36	0.000048
	Expert	6/50	49 / 17 / 6	35.62	0.11	0.0016
	Teacher	10/50	47 / 25 / 10	38.10	0.0078	0.000040

Table 8: Serving and run settings of Qwen3.8-Flash-Next and Ling-3.0-flash. Thinking is each model's default, on, as for DeepSeek Flash, whose reasoning effort is left at the provider default. Rendered memory is the length of the notebook and every card-table row rendered as one block, the most the player can be shown. A decision shows the notebook and only the rows of the cards in view.

	Qwen3.8-Flash-Next	Ling-3.0-flash
Checkpoint	`VnimanieAI/Qwen3.8-Flash-Next-W4A16`	`inclusionAI/Ling-3.0-flash-int4`
Quantization	int4 weights, 16-bit activations, group 128	int4 weights, 16-bit activations, group 32
Server	vLLM 0.30.0, tensor parallel 4, 4 NVIDIA L40	vLLM 0.30.0, tensor parallel 4, 4 RTX 4090 (48 GB)
Max model length	40,960 for None, 262,144 for the writer and the other arms	262,144
Self-batch material	the player's own None games on EXM0001–0010 and EXM0013–0022	
Exam seeds	None, Statistics, Expert: the 50 seeds of the DeepSeek Flash None arm, EXM0001–0010, 0013–0030, 0051–0072. Teacher: TCH20260922E0001–0050, the seeds of the DeepSeek Flash Teacher arm. Self-batch: EXM0023–0072 (50), disjoint from its material	
Self-batch writer	the player model, with the bundle pipeline and prompts of the DeepSeek Flash Self-batch-20 writer: 1 notebook call and 16 card-table calls, 64,000 output tokens per call, temperature 0.3	
Writer penalty	presence penalty 1.5	none
Self-batch notebook	4,275 characters, 124 card-table rows	2,404 characters, 122 card-table rows
Rendered memory	Self-batch 29,600 (Qwen3.8-Flash-Next) and 29,747 (Ling-3.0-flash) characters. Statistics 26,892, Expert 30,828, Teacher 33,206 for both players	
Requests at limit	None 12/3,300, Self-batch 101/2,935, Statistics 177/3,979, Expert 83/3,284, Teacher 23/3,641	None 1/2,455, Self-batch 1/2,283, Statistics 0/3,247, Expert 1/2,978, Teacher 0/3,258

The Statistics, Expert and Teacher files were built for DeepSeek Flash and name it as their player. For the other two players the harness therefore introduces their notebook as "written by an earlier player of this same setting, after their own runs", and adds the matching note to the system prompt. Self-batch is written by the player itself and is rendered in the first person, as every memory is for DeepSeek Flash. The three files are byte-identical to the ones the DeepSeek Flash arms read, checked by sha256 at every launch, and none of them was written for, or from the games of, Qwen3.8-Flash-Next and Ling-3.0-flash. The memory files and their sha256 are listed in `analysis/multiplayer_20260923/replication_table.md`.

Table 9: Two-sided Fisher exact p on the clear rate for every pair of arms. Denominators are 50 for seven arms and 20 for Self-seq.

	Self-batch	Self-seq	Self-batch-20	Self-batch-20 (Opus)	Statistics	Expert	Teacher
None	1.00	1.00	0.83	1.00	0.024	0.024	0.0023
Self-batch	–	1.00	1.00	1.00	0.041	0.041	0.0046
Self-seq	–	–	1.00	1.00	0.12	0.12	0.034
Self-batch-20	–	–	–	1.00	0.068	0.068	0.0088
Self-batch-20 (Opus)	–	–	–	–	0.041	0.041	0.0046
Statistics	–	–	–	–	–	1.00	0.55
Expert	–	–	–	–	–	–	0.55

Table 10: Two-sided Mann–Whitney U p on the score for every pair of arms.

	Self-batch	Self-seq	Self-batch-20	Self-batch-20 (Opus)	Statistics	Expert	Teacher
None	0.19	0.77	0.39	0.53	0.0081	0.012	0.00051
Self-batch	–	0.20	0.60	0.48	0.10	0.17	0.010
Self-seq	–	–	0.36	0.46	0.030	0.039	0.0053
Self-batch-20	–	–	–	0.84	0.051	0.080	0.0050
Self-batch-20 (Opus)	–	–	–	–	0.033	0.047	0.0027
Statistics	–	–	–	–	–	0.89	0.40
Expert	–	–	–	–	–	–	0.33

B STATISTICAL DETAIL

This appendix holds the tests behind Sections 4.1 and 4.4. Every arm is scored on its full set of valid exam games and every comparison is unpaired. Fisher is the two-sided exact test on the clear rate and Mann–Whitney the two-sided U test on the score. Per-arm totals are in Table 6 and the per-block totals behind them in Table 5.

B.1 EVERY PAIR OF ARMS

Tables 9 and 10 give all 28 pairwise tests, on the clear rate and on the score. The two endpoints disagree on four pairs at the 0.05 level, two in each direction. Self-seq against Statistics and against Expert is $p = 0.12$ on the clear rate and $p = 0.030$ and $p = 0.039$ on the score. Self-seq carries the lowest mean score of the eight arms, 38.45, below No memory at 39.68, while its clear rate matches No memory. Self-batch against Statistics and against Expert goes the other way, $p = 0.041$ on the clear rate and $p = 0.10$ and $p = 0.17$ on the score. Self-batch takes 35 games into act 3, near Statistics and Expert at 37 and 34, but clears 15 of them against 26 for each. Of the 20 games it loses in act 3, 17 end at the act-3 boss on floor 50, which the score places one floor below a clear.

B.2 MULTIPLE COMPARISONS

The seven comparisons of a memory arm against No memory form one family. Holm adjustment over that family gives $p = 0.016$ for Teacher and $p = 0.15$ for both Statistics and Expert. The other four stay at 1.00. Teacher is the one arm below 0.05 after the correction.

B.3 THE OTHER TWO PLAYERS

For each of Qwen3.8-Flash-Next and Ling-3.0-flash the four memory arms against No memory form one family. Holm adjustment gives $p = 0.079$ for Teacher with Qwen3.8-Flash-Next, against 0.42 for Statistics and 1.00 for Self-batch and Expert, so no Qwen3.8-Flash-Next arm stays below 0.05 after the correction. With Ling-3.0-flash it gives $p = 0.031$ for Teacher, against 0.34 for Expert and above 0.7 for Statistics and Self-batch. Teacher against Statistics is 18 against 14 of 50 with Qwen3.8-Flash-Next, $p = 0.52$ on the clear rate and $p = 0.98$ on the score. With Ling-3.0-flash it is 10 against 4, $p = 0.15$ and $p = 0.55$. Teacher against Expert on the score is $p = 0.017$ with Qwen3.8-Flash-Next and $p = 0.17$ with Ling-3.0-flash. Between players under Teacher, Fisher's exact test on the clear rate gives $p = 0.027$ for DeepSeek Flash against Qwen3.8-Flash-Next, $p =$

Table 11: Common-state replay, all quantities. 148 nodes, 10 samples per node per condition. Brackets are 95% intervals from resampling nodes. Self-seq is the sequential self notebook, and Self-batch and Teacher read the frozen memory files of their exams. Take rate is the share of samples that take a card. Modal flip is the share of nodes whose most frequent choice differs from that of None, and take/skip flip the share where one of the two takes a card and the other skips. Agreement is the share of samples that match the most frequent None choice at their node, so the None entry is its agreement with itself. Rule-citation rate is the share of replies that cite a numbered rule. Decision entropy is the entropy in bits of the choice over the 10 samples of a node, averaged over nodes.

	None	Self-batch	Self-seq	Expert	Statistics	Teacher
Take rate, all nodes	0.901 [0.861, 0.936]	0.925 [0.891, 0.955]	0.901 [0.859, 0.936]	0.722 [0.662, 0.776]	0.961 [0.934, 0.984]	0.856 [0.806, 0.903]
Take rate, deck $<$20	0.996 [0.991, 1.000]	0.993 [0.980, 1.000]	0.979 [0.957, 0.995]	0.979 [0.959, 0.993]	1.000 [1.000, 1.000]	0.945 [0.896, 0.982]
Take rate, deck 20–29	0.919 [0.860, 0.965]	0.950 [0.902, 0.987]	0.923 [0.860, 0.975]	0.692 [0.598, 0.777]	0.950 [0.894, 0.992]	0.890 [0.812, 0.958]
Take rate, deck $\geq$30	0.745 [0.637, 0.850]	0.797 [0.693, 0.887]	0.762 [0.650, 0.860]	0.400 [0.295, 0.512]	0.923 [0.853, 0.978]	0.687 [0.565, 0.807]
Take/skip flip vs None	0.000	0.054 [0.020, 0.095]	0.088 [0.047, 0.135]	0.189 [0.128, 0.257]	0.068 [0.027, 0.108]	0.122 [0.074, 0.176]
Modal flip vs None	0.000	0.365 [0.291, 0.446]	0.358 [0.284, 0.432]	0.405 [0.324, 0.486]	0.405 [0.324, 0.486]	0.480 [0.399, 0.561]
Agreement with None	0.786 [0.755, 0.816]	0.616 [0.551, 0.678]	0.614 [0.554, 0.670]	0.562 [0.500, 0.624]	0.595 [0.529, 0.662]	0.498 [0.428, 0.568]
Rule-citation rate	0.000	0.001 [0.000, 0.002]	0.003 [0.000, 0.008]	0.036 [0.024, 0.049]	0.001 [0.000, 0.002]	0.000
Decision entropy	0.682 [0.598, 0.767]	0.469 [0.391, 0.549]	0.560 [0.479, 0.641]	0.636 [0.553, 0.718]	0.394 [0.319, 0.470]	0.386 [0.309, 0.465]
Reply chars (mean)	422.3	430.2	428.8	447.9	437.2	425.4
Reasoning chars (mean)	3485.5	3051.3	4058.8	4086.0	3155.5	2882.9

Table 12: Two replayed nodes, 10 samples per condition.

Node	Option	None	Self-batch	Self-seq	Expert	Statistics	Teacher
EXM0019:117	take Sever Soul+	5	1	0	1	10	0
	skip rewards	5	9	10	9	0	10
EXM0007:113	take Second Wind	1	0	0	0	0	0
	take Anger	5	10	0	0	10	0
	take Cleave	1	0	9	0	0	6
	skip rewards	3	0	1	10	0	4

0.12 for Qwen3.8-Flash-Next against Ling-3.0-flash and $p = 0.0001$ for DeepSeek Flash against Ling-3.0-flash.

C Replay and upgrade audit

Full replay table. Table 11 gives every quantity of the common-state replay with its interval, plus the mean reply and reasoning length per condition. The 148 nodes are every third card-reward decision, ordered by seed and step, of the 20 No memory games on EXM0001–0020. Two of those games, EXM0011 and EXM0012, later stopped on a screen with no parsable option and are not among the arm's 50 exam games. The other 18 are among the 50 games that the Self-batch and Teacher writers read. Each node is replayed 10 times per condition, which is 1480 rows per condition and 8880 rows in total. Every interval is a percentile interval over resampled nodes. One of the 1480 Self-batch replies failed to parse, and no reply failed in the other five conditions.

Two node examples in detail. Two nodes show the three channels of Section 4.2 in a single state. Where a 30-card deck is offered one upgraded card, Self-batch, Self-seq, Expert and Teacher skip, No memory splits 5 to 5 and Statistics takes. Where a 30-card deck is offered three cards, Self-seq takes a different card from the one No memory, Self-batch and Statistics take, Expert skips, and Teacher splits between Self-seq's card and skipping. Table 12 tallies all 10 samples of each condition at these two nodes. Both nodes sit at deck size 30, the stratum where the conditions separate most.

Upgrade audit. The audit covers the 200 frozen exam games of Self-batch, Self-batch-20, Self-batch-20 (Opus) and Expert, plus the 50 No memory games that the Self-batch writer read. It counts explicit card-to-upgrade menus only and excludes automatic, random, event-all and temporary combat upgrades. It verified 1,028 raw upgrade records. Table 13 lists the five most frequent targets per arm, and Table 15 follows the Strike rules of Section 4.2 from the writers' input to the player's upgrades. Table 14 gives the card-take rate by act and how often each notebook makes the act, the deck size or HP a condition.

Table 13: Top five explicit upgrade targets per arm over 50 exam games each.

Self-batch (204 upgrades)	Self-batch-20 (221)	Self-batch-20 (Opus) (227)	Expert (175)
Perfected Strike 30	Anger 41	Perfected Strike 29	Perfected Strike 30
Anger 24	Strike 28	Heavy Blade 22	Heavy Blade 16
Heavy Blade 24	Pommel Strike 22	Demon Form 16	Shrug It Off 14
Demon Form 15	Shrug It Off 22	Bash 16	Demon Form 12
Pommel Strike 14	Heavy Blade 21	Pommel Strike 16	Inflame 10

Table 14: Act-dependent behaviour and act-dependent rules. Card-take rate is the percentage of card-reward offers at which the player took a card, over all valid exam games of the arm, and every cell holds at least 201 offers. Rule conditions count the notebook's rule units (its tagged rule sentences) that make the act, the deck size or HP a condition, out of all its rule units. None has no notebook, and the Self-batch and Teacher notebooks were not tagged.

	Card-take rate by act, %			Rule conditions		
Arm	act 1	act 2	act 3	act	deck	HP
None	98	93	80	–	–	–
Self-batch	98	92	83	–	–	–
Self-batch-20	98	94	86	2/50	4/50	8/50
Self-batch-20 (Opus)	96	85	68	3/52	8/52	13/52
Expert	91	70	52	15/50	9/50	15/50
Teacher	92	90	68	–	–	–

Table 15: The Strike chain, link by link, for the two batch memories that DeepSeek Flash wrote.

Link	Content
Writer input	The Self-batch-20 writer read a merged per-card ledger over its 20 source games. Strike was played 1,147 times there, rank 1 in 17 of 20 games, ahead of Defend at 885 and Anger at 330. The Self-batch writer read a ledger of the same form over its 50 No memory games, with Strike first at 2,011 plays, ahead of Defend at 1,777 and Bash at 649. The raw event logs carry act and copy information, and neither ledger has a per-act field.
Self-batch-20 rule	"Upgrade order: your most-played cheap attack (Strike, Twin Strike, Anger), then Shrug It Off or another Block card, then Heavy Blade or a Strength card."
Self-batch rule	The notebook calls Defend and Strike the executor's "most-played cards in nearly every run, Bash close behind". The card-table row for Strike says the five starter copies are "what the shell plays every turn, and Strength multiplies each one. Smith [upgrade] one early".
Leak check	"most-played", "busiest", "upgrade order", "cheap attack" and "Smith one early" each occur 0 times in the prompts either writer was given, so the prompts do not supply the rules' wording. "Notebook" occurs 9 times in each, where the prompt names the text to be written.
Player behaviour	Explicit upgrades targeting Strike: Self-batch 11 of 204 (10 in act 1, 1 in act 3, over 11 games), Self-batch-20 28 of 221 (23 in act 1, 4 in act 2, 1 in act 3, over 23 games), Self-batch-20 (Opus) 0 of 227, Expert 0 of 175. In the games the writers read, the player picked Strike once in 201 explicit upgrades over the 50 No memory games and 3 times in 92 over the 20 source games of Self-batch-20.
Self-seq notebooks	None of the 21 notebook snapshots of Self-seq, written one game at a time from the same 20 source games as Self-batch-20, states the upgrade-priority rule.

Table 16: The 21 checkable Teacher rules with their class and each player's compliance, Teacher with No memory in parentheses. DS, Qwen, Ling: DeepSeek Flash, Qwen3.8-Flash-Next, Ling-3.0-flash. The letter of a rule gives its finer class. An F rule names the answer outright, and a T rule sets one condition read off a number on the screen. An M rule sets several conditions that must hold together, and a G rule gives a target with a reason rather than a screen rule. F and T rules are simple, M and G rules compound.

	Rule	Class	DS	Qwen	Ling
F1	Skip list: never take a Skip-tier card	simple	.98 (.78)	.99 (.84)	.98 (.74)
F2	A-tier: take on sight (act 1)	simple	.93 (.61)	.93 (.62)	.87 (.60)
F3	Neow: boss relic swap first	simple	.98 (.38)	.98 (.00)	.98 (.00)
F4	Boss relics: avoid list not taken	simple	.96 (.66)	.95 (.76)	.95 (.80)
F5	Never upgrade Strike or Defend	simple	1.00 (.98)	1.00 (.95)	.95 (.55)
F6	Upgrade order: top listed card on the menu	simple	.62 (.26)	.74 (.28)	.58 (.07)
F7	Shop: Remove Card first when affordable	simple	.75 (.56)	.69 (.53)	.65 (.04)
T1	Rest above 80% HP: Smith	simple	.99 (.97)	.95 (.92)	.86 (.62)
T2	Rest below 50% HP: Rest	simple	1.00 (1.00)	1.00 (1.00)	1.00 (.92)
T3	Never an elite below 55% HP	simple	1.00 (.78)	.94 (.82)	1.00 (.94)
T4	Named events (HP gates, refuse, reach limits)	simple	.90 (.60)	.81 (.58)	.81 (.78)
T5	Conditional A-tier card not taken when its condition fails	simple	.79 (.77)	.84 (.75)	.82 (.67)
T6	From 30 cards: no non-A card	simple	.83 (.59)	.89 (.69)	.94 (.55)
M1	Rest at 50-80% HP: Smith only if no elite or boss within two floors and $<$ 1/3 upgraded	compound	.71 (.52)	.60 (.45)	.72 (.43)
M2	Elite taken if and only if HP $\geq$ 70%, damage source, Rest or shop before the next fight	compound	.85 (.70)	.74 (.64)	.71 (.63)
M3	Boss relic: best eligible option of the ranked list	compound	.71 (.43)	.65 (.35)	.56 (.23)
M4	From 30 cards: an A-tier card on offer is still taken	compound	.72 (.58)	.47 (.35)	.35 (.65)
M5	Enter the act 2 boss at $\geq$ 75% HP	compound	.89 (.72)	.78 (.84)	.85 (.73)
G1	Act 2 prep list: key damage card taken when offered (to floor 33)	compound	.64 (.48)	.56 (.32)	.48 (.37)
G2	Key damage card chosen over another A-tier card on the same screen	compound	.63 (.55)	.49 (.31)	.43 (.48)
G3	Elites are income: elite taken when all conditions hold	compound	.80 (.73)	.69 (.63)	.49 (.54)

D RULE CLASSES UNDER THE TEACHER MEMORY

Table 16 lists the 21 rules of the Teacher notebook whose compliance can be read from the logs, the class each was given, and each player's compliance under Teacher and under No memory. A simple rule names the answer outright or gates it on one number printed on the same screen. A compound rule needs several conditions to hold together, or gives a target and leaves the reader to pursue it over several screens. The classes were fixed before the compliance was computed. The difference of Section 4.4, the change from No memory to Teacher in the between-player gap for compound rules minus that for simple rules, is 0.17 (95% interval 0.08 to 0.25). Moving T5 and T6, the two most debatable rules, to compound gives 0.13 (0.05 to 0.21).

The convergence holds beyond the 16 cards of Figure 4. Over all 72 cards offered at least 8 times to every player in every arm, the mean between-player spread of the take rate is 24 points without memory, 15 under Expert and 12 under Teacher. The mean pairwise correlation is 0.46, 0.77 and 0.91. Under either memory, sampling noise alone leaves about 10 points between three players that share one true rate per card.

E TEACHER MEMORY DETAIL

The teacher's 11 self-reported findings. Table 17 lists all 11 findings the teacher raised against its own memory and the verdict of the independent check. Eight are confirmed. Section 4.3 uses findings 1, 2, 4, 8, 9 and 11, and finding 3 in narrowed form. Findings 6 and 7 are not confirmed by the independent check and are not used.

Table 17: The 11 findings the Teacher raised against its own memory, with the verdict of the independent re-check.

#	Subject	Verdict
1	Blood for Blood note says Hemokinesis does more damage	Confirmed: 18 against 15 in the reference text
2	Two notebook lines overrule the brief's act-1 example	Confirmed: Bludgeon and Hemokinesis on the take-on-sight list
3	"Randomness is free" looser in the card table than in the diagnosis	Partially confirmed: narrow to "without a gate on the randomness itself"
4	A claim conditional on the executor determinizing hidden information	Confirmed: all 50 source summaries record it
5	The simulator's built-in heuristic agrees on Runic Dome and not on the rest	Confirmed
6	Apotheosis advice assumes control the player lacks	Not confirmed by the independent check, not used: the card was played at floor 46 of one run
7	The card table is friendlier to Barricade than the notebook	Not confirmed by the independent check, not used: no notes/table conflict, and one Barricade run is a victory
8	Thresholds are the teacher's, not the data's	Confirmed: none appears in the reference text or in the human input
9	A mean stated as a bound	Confirmed: Strikes and Defends left in the final deck average 3.9 for winners and 6.2 for losers, and 8 of 14 winners hold 4 or fewer
10	Deck-size gates sit against the data	Confirmed: final deck sizes run 33–43 for winners and 18–41 for losers
11	Three mechanical imprecisions	Confirmed, all three