JY-CRPG-BENCH: A LONG-HORIZON BENCHMARK FOR FRONTIER MODELS IN AN OPEN-WORLD RPG

Anonymous authors
Paper under double-blind review

ABSTRACT

A long-horizon task requires a model to perceive its environment, decide for itself what is worth doing, and hold one plan across hours of dependent decisions. We present JY-CRPG-BENCH, which places a frontier model inside Heroes of Jin Yong, a 1996 open-world role-playing game whose ending requires the fourteen books of the Jin Yong novels. The model plays it as a person does, from the screen and the keyboard alone, which requires reading 320×200 isometric pixel art and 16-pixel Traditional Chinese and knowing the novels its world is built from. It has no quest log, marker or score, so the model must keep its own map and goals, and it must win turn-based battles against the game AI. Progress is scored by milestones of the game itself and by completion time. Under a 60-minute budget, in which a human speedrun finishes the game, the best of 12 frontier models reaches 10 of the 11 milestones in the only session that wins a battle, and no session holds a book. The models stall at four points: leaving the first area, finding the entrances on the world map, acting on instructions given in dialogue, and winning the first battle. Two of these act as filters that a session passes early or not at all, and four-hour sessions of 9 models get no further, while experienced human players hold a book within 105 minutes.

1 INTRODUCTION

Frontier models solve competition mathematics and write working compilers (Google DeepMind, 2025; Carlini, 2026), yet the tasks they complete reliably are still measured in hours of human work (Kwa et al., 2025). A long horizon is a property of the task, the number of dependent steps its completion requires (Chen et al., 2026), and several of the benchmarks that measure it are text environments, business simulations (Backlund & Petersson, 2025; Fan et al., 2026) or interactive fiction (Phan et al., 2025). A game adds a requirement they leave out: the model has to recover its own state from raw perception, because nothing in the environment reports it. The game benchmarks that keep perception each relax another part of the problem: BALROG adds a text rendering of the state (Paglieri et al., 2025), ARC-AGI-3 removes language (ARC Prize Foundation, 2026), and MineExplorer hands the model its objectives (Ju et al., 2026). None pairs the screen and keyboard of a human player with a script other than English, as Table 1 shows.

We present JY-CRPG-BENCH, built on Heroes of Jin Yong, a 1996 Chinese role-playing game that we run unmodified under emulation. Its ending requires recovering the fourteen books of the Jin Yong novels from an open world, and Figure 1 shows how it is played. We chose the game because it tests five capabilities of a frontier model at once.

1. The model must read the screen. The frame is 320×200 pixel art in a $45°$ isometric projection, and every objective, prompt and statistic on it is Traditional Chinese in 16-pixel glyphs, a script that multimodal evaluation has largely neglected (Tam et al., 2025). No text is supplied to the model.
2. The model must bring knowledge the game never states. The world is assembled from the fourteen novels, so who a character is, which faction to join and which storyline leads to a book are known to a reader of the novels and written nowhere on screen.
3. The model must set its own goals. The game keeps no quest log, draws no marker and reports no score, so the model maintains its own map and plan and decides for itself what is worth doing.

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.

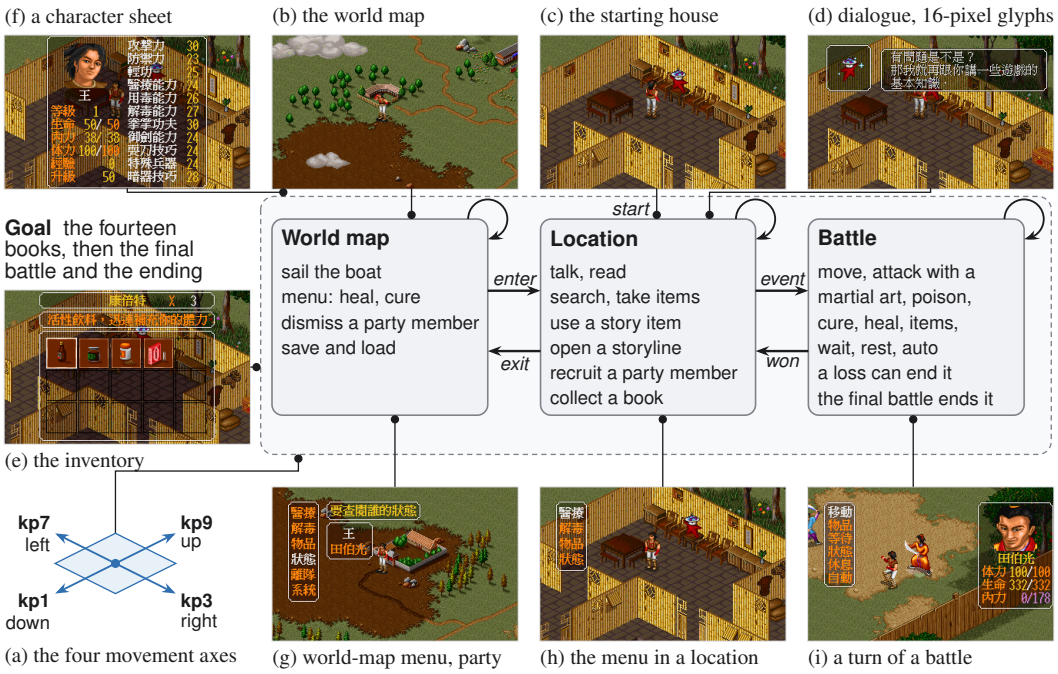


Figure 1: The game as a state chart of three states, each repeating the steps it lists, with the transitions between them. Every frame is one the model receives, upscaled three times without interpolation, and points at the state it shows, or at the dashed outline when it holds in every state.

ARC-AGI-3 places the same goal-setting requirement at the centre of agentic intelligence (ARC Prize Foundation, 2026).

4. The model must act through a keyboard alone, exposed through an API that it calls as a tool. Movement runs along four diagonals, no key moves straight up the screen, and every system of the game is a menu.
5. The model must win battles. Battles are turn-based on a grid against the AI of the game, with position, range and internal energy to manage, and a party at level one can lose them.

Together these form one long-horizon task, a persistent world with sparse and delayed reward over thousands of dependent steps. The benchmark gives the model only what a person playing the game has, the frame on screen and the keys of the keyboard, with no state interface, text rendering, high-level action or memory built for a model, so its difficulty is that of the game as a person plays it. Section 2 reviews related benchmarks, Section 3 describes the benchmark, and Section 4 reports the results.

Table 1: Game benchmarks for language-model agents. Observation is what the model receives before any upscaling, action is what it sends, and the last column is the record that verifies progress.

Benchmark	Environment	Observation	Action	Language	Progress from
JY-CRPG-BENCH	one 1996 CRPG	320×200 px	keys	Trad. Chinese	memory, save, replay
BALROG	6 research games	text, render	text	English	env. reward
TextQuests	interactive fiction	text	text	English	game state
ARC-AGI-3	abstract puzzles	64×64 grid	6 actions, click	none	env. score
VideoGameBench	1990s action games	native px	keys, mouse	English	checkpoint images
GameWorld	34 browser games	browser	mouse, keys, API	English	game state
MineExplorer	Minecraft	3D render	skill calls	English	milestone rules

2 Related Work

2.1 Long-Horizon Evaluation

Vending-Bench established long-term coherence as the property that separates long-horizon tasks from long chains of short ones (Backlund & Petersson, 2025), and E-Commerce Bench extends the setting to a year of deterministic business operation (Fan et al., 2026). Both keep perception out of the loop. TextQuests poses open-ended play over a long horizon in interactive fiction, without vision (Phan et al., 2025), and TextAtari stretches horizons to 10^5 steps over text renderings (Li et al., 2025). ARC-AGI-3 isolates adaptation to novel tasks by excluding language and external knowledge (ARC Prize Foundation, 2026), and AutumnBench probes world-model quality through environment-level queries (Warrier et al., 2025). UltraHorizon makes exploration under hidden rules the task itself (Luo et al., 2025), and Kapoor et al. (2026) observe that benchmark design has favoured tasks that are precisely specified, automatically graded and short, and call for open-world evaluation. Navigation is a documented weakness even without perception: one-shot maze navigation fails by ten cells a side for every model tested (Jiang et al., 2026), and from coordinates alone the failures are attributed to looping (Einarsson, 2025). Long horizons also complicate measurement: the ABC checklist catalogues task-validity and outcome-validity failures across agentic benchmarks (Zhu et al., 2025), and an audit of agent traces finds reward hacking in a majority of them on two suites (Shao et al., 2026).

2.2 Game Environments as Agent Benchmarks

Games have been the standard setting for sequential decision-making since the Arcade Learning Environment (Bellemare et al., 2013), and language and vision-language agents have produced a second generation of game benchmarks. BALROG aggregates six reinforcement-learning environments under one protocol and reports a persistent gap between what a model knows about a game and what it does inside one (Paglieri et al., 2025). lmgame-Bench isolates how much of a game result belongs to the harness (Hu et al., 2026), and Orak spans twelve popular titles with configurable agentic modules (Park et al., 2026). VideoGameBench is the nearest benchmark to ours, running real 1990s games from raw pixels under a no-scaffold rule (Zhang et al., 2025a). GameWorld standardises the action interface across 34 browser games and pairs every task with a state-verifiable metric (Ouyang et al., 2026). MineExplorer evaluates open-world exploration in Minecraft and finds that models handle single-hop tasks but degrade once hidden prerequisites must be coordinated over a longer trajectory (Ju et al., 2026).

The nearest precedents in horizon are the public Pokémon runs. Claude 3.7 Sonnet reached three gym badges from screen pixels with a memory scaffold, where an earlier model had not left the first house (Anthropic, 2025). Gemini 2.5 Pro finished Pokémon Blue with a harness that translated the RAM of the game into text (Comanici et al., 2025). The PokeAgent Challenge turned the setting into a speedrunning track scored by completion time over milestones (Karten et al., 2026a). PokeGym builds a long-horizon benchmark on the open-world game Pokémon Legends: Z-A with no access to game state (Zhang et al., 2026a), and StarBench one on the turn-based Honkai: Star Rail (Zhang et al., 2025b). Crafter established the single-environment, achievement-based design we follow (Hafner, 2022), and NetHack remains the deepest environment in use for learned agents, with runs of tens of thousands of turns (Küttler et al., 2020). Cradle and Voyager drive games through a foundation model with a skill library or a code API (Tan et al., 2024; Wang et al., 2024a).

3 Benchmark Design

3.1 Environment

Heroes of Jin Yong, published by Heluo Studio in 1996, is an open world of two tiers: one world map of 480×480 tiles (scarsty & contributors, 2026) and 84 locations attached to it, such as buildings, sects and caves. Figure 1 draws the game as three states. The player enters a location from the world map and exits back to it, and inside a location a story event or a challenge starts a battle. A battle is turn-based on a grid: each turn a character moves along the four axes and then strikes, uses an item, waits or rests, and skills draw on internal energy. A won battle returns to the location, and a lost one either ends the game or lets the story continue. The game states its goal only through its story: enter

the world of the Jin Yong novels, join its factions, learn its martial arts, and recover the fourteen books that end the game. Appendix G gives the Chinese original of every name from the game.

The game fixes the order of its opening and leaves the campaign open. It starts in a small house with a fenced yard, a searchable chest, one speaking character and one exit tile, and beyond the exit the path south across the world map leads to the hermit. His house, the starting house and five inns are the 7 locations open at the start. Figure 9 in Appendix B shows the route. A cabinet beside the hermit holds the compass, the only source of coordinates the game offers, and his conversation opens all but 4 of the other locations, among them the house of the first party member, who joins after a yes-or-no prompt. The opening ends there without a battle, and the order of play after it is left to the player. The campaign runs through battles, the experience and levels that victories bring, and the chains of items, factions and battles that lead to the books. With all fourteen books and enough reputation, the ending is a final battle (bocelli, 2023).

3.2 Observation and Action Space

The model observes the raw framebuffer at 320×200 and acts by pressing keys. A scored session serves the native frame without upscaling, annotation, set-of-marks overlay or accessibility tree (cf. Xie et al., 2024). The frame combines several documented weaknesses of vision-language models. Low resolution reduces their accuracy (Niu et al., 2025), and images far from the training distribution degrade recognition even of familiar categories (Lin et al., 2026). Text rendered as pixels is understood less well than the same text supplied as tokens (Liu et al., 2026), and precise spatial reading fails once primitives are small or adjacent (Rahmanzadehgervi et al., 2024; Wang et al., 2024b). The action space is the keyboard of 100 keys, and Figure 8 shows the few the game responds to. Movement runs along the four numpad diagonals, with the arrow keys aliased onto the same axes, so a model that presses `up` expecting to walk up the screen moves up and to the right.

The single action presses a key, or a list of up to 100 keys in order, holding each for a requested number of emulated frames. It returns once the screen has settled, including after a change of location, so the frame the model receives is the one to act on. There is no call that advances the game without input. The model can save and load through the menu of the game but cannot restore snapshots of the emulator, and Table 2 lists the calls. Appendix B states the settle rule in frames, and Appendix C gives the reasons for the list action and for the other choices of the protocol.

3.3 Harness and Session Protocol

The benchmark supplies the instructions and the control surface, and the model plays through an agent harness. The hour sessions ran inside pi[1], a minimal coding agent, installed without modification and set to its high thinking level. Its four tools read, write and edit files and run a shell, from which the model drives the control surface with HTTP calls. The model receives one prompt, which names the instructions and sets the rules of the session: it may not read other conversations, look up walkthroughs or code online, or restart, and it reads and writes files only in a folder of its own. The instructions cover the session call, the calls that constitute play, the opening route as far as the compass, and a manual of the controls and systems of the game. They are published as `agents.md` in Chinese, which every model in this paper reads, and in an English translation, and Appendix F gives the prompt and the instructions in both languages side by side. An action returns only its metadata unless the model asks for the frame in the same reply, so the model looks at the screen when it decides to. The benchmark adds no memory scaffold, no map, no pathfinder and no access to the state of the game. A model keeps any memory between actions in files it writes in that folder, and a scored session receives no field that reports its progress until it ends, so a model cannot query its own score. A session that reads the catalogue or the timelines of other sessions does not count. The leaderboard ranks models under one fixed harness, since harness choice changes game results (Hu et al., 2026; Karten et al., 2026b) and on long-horizon tasks the variance a harness induces can exceed the variance between models (Zhang et al., 2026b).

A session is created with a declared model name and a playtime, and its clock starts at the first playable moment. The budget is played time, as for the human references and as in ARC-AGI-3 and TextQuests (ARC Prize Foundation, 2026; Phan et al., 2025). The game advances only on input,

[1] `https://github.com/badlogic/pi-mono`

so, as on a chess clock, a model thinks as long as it likes before each action and the time counts against its budget. Token counts are not comparable across closed multimodal models, whose image encoders are not published. Each session is a separate emulator process with a private copy of the game and its save slots reset to a new game, so no session inherits the progress of another. A session ends when its budget lapses or the game is completed. It then renders its recording to a time-compressed video, computes its metrics, and appends itself to the public catalogue with the video and its keypress timeline. Figure 2 shows the apparatus, and Appendix B lists the emulation parameters.

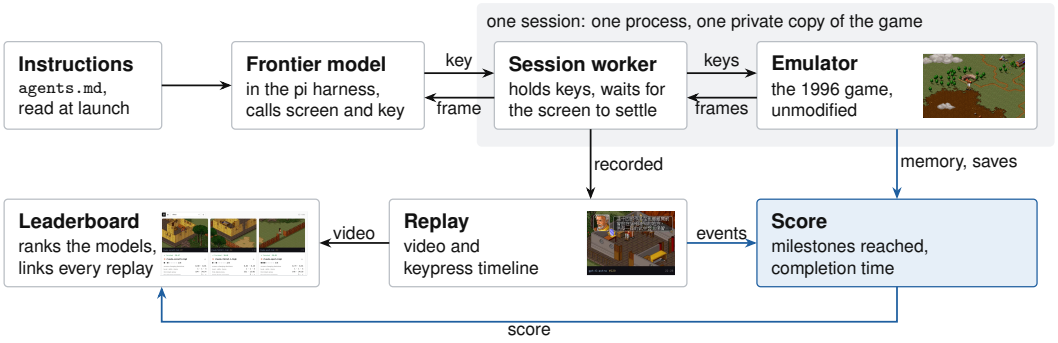


Figure 2: A frontier model in pi plays through the control API of its own session. Blue marks the measurement, which reads the memory of the emulator, the saves of the game and the on-screen events of the replay, and which the model cannot see.

3.4 Progress Measurement

Progress is read from records the model never sees: the state of the game, in the memory of the emulated machine and in the saves the game writes, and the published replay. The state gives the inventory, the compass, the books, the party and the level and experience of the leader. The game keeps seven events only on screen: the model enters a location from the world map, speaks with the hermit, opens the item screen with the compass in it, enters a battle, loses it or wins it, or answers the prompt of a party member who asks to join. Each draws fixed art at a fixed screen position, so these events are read by matching every frame of the replay against that art. A milestone, once observed, stays credited, so an item picked up and later used still counts. Appendix D lists the record behind each milestone and checks the readings against one another.

The milestone scale has 11 steps, following the achievements of Crafter (Hafner, 2022): six in the opening and five in the campaign. The opening milestones are the exit tile, the chest in the first room, a location entered from the world map, the conversation with the hermit, the compass in his cabinet, and the party member who joins through dialogue. The campaign milestones are a battle entered, a battle ended, won or lost, experience gained, the second level, and the first of the fourteen books. The milestones are listed in the order of the opening route, with the optional chest after the exit, and a session is credited with every milestone it reaches, so the count does not depend on the path taken.

The milestones extend to the ending of the game. The fourteen books are required before the final battle, and the saves record every book held, so the count of books continues the scale beyond the last milestone. The ending, read from the replay, gives the completion time that ranks the sessions that finish.

4 Evaluation

We evaluate 12 models from 5 vendors at the 60-minute budget, which is sized to the opening, with 42 sessions on record and 2 to 6 per model. A model is credited with every milestone any of its sessions reached, and a session created under a variant of a model name, a spelling or the thinking level, is listed under the model. The service ends a session after ten minutes without an action, which stopped 11 of the 42 sessions before the hour. The baseline is a random policy that plays the

same budget through the same control surface without reading the screen (Mnih et al., 2015; Hafner, 2022; ARC Prize Foundation, 2026). Its 2 sessions replay one key sequence drawn from seed 1996, and Appendix C gives the distribution of its keys. The human references are read from published videos of the game, 4 speedruns and 3 playthroughs by experienced players, scored by the milestone readings of a replay, which Appendix D describes.

4.1 Main Results

Of the 12 models, all 12 reach the world map, 10 pick up an item, 5 speak with the hermit, and two take the compass, enter a battle and see it end. `claude-opus-5.5` reaches 10 of the 11 milestones in one of its 2 sessions, which recruits a party member and wins a battle whose experience lifts the hero to the second level, and `gpt-6-astra` reaches 8 across its 4 sessions. No session holds a book. Figure 3 shows, for each model, the share of sessions that reached each milestone, with the human references in its two top rows. The speedruns, all of which finish the game in 60 to 126 minutes, cross onto the world map within the first minute in 32 to 107 keypresses and hold the first book between minutes 26 and 72. The playthroughs, one of which reaches the ending, hold it between minutes 58 and 105.

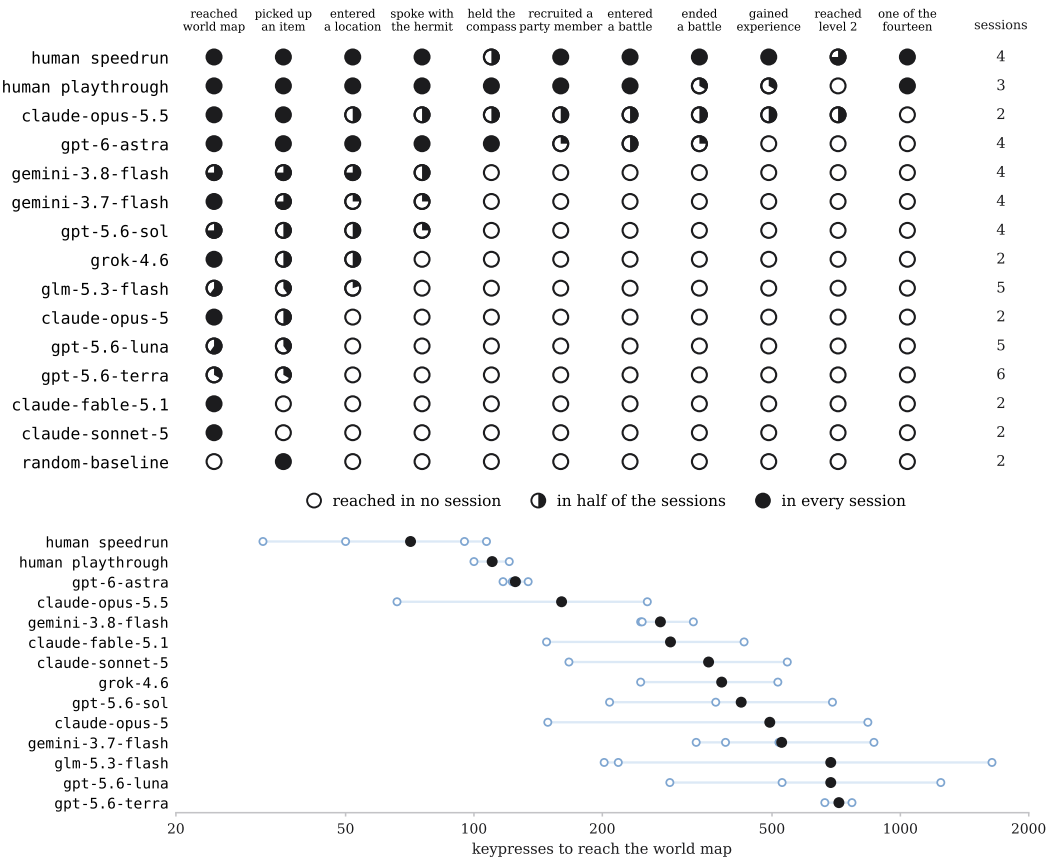


Figure 3: Milestones per model. The rows of the upper panel are ordered by the milestones reached and the rows of the lower panel by the mean keypresses to the world map, with the random baseline last. In the upper panel each marker shows the share of sessions in which the model reached the milestone, with the number of sessions at the right. The lower panel shows the keypresses to the world map across the sessions that crossed, on a log scale, with a light dot per session and a black dot for their mean. The two human rows are read from published videos, and Appendix D gives the record behind each milestone.

4.2 Behavioural Analysis

The random policy reaches 1 of the 11 milestones in 60 minutes, since a random walk in a small room reaches the chest within the hour. The world map is the first milestone a random walk does not reach. The chest therefore separates models from one another but not from the baseline: 10 models searched it and 2 never did. Beyond the chest the models stall at four points, and Figure 10 in Appendix D shows frames from each.

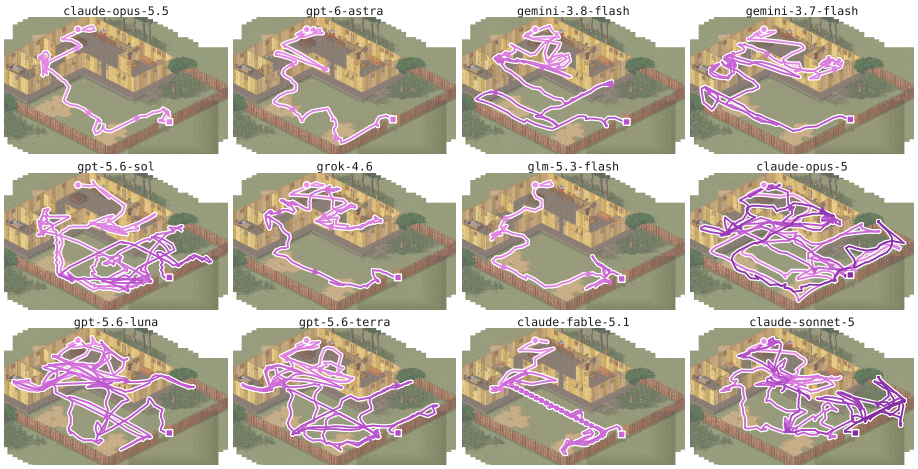


Figure 4: The walk of every model through the starting house in its session with the most milestones, from the spawn tile (circle) to the crossing (square). The colour gives the minute of play on the scale of Figure 5 and the arrowheads the direction. The panels follow the upper panel of Figure 3, and Appendix D describes how a walk is read.

The first is the gate of the starting house, where the model must map its keys onto the isometric view. The yard is closed by a fence with one gate, and Figure 4 draws the walk of every model through it. Every model crosses onto the world map, in 32 of the 42 sessions, and 7 models cross in every session they played, but the gate is found by trial. The median crossing took 308 keypresses, and `glm-5.3-flash` pressed 1640 keys in one session, walking into the fence until minute 37. The count varies more within a model than between models, as the lower panel of Figure 3 shows: the crossings of `glm-5.3-flash` differ by a factor of 8 and the means of the models by a factor of 6.

The second is finding the entrances on the world map, the same problem at a larger scale, since the entrance of a walled house is one tile on one face. After the crossing the rest of the hour is spent on the world map, and Figure 5 draws three walks on it. 7 models entered a location from it, for all but `claude-opus-5.5` and `gpt-6-astra` an inn or the house of the hermit, and the other 5 stayed on the world map or went back into the starting house. `claude-fable-5.1` crossed at minute 13, then spent 1261 actions and 44 minutes on the world map without entering a location, passing the gate of the hermit repeatedly.

The third is the dialogue of the hermit, where the model must turn what it reads into an action. The instructions name the cabinet beside him, and his conversation names it again. 5 models reach the conversation, in 9 sessions. `gemini-3.8-flash` pressed through it for 15 minutes and `gemini-3.7-flash` for 1, and neither searched the cabinet. `gpt-6-astra` reached him in every one of its 4 sessions, between minutes 11 and 19, and the 4 other models that reached him did so in at most half of their sessions. The session of `claude-opus-5.5` in Figure 5 searched the cabinet and used what it found, reading its coordinates from the compass 5 times.

The fourth is the first battle, where the model has to recover from a defeat. All 3 sessions that entered a battle entered it in the house of Yan Ji. `gpt-6-astra` lost it at minute 28 in one session and sent its last key at minute 42 with it still open in another, and `claude-opus-5.5` lost it at minute 17. What separates the one win is what that session did next, as Figure 6 shows: it recruited a party member at minute 22, returned to Yan Ji at minute 26 and won at minute 28, passing through all

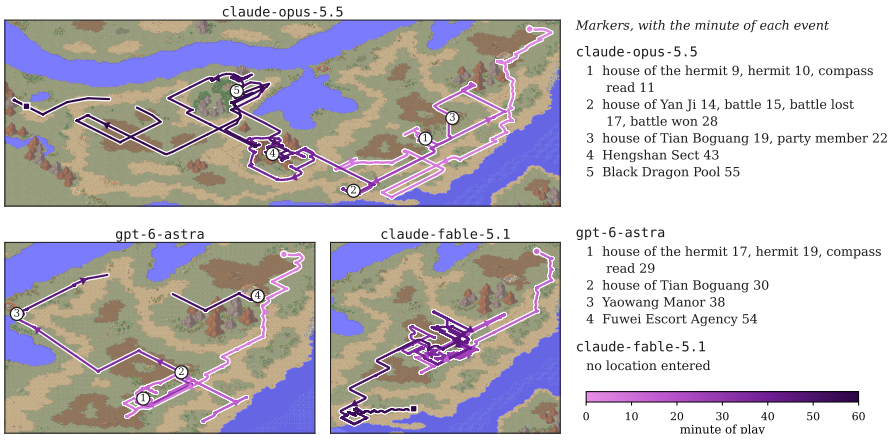


Figure 5: Three walks on the world map in the first hour, coloured by the minute of play: the session of `claude-opus-5.5` with the most milestones, the session of `gpt-6-astra` that entered the most locations, and the session of `claude-fable-5.1` that stayed longest on the world map without entering a location. A numbered marker stands at the last point on the world map before a location entered or an event read from the replay, and the column at the right lists the markers with their minutes. A gap in a path is a visit to a location or a loaded save.

three states of Figure 1. No other session that lost a battle returned to it, and Appendix E compares the battles.



Figure 6: The session of `claude-opus-5.5` that won a battle: the defeat of the hero alone, the party chosen on its return, an item that restores internal energy, and the battle won.

4.3 Four-Hour Sessions and Filters

A four-hour session counts when the service ended it at the budget with its last key in the final 15 minutes. Of the 48 attempts, 10 by 9 models count, and Table 5 lists them. Appendix B gives the reasons the others do not, among them an attempt of `gemini-3.8-flash` that broke the rule of its prompt by reading the timelines of earlier sessions.

Figure 7 follows all 52 model sessions, hour and four-hour, through the steps a playthrough passes in order. Every step thins the population, and two act as filters: sessions pass them within the first 30 minutes or not at all. Of the 18 sessions that entered a location, 10 reached the hermit, all within 18 minutes, and none after 30 minutes, although the others played 8 more hours beyond that point. Of the 4 that entered a battle, one won, within 30 minutes, and none after, although the others played 3 more hours. After their first hour the four-hour sessions pass no filter, and the one among them that lost a battle, of `claude-opus-5.5`, did not return to it in 183 more minutes and 11099 actions.

5 Conclusion

We presented JY-CRPG-BENCH, a long-horizon benchmark in which a frontier model plays a 1996 open-world role-playing game with the screen and keyboard of a human player, scored by events of the game itself. Progress over this horizon is gated by filters. A session reaches the hermit or wins its first battle within its first 30 minutes or never does, and neither more played time nor longer

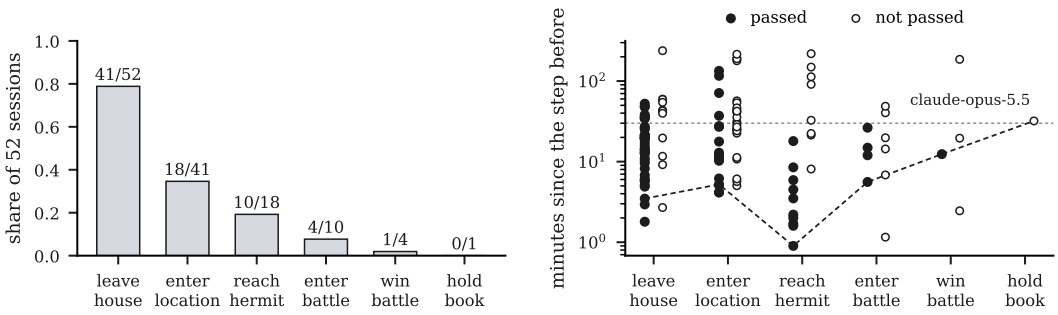


Figure 7: All model sessions through the steps a playthrough passes in order. Left, the sessions that passed each step, of those that passed the step before. Right, for each of those, the minutes from the step before to passing it (filled) or to its last key (open); the dotted line marks 30 minutes, and the dashed line traces the session of `claude-opus-5.5` that won a battle.

thinking changes that: the four-hour sessions pass no filter after their first hour, and the milestones a session reaches barely rise with the time its model thinks before each action. A stuck model spends the extra time repeating what it has tried: it walks into the same fence, circles the same gate or presses through the same dialogue. A long horizon thus measures which filters a model passes, and experienced human players pass all of them and hold a book within 105 minutes. Each filter asks for a capability of any general agent: grounding its actions in an unfamiliar view to find an entrance one tile wide, turning a line it reads into the action it names, and revising its plan from the outcome of its own actions. The one session that won a battle did the last: it lost, recruited a party member and returned to win, although no instruction describes that plan. Since more time at test does not supply these capabilities, they have to come from training across many environments that demand them, including small text in scripts other than Latin. Longer sessions and a first-time human player under the same instructions will extend the scale into the campaign.

AI USE STATEMENT

We used generative AI tools, coding agents built on frontier language models, throughout this work. They implemented the benchmark service, the replay scanner and the scripts that generate every number, figure and table in this paper from the session records. They recovered and reformatted session records, supported the qualitative analysis of the replays, and proposed and refined parts of the measurement design. They searched for and formatted related work, and they drafted and edited the text of this paper under our direction. We have not used generative AI tools to generate synthetic data, to formulate or prove mathematical claims, or to propose hypotheses. The models evaluated in the benchmark are the object of study. We have reviewed all AI-assisted work. Every number in the paper is regenerated from the session records at each build, every typed claim about the records is verified by a consistency script, and the frames and readings behind the qualitative analysis were checked against the replays. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The benchmark runs a commercial 1996 game that is not redistributed: every host supplies its own copy, and the frames in this paper identify the interface under study, as Appendix A explains. The work involves no human participants, and the public catalogue carries the metrics, videos and keypress timelines of model sessions and no personal data.

REPRODUCIBILITY STATEMENT

The benchmark code, the instructions the models receive, the replay scanner and the scripts that generate every number, figure and table in this paper will be released with the paper, and an anonymised copy of the code base is in the supplementary material. Appendix F gives the instructions in full. The release includes the records, videos and keypress timelines of every session in this paper. Ap-

pendix B lists the emulation parameters and the dates of the sessions, Appendix C gives the reasons for the protocol choices, and Appendix D the readings behind every milestone. The game itself is not redistributable, as Appendix A states.

REFERENCES

Anthropic. Claude's extended thinking. Research post, section "Claude plays Pokémon", https://www.anthropic.com/research/visible-extended-thinking, 2025.

ARC Prize Foundation. ARC-AGI-3: A new challenge for frontier agentic intelligence. *arXiv preprint arXiv:2603.24621*, 2026.

Axel Backlund and Lukas Petersson. Vending-Bench: A benchmark for long-term coherence of autonomous agents, 2025. URL https://arxiv.org/abs/2502.15840.

Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The Arcade Learning Environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279, 2013.

bocelli. A complete walkthrough of the original Heroes of Jin Yong (in Chinese). https://forum.gamer.com.tw/C.php?bsn=10859&snA=2778, 2023.

Nicholas Carlini. Building a C compiler with a team of parallel Claudes. https://www.anthropic.com/engineering/building-c-compiler, 2026.

Mingguang Chen, Licheng Wang, and Bo Qu. The horizon gap: Planning, memory, execution, training, and evaluation for long-horizon LLM agents, 2026. URL https://arxiv.org/abs/2608.06663.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261.

Hafsteinn Einarsson. MazeEval: A benchmark for testing sequential decision-making in language models, 2025. URL https://arxiv.org/abs/2507.20395.

Wei Fan, Xinjie Shen, Xudong Guo, Jianhong Tu, Yang Su, Yinger Zhang, Lianghao Deng, Fengyu Wang, Baohua Dong, Yangqiu Song, and Dayiheng Liu. E-Commerce Bench: Evaluating LLM agents on long-horizon autonomous business operation, 2026. URL https://arxiv.org/abs/2608.30730.

Google DeepMind. Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad. https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad, 2025.

Danijar Hafner. Benchmarking the spectrum of agent capabilities. In *International Conference on Learning Representations*, 2022.

Lanxiang Hu, Mingjia Huo, Yuxuan Zhang, Haoyang Yu, Eric P Xing, Ion Stoica, Tajana Rosing, Haojian Jin, and Hao Zhang. lmgame-Bench: How good are LLMs at playing games? In *International Conference on Learning Representations*, 2026.

Yuhan Jiang, Peng Luo, and Liqiu Meng. Lost in Aggregation: A multi-scale diagnostic benchmark for LLM spatial navigation, 2026. URL https://arxiv.org/abs/2606.22219.

Tianjie Ju, Yueqing Sun, Zheng Wu, Wei Zhang, Yaqi Huo, Xi Su, Qi Gu, Xunliang Cai, Gongshen Liu, and Zhuosheng Zhang. MineExplorer: Evaluating open-world exploration of MLLM agents in Minecraft, 2026. URL https://arxiv.org/abs/2605.30931.

Sayash Kapoor, Peter Kirgis, Andrew Schwartz, Stephan Rabanser, J. J. Allaire, Rishi Bommasani, Harry Coppock, Magda Dubois, Gillian K Hadfield, Andrew B. Hall, Sara Hooker, Seth Lazar, Steve Newman, Dimitris Papailiopoulos, Shoshannah Tekofsky, Helen Toner, Cozmin Ududec, and Arvind Narayanan. Open-world evaluations for measuring frontier AI capabilities, 2026. URL https://arxiv.org/abs/2605.20520.

Seth Karten, Jake Grigsby, Tersoo Upaa Jr, Junik Bae, Seonghun Hong, Hyunyoung Jeong, Jaeyoon Jung, Kun Kerdthaisong, Gyungbo Kim, Hyeokgi Kim, Yujin Kim, Eunju Kwon, Dongyu Liu, Patrick Mariglia, Sangyeon Park, Benedikt Schink, Xianwei Shi, Anthony Sistilli, Joseph Twin, Arian Urdu, Matin Urdu, Qiao Wang, Ling Wu, Wenli Zhang, Kunsheng Zhou, Stephanie Milani, Kiran Vodrahalli, Amy Zhang, Fei Fang, Yuke Zhu, and Chi Jin. The PokeAgent Challenge: Competitive and long-context learning at scale, 2026a. URL https://arxiv.org/abs/2603.15563. NeurIPS 2025 Competition Track.

Seth Karten, Joel Zhang, Tersoo Upaa Jr, Ruirong Feng, Wenzhe Li, Chengshuai Shi, Chi Jin, and Kiran Vodrahalli. Continual harness: Online adaptation for self-improving foundation agents, 2026b. URL https://arxiv.org/abs/2605.09998.

Heinrich Küttler, Nantas Nardelli, Alexander Miller, Roberta Raileanu, Marco Selvatici, Edward Grefenstette, and Tim Rocktäschel. The NetHack learning environment. *Advances in Neural Information Processing Systems*, 33:7671–7684, 2020.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Chris Painter, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long software tasks, 2025. URL https://arxiv.org/abs/2503.14499. NeurIPS 2025.

Wenhao Li, Wenwu Li, Chuyun Shen, Junjie Sheng, Zixiao Huang, Di Wu, Yun Hua, Wei Yin, Xiangfeng Wang, Hongyuan Zha, et al. TextAtari: 100k frames game playing with language agents. *arXiv preprint arXiv:2506.04098*, 2025.

Ling Lin, Yang Bai, Heng Su, Congcong Zhu, Yaoxing Wang, Yang Zhou, Huazhu Fu, and Jingrun Chen. OODBench: Out-of-distribution benchmark for large vision-language models, 2026.

Qing'an Liu, Juntong Feng, Yuhao Wang, Xinzhe Han, Yujie Cheng, Yue Zhu, Haiwen Diao, Yunzhi Zhuge, and Huchuan Lu. VISTA-Bench: Do vision-language models really understand visualized text as well as pure text?, 2026.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, Zixuan Hu, Hongze Mi, Yibo Wang, Naiqiang Tan, Hong Chen, Yi R. Fung, Chun Yuan, and Li Shen. UltraHorizon: Benchmarking agent capabilities in ultra long-horizon scenarios, 2025. URL https://arxiv.org/abs/2509.21766.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Junbo Niu, Yuanhong Zheng, Ziyang Miao, Hejun Dong, Chunjiang Ge, Hao Liang, Ma Lu, Bohan Zeng, Qiahao Zheng, Conghui He, and Wentao Zhang. Native visual understanding: Resolving resolution dilemmas in vision-language models, 2025.

Mingyu Ouyang, Siyuan Hu, Kevin Qinghong Lin, Hwee Tou Ng, and Mike Zheng Shou. GameWorld: Towards standardized and verifiable evaluation of multimodal game agents, 2026. URL https://arxiv.org/abs/2604.07429.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wołczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, et al. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In *International Conference on Learning Representations*, 2025.

Dongmin Park, Minkyu Kim, Beongjun Choi, Junhyuck Kim, Keon Lee, Jonghyun Lee, Inkyu Park, Byeong-Uk Lee, Jaeyoung Hwang, Jaewoo Ahn, et al. Orak: A foundational benchmark for training and evaluating LLM agents on diverse video games. In *International Conference on Learning Representations*, 2026.

Long Phan, Mantas Mazeika, Andy Zou, and Dan Hendrycks. TextQuests: How good are LLMs at text-based video games? *arXiv preprint arXiv:2507.23701*, 2025.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2024. URL https://arxiv.org/abs/2407.06581.

scarsty and contributors. kys-cpp: a C++ reimplementation of Heroes of Jin Yong. GitHub repository, https://github.com/scarsty/kys-cpp, 2026.

Jiaqi Shao, Hanck Chen, Wei Zhang, Maxm Pan, and Bing Luo. Do agent benchmarks measure capability? Protocol validity in the age of agentic AI, 2026. URL https://arxiv.org/abs/2607.22368.

Zhi Rui Tam, Ya-Ting Pai, Yen-Wei Lee, and Yun-Nung Chen. VisTW: Benchmarking vision-language models for Traditional Chinese in Taiwan, 2025. URL https://arxiv.org/abs/2503.10427.

Weihao Tan, Wentao Zhang, Xinrun Xu, Haochong Xia, Ziluo Ding, Boyu Li, Bohan Zhou, Junpeng Yue, Jiechuan Jiang, Yewen Li, Ruyi An, Molei Qin, Chuqiao Zong, Longtao Zheng, Yujie Wu, Xiaoqiang Chai, Yifei Bi, Tianbao Xie, Pengjie Gu, Xiyun Li, Ceyao Zhang, Long Tian, Chaojie Wang, Xinrun Wang, Börje F. Karlsson, Bo An, Shuicheng Yan, and Zongqing Lu. Cradle: Empowering foundation agents towards general computer control, 2024. URL https://arxiv.org/abs/2403.03186.

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*, 2024a. URL https://arxiv.org/abs/2305.16291.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Yixuan Li, and Neel Joshi. Is a picture worth a thousand words? Delving into spatial reasoning for vision language models, 2024b. URL https://arxiv.org/abs/2406.14852. NeurIPS 2024.

Archana Warrier, Dat Nguyen, Michelangelo Naim, Moksh Jain, Yichao Liang, Karen Schroeder, Cambridge Yang, Joshua B Tenenbaum, Sebastian Vollmer, Kevin Ellis, et al. Benchmarking world-model learning with environment-level queries. *arXiv preprint arXiv:2510.19788*, 2025.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, et al. OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. *Advances in Neural Information Processing Systems*, 37:52040–52094, 2024.

Alex L Zhang, Thomas L Griffiths, Karthik R Narasimhan, and Ofir Press. VideoGameBench: Can vision-language models complete popular video games? *arXiv preprint arXiv:2505.18134*, 2025a.

Haoran Zhang, Chenhao Zhu, Sicong Guo, Hanzhe Guo, Haiming Li, and Donglin Yu. StarBench: A turn-based RPG benchmark for agentic multimodal decision-making and information seeking, 2025b. URL https://arxiv.org/abs/2510.18483.

Ruizhi Zhang, Ye Huang, Yuangang Pan, Chuanfu Shen, Zhilin Liu, Ting Xie, Haijun Lei, and Lixin Duan. Mastering PokeGym: Graph-guided multimodal evolution at test time, 2026a. URL https://arxiv.org/abs/2604.08340.

Yunbei Zhang, Janet Wang, Yingqiang Ge, Weijie Xu, Jihun Hamm, and Chandan K. Reddy. Stop comparing LLM agents without disclosing the harness, 2026b. URL https://arxiv.org/abs/2605.23950.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, et al. Establishing best practices in building rigorous agentic benchmarks. *Advances in Neural Information Processing Systems*, 38, 2025.

A Release and Licensing

The benchmark service, its public submission system and the leaderboard are released on acceptance of this paper, and the supplementary material holds an anonymised copy of the code base for review. The benchmark code is released under MIT. The emulation core (`dosbox_pure_libretro`) is GPLv2 and is loaded at runtime through the libretro C API. The bundled macOS core is built from `schellingb/dosbox-pure` at `7f6e8fb`, and the hosted service ships a Linux build pinned by image digest. The game is the 1996 commercial release, copyright Soft-World International and Heluo Studio. It is not redistributable or tracked in the repository, and every host supplies its own copy. The copy the paper was measured with is identified by two SHA-256 digests, one of the program file the launcher loads and one of a manifest over the 616 shipped files without the three rewritable save slots, with one line per file, sorted by name, giving digest and name.

```
Z.DAT  0034f836def2b287e62b0902f08cab22b61b397f24d44d6038274f1750a007ce
manifest  29ef75ab32a7905e0b274a9bba9e9ea0ff1b081e74aeacedf86350604aedc558
```

The public catalogue carries only the metrics, videos and replay timelines we measured, and nothing that substitutes for the game data. The frames of the game in this paper identify the interface under study, as prior work showing frames of copyrighted games has done (Paglieri et al., 2025).

B Interface and Session Details

Table 2 lists the control surface, and Figure 8 shows the keyboard the API exposes. Emulation is DOSBox Pure through libretro at a fixed 26800 cycles with no sound card, so the observation is visual only. The published video runs at 8 times speed, or faster for long runs, with a strip naming the model, the action number, the keys held and the clock, beside a timeline file with every keypress and its hold.

Each key is held for a requested number of emulated frames, ten by default, because the game samples its keyboard once per loop. The action returns once the screen has settled: the game has 30 frames to react, and the wait then ends after 9 frames without change, when the picture cycles as in an idle animation, or after 120 frames regardless. A black frame marks a change of location, and the action waits through it and settles once more.

The hour sessions ran from 10 to 25 September 2026 and the four-hour sessions from 16 to 22 September 2026. Table 3 lists the effort of every model in the quantities the service records: sessions, actions, keys per action and the time between actions, and Table 4 gives each milestone in actions and in minutes, the medians over the sessions that reached it. The sessions took between 31 and 1994 actions. Neither the time a model thinks between actions nor how often it looks set the order of the milestones reached: over the 42 model sessions, their rank correlation is 0.10 with the median think time and -0.15 with the keys per action, and over the 26 sessions whose record counts the screen reads, where the median read the screen 1.01 times per action, it is -0.18 with the reads per action. `gpt-5.6-terra` sent lists of up to 100 keys in one action, 22 on average, so it looked at the screen once for every 22 keys it pressed.

The catalogue also holds sessions the paper leaves out: 2 of `gemini-3.6-flash`, a model with two later versions in this paper, and the 38 four-hour attempts that do not count, of which 5 sent no action, 3 sent one or two, 27 sent their last key before minute 225, 2 were declared under names that identify no model of the paper, and one read the timelines of earlier sessions. That attempt, while stuck in the opening, fetched the keypress timelines of earlier sessions with its shell, read the key sequences of the fastest, and went on to hold the compass and recruit a party member. Reward hacking of this kind is common in agent traces (Shao et al., 2026), and a long horizon gives a stuck model hours to look for it. The release lists each attempt with its reason.

Table 2: The control surface. A check marks a call the model can use in a scored session; the snapshot calls serve development.

Call	Meaning	Model can use
`GET /api/screen`	the current frame	✓
`GET /api/help`	the instructions	✓
`GET /api/keys`	every accepted key name	✓
`GET /api/slots`	emulator snapshots on disk	×
`POST /api/key {"key":"kp3","hold":10}`	press one key, or several in order	✓
`POST /api/save {"name":"..."}`	take a snapshot	×
`POST /api/load {"name":"..."}`	restore a snapshot	×

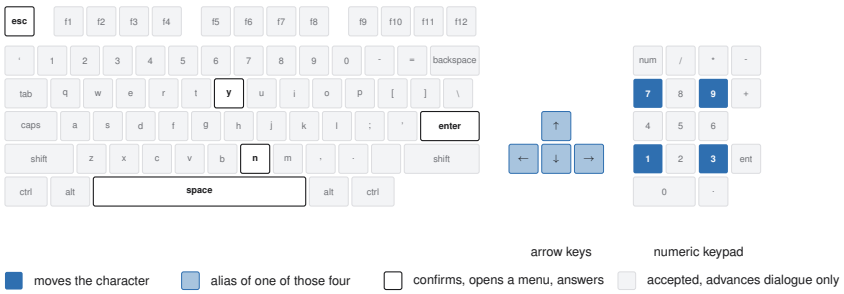


Figure 8: The keyboard the API exposes, with the nine keys the game responds to outside dialogue highlighted; any key advances a dialogue.

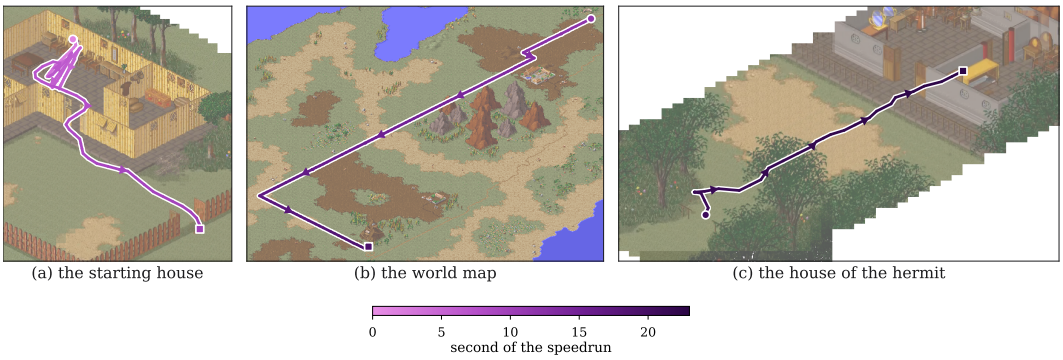


Figure 9: The opening route as far as the hermit, from a human speedrun: (a) the starting house from the spawn tile to the gate, (b) the world map from the gate to the house of the hermit, and (c) that house from its door to the hermit. The walks are read and drawn as in Figures 4 and 5, coloured by the second of the speedrun; the house of the hermit is stitched from the frames of the speedrun.

C Design Decisions

An action accepts one key or a list of up to 100, because a menu path or a repeated step is one decision. A list costs one round trip and one settle wait for the whole list. The coupling between perception and action is set by how often a model looks, since an action returns no picture unless asked, and a model that sends single keys without asking for the screen is as open-loop as one that sends twenty. The cap bounds the longest list, and every list is recorded in the public timeline.

The observation is the frame the game draws, at its own resolution, since reading sixteen-pixel glyphs and an isometric map is part of the task. The instructions list the on-screen terms in the Traditional Chinese the game draws.

The baseline is a seeded random policy because it measures what the milestone structure yields to chance: the small first room puts the chest within reach of a random walk, and no later milestone is.

Table 3: Effort per model in the order of the upper panel of Figure 3; actions and seconds are medians over sessions.

Model	Sessions	Actions per session	Keys per action	Seconds between actions
claude-opus-5.5	2	1116	2.4	2
gpt-6-astra	4	254	3.7	11
gemini-3.8-flash	4	174	3.1	18
gemini-3.7-flash	4	170	4.9	16
gpt-5.6-sol	4	332	3.1	8
grok-4.6	2	472	1.4	5
glm-5.3-flash	5	215	1.8	14
claude-opus-5	2	303	6.5	12
gpt-5.6-luna	5	155	6.9	11
gpt-5.6-terra	6	126	11.8	12
claude-fable-5.1	2	874	1.4	10
claude-sonnet-5	2	369	2.8	9
random-baseline	2	2118	1.0	1

Table 4: Each milestone in actions and minutes: the sessions that reached it and the median action count and minute at which they did.

Milestone	Sessions	Median actions	Median minute
reached the world map	32	86	16
picked up an item	22	20	3
entered a location	13	145	27
spoke with the hermit	9	101	19
held the compass	5	129	21
recruited a party member	2	270	37
entered a battle	3	158	27
ended a battle	2	172	22

Table 5: The four-hour sessions that count, one row per session: the minute of the last key, the actions sent, the minute at which each milestone was first reached, and the milestones credited. The crossing is timed by the service and the other milestones by the replay; a dash marks a milestone not reached.

Model	Last key	Actions	Map	Item	Location	Hermit	Compass	Party	Battle	Ended	Milestones
claude-opus-5.5	229.3	11850	2.9	1.2	13.2	31.2	33.6	–	43.2	46.7	7
	234.2	5310	19.1	5.2	–	–	–	–	–	–	2
gpt-5.6-sol	239.5	782	19.2	210.9	90.3	–	–	–	–	–	3
deepseek-v4.1-flash	239.8	6026	10.3	–	21.2	–	–	–	–	–	2
gemini-3.7-flash	231.1	584	52.4	3.6	–	–	–	–	–	–	2
glm-5.3-flash	238.1	372	48.6	3.2	–	–	–	–	–	–	2
gpt-5.6-luna	238.9	530	9.6	–	125.8	–	–	–	–	–	2
grok-4.6	239.4	2857	14.5	–	148.4	–	–	–	–	–	2
qwen3.8-27b	239.0	203	47.6	1.6	–	–	–	–	–	–	2
glm-5.3	238.4	378	–	45.1	–	–	–	–	–	–	1

Each action draws one of the 13 keys the instructions teach, the four diagonal movement keys with 67 percent of the draws and the confirm key with 12 percent, 1.1 seconds apart on average, until the budget lapses.

D Measurement Details

The world map is read from its frame; the item from the item message; the location from the name banner drawn on entry, the starting house excepted; the hermit from his portrait; the compass from the inventory or the item screen; the party member from the party size or a yes to the recruitment prompt; the battle from the card of the acting character; an ended battle from the defeat or battle-won message; experience and level from the character sheet in the save or the messages a won battle draws; and a book from the inventory or the items the hero carries.



(a) `gpt-6-astra` (b) `gpt-6-astra` (c) `claude-opus-5.5`

(d) `claude-opus-5.5` (e) `gemini-3.8-flash` (f) `glm-5.3-flash`

(g) `claude-fable-5.1` (h) `grok-4.6` (i) `gpt-6-astra`

Figure 10: Frames from the replays behind the readings and the walks of Section 4, each with the strip naming the model, the action count and the clock: (a) the compass coordinates on the item screen; (b) the conversation with the hermit; (c) the recruitment prompt of Tian Boguang; (d) the message of a battle won; (e) the hermit naming the cabinet; (f) the yard behind the fence; (g) passing the gate of the hermit; (h) talking to an innkeeper; (i) the defeat message.

Every replay frame is matched by normalised cross-correlation against nine templates, the fixed art of the on-screen events, and Figure 10 shows frames with them. A template counts when it scores above 0.9 for 0.8 seconds of play: two consecutive frames of a replay at 8 times speed and one frame of a replay at 24 times. The game shows such art for a second or more, so a shorter match is a screen transition. Four messages are dismissed at the next key, so a single frame counts: the item message and the three that close a won battle, which announce the win, the experience gained and a new level. A recruitment prompt counts as a recruitment when the first key pressed after it appears is a yes. The walks of Figures 4, 5 and 9 place every frame of a replay or video at the offset of highest normalised cross-correlation on a panorama of the starting house, stitched from the replays, or on the world map, rendered from the data files of the game. The world map scrolls to keep the hero at one screen position, which a compass reading of a replay fixes, so the offset of a frame gives his tile.

The human rows of Figure 3 are read from 7 published videos, 4 speedruns and 3 playthroughs, listed in the release with the frame behind every reading. A capture is admitted when the first room of the starting house in a benchmark replay matches it at 0.90 or above, which locates the game frame and excludes the remakes of the game. The capture is scaled to the native frame and matched against the templates as a replay is, over a window of two pixels and at a threshold set per video in the gap of its score distribution. The clock starts where a benchmark session starts, on the first frame in which the hero stands on his spawn tile, and the crossing is the first black frame after it. An action is a tile step, read from the shift of the background on the lattice of the isometric tile, or a screen change without a step, such as a line of dialogue. A won battle and a level are read from the messages that announce experience gained and a new level, a book from the item message that

names it, and the ending from the closing conversation with the hermit. A recruitment is read from the party list or from a battle in which a party member acts.

The readings agree wherever two records exist. Each of the 21 sessions whose inventory shows a new item also shows the item message. The crossing count read from the replay agrees with the count the service recorded on all 20 sessions that carry both. Of the 23 sessions with both an inventory reading and a replay, every one of the 4 whose inventory holds the compass shows the coordinate line on the replay, and no other session does. Of the 32 sessions with a preserved save and a replay, every one of the 4 whose save shows the locations the hermit opens also shows his portrait on the replay. Where an event occurs, the five held templates score at least 0.93 and the four single-frame messages at least 0.90; elsewhere they score at most 0.71 and 0.52. The first reading of every template in the hour sessions, 24 in all, was also checked by eye against its frame, and each shows the event.

E Battle Analysis

A battle is turn-based tactics on the grid of the location. The player chooses which party members fight, and on its turn a character moves along the four axes and then attacks with a martial art, uses poison, a cure, medicine or an item, waits, rests or hands the battle to the game. An attack draws on internal energy, so a character without it has no attack in its menu. Table 6 lists every battle on record, and Figure 11 sets the two battles of one session side by side.

All 5 battles are fought against Yan Ji in the yard of his house. Of the 4 fought by the hero alone, 3 ended in defeat and one stayed open when the session stopped sending keys. The battle itself does not separate the win from the defeats: in the won battle 83 percent of the keys confirmed the highlighted command, and the one step on the replay beyond confirming is an item that restores internal energy. What separates it is the plan around the battle. The session of `claude-opus-5.5` is the only one that returned to a battle it had lost, and it returned with a party member recruited in between. The 2 other sessions that lost a battle never fought again, one of them with 193 minutes of its budget left. The sessions that recruited Duan Yu fought no battle afterwards.

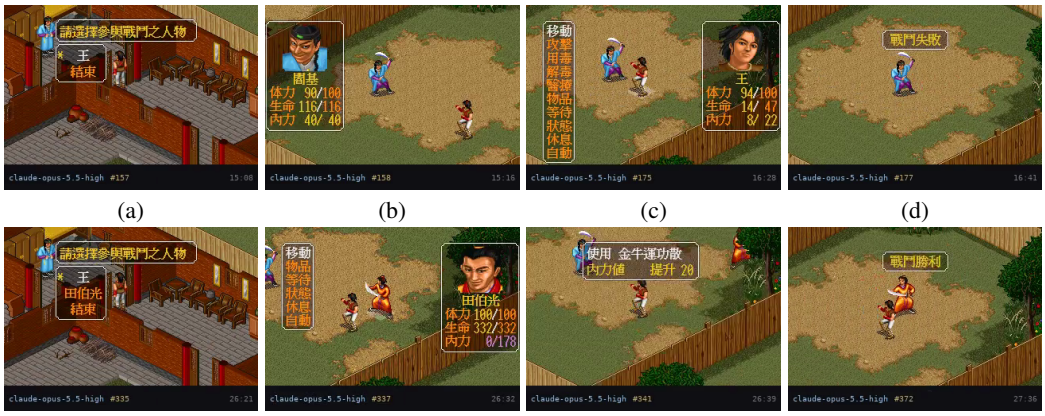
Table 6: Every battle on record, with the minutes it began and ended, and the actions and keys the model sent while it lasted, with the share of keys that confirm.

Model	Budget	Party	Began	Ended	Outcome	Actions	Keys	Confirm
`claude-opus-5.5`	60 min	hero	15.2	16.7	lost	25	45	49 %
`claude-opus-5.5`	60 min	hero, Tian Boguang	26.4	27.6	won	44	52	83 %
`gpt-6-astra`	60 min	hero	39.5	42.0	open	6	20	15 %
`gpt-6-astra`	60 min	hero	26.9	27.9	lost	6	11	64 %
`claude-opus-5.5`	240 min	hero	43.2	46.7	lost	67	86	52 %

F Prompt and Instructions

The first row is the prompt each model is sent to start a session, and the rows after it are the instructions the prompt names. Every model in this paper reads the Chinese in the right column; the left column is its English translation, in the names of Appendix G. In the prompt, `{PLAYTIME}` names the budget and `{MODEL_NAME}` the model. The instructions are those of the 60-minute budget. Those of the four-hour budget differ only in the minutes value of the session call. The addresses of the site and of the hosted service are withheld as `SITE_URL` and `SERVICE_URL`.

English translation	**Chinese, as sent**
Prompt	提示词
Read `SITE_URL/{PLAYTIME}/agents.md` and play as it says. Do not read the contents of other conversations, do not go online for walkthroughs or code, and do not restart or create other tasks. You may create only one new folder, under `/tmp/`, named after the	读 `SITE_URL/{PLAYTIME}/agents.md`，照着玩。注意不要读取其他对话的内容，不要联网查攻略和代码，不要重启或者新建其他任务。你只能在 `/tmp/` 目录下创建一个新的文件夹，session_id 文件夹，session_id 在游戏创建后



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 11: The two battles of the session of `claude-opus-5.5` against Yan Ji. Lost, top: (a) the hero alone chosen to fight; (b) Yan Ji at full health; (c) the hero at low health; (d) the defeat message. Won, bottom: (e) the hero and Tian Boguang chosen; (f) Tian Boguang without internal energy and without an attack in his menu; (g) an item that raises internal energy; (h) the battle-won message.

session_id, which the server tells you once the game is created. Then read, write or execute files only inside that folder. Model name `{MODEL_NAME}`/. Whatever problem you meet, keep playing without stopping, and advance the game as far as you can.

服务器会告诉你。然后仅能在这个文件夹内读写或执行文件。模型名 `{MODEL_NAME}`/。无论遇到什么问题，请一直玩下去不要停下来，尽最大可能推进这个游戏。

---

agents.md

You are about to play 金庸群俠傳 (Heroes of Jin Yong), the original 1996 DOS game by 河洛工作室, running unmodified under emulation. You send keys and fetch the screen. The game is in Traditional Chinese.

Start

```
curl -s -X POST SERVICE_URL/session \
     -H 'content-type: application/json' \
     -d '{"agent":"YOUR-MODEL-NAME","minutes":
    60}'
```

`agent` names the model and its thinking level. `minutes` is the playtime of the run; this brief is the 60 minute one. The reply carries `base_url`, called `$BASE` below. Every call goes there.

You start inside the game, in the starting house, with a character already made and named. Do not change the name or touch the 注音 input method.

Skill: play 金庸群俠傳 (Heroes of Jin Yong)

The original 1996 DOS game by 河洛工作室, running under emulation at $BASE. You send keys and fetch the screen.

The loop

Acting and looking are separate calls. A key press waits for the screen to settle and returns metadata; `GET /api/screen` returns the picture. Add `?image=1` to an action to get the picture in the same reply.

The game is entirely in Traditional Chinese. Objectives, choices and prompts that expect a specific key are all in its text.

API

```
GET  $BASE/api/screen
    look, pressing nothing
```

agents.md

你即将游玩《金庸群俠傳》，1996 年河洛工作室的原版 DOS 游戏，未经修改，跑在模拟器上。你送出按键，并取得画面。游戏是繁体中文。

开局

```
curl -s -X POST SERVICE_URL/session \
     -H 'content-type: application/json' \
     -d '{"agent":"YOUR-MODEL-NAME","minutes":
    60}'
```

`agent` 填模型名和思考档位。`minutes` 是这一局的游玩时长，这份说明对应 60 分钟。回应里的 `base_url` 以下称 `$BASE`，所有呼叫都送到那里。

开局时你已经在游戏里，站在开场房间，角色已经建好并有名字。不要改名，也不要碰注音输入法。

技能：游玩《金庸群俠傳》

1996 年河洛工作室的原版 DOS 游戏，以模拟器执行于 $BASE。你送出按键，并取得画面。

运作方式

动作和看画面是两次呼叫。送按键会等画面稳定后回传状態；`GET /api/screen` 回传画面。在动作 URL 加 `?image=1`，可在同一回应取得动作后的画面。

游戏全是繁体中文。目标、选项和等待特定按键的提问都在文字里。

API

```
GET  $BASE/api/screen
     只看画面，不按任何键
```

```
POST $BASE/api/key   {"key":"kp3"}
    press one key; +"hold" frames
POST $BASE/api/key   {"key":["kp9","enter"]}
    press several, in order
GET  $BASE/api/help
    this skill
```

`/api/screen` returns JSON with `image`, a base64 PNG data URI; `?format=png` returns the raw bytes. `/api/key` is the only action: `key` is one key name or a list pressed in order, so a repeat is a list of the same key and a menu path is a list. It returns once the screen has settled, a change of location included, with `ok`, `action` and `frame`, the number of the picture that followed. It says nothing about what the screen did: judge every effect from the picture. There is no wait call: the game moves only on a key, and an action waits for the result.

Each call reads only the fields shown. Any other field is refused with a 400 naming it.

```
curl -s -X POST $BASE/api/key -H
    'content-type: application/json' \
     -d '{"key":"enter"}'
```

Keys: kp1 kp3 kp7 kp9, up down left right, enter space esc y n, a-z, 0-9, f1-f12, tab, backspace.

Movement

The world is isometric: the four movement axes are diagonals on screen. The numpad names match the visible direction and are identical to the arrows:

```
kp7  ↖ up-left      kp9  ↗ up-right
    (kp7 == left, kp9 == up)
kp1  ↙ down-left    kp3  ↘ down-right
    (kp1 == down, kp3 == right)
```

Use `kp7/kp9/kp1/kp3`. On a clear path, alternating two directions moves horizontally or vertically across the screen:

```
screen-right : kp3, kp9, kp3, kp9, ...
    screen-left : kp7, kp1, ...
screen-down  : kp3, kp1, kp3, kp1, ...
    screen-up   : kp7, kp9, ...
```

`hold` is the number of frames the key stays down, not tiles travelled; it does not follow paths, turn or avoid obstacles. Use short taps where the route is unclear and longer holds on a confirmed clear stretch. A `hold` below 5 is refused: the game reads the keyboard once per loop, and a press released within one loop is lost. The default is 10.

Interacting

- enter and space confirm, advance dialogue and investigate. For a person or container, stand on an adjacent tile facing the target, then press enter or space. Story events triggered by stepping on a tile are separate.
- Any key advances ordinary dialogue. Answer choices and （Y／N） prompts with y and n.
- esc opens the menu. In a building: 醫療 / 解毒 / 物品 / 狀態. On the world map also 離隊 (dismiss a party member) and 系統 (save, load, quit). The game saves only from the world map.

The world

You play 小蝦米, who wakes inside the world of Jin Yong's novels. The way home is to find the fourteen books scattered across the land. Characters from the novels can be recruited and their martial arts learned. Battles are turn-based between teams, in an order set by 輕功. A fallen character, a lost battle and the end of

```
POST $BASE/api/key   {"key":"kp3"}
    按一个键；可加 "hold"（帧数）
POST $BASE/api/key   {"key":["kp9","enter"]}
    依序按多个键
GET  $BASE/api/help
    本技能说明
```

`/api/screen` 回传 JSON，`image` 是 base64 的 PNG data URI；加 `?format=png` 直接取得 PNG 位元组。`/api/key` 是唯一的动作：`key` 是一个键名或依序按下的键名列表，重复按键就是同一个键的列表，选单路径也是一个列表。它会等画面稳定（包括场景切换）才回传 `ok`、`action` 与 `frame`（接下来那张画面的编号），不说明画面发生了什么：一切效果从画面判断。没有等待呼叫：游戏只在按键时前进，动作本身已等到结果。

每个呼叫只读取上面列出的栏位，其他栏位会以 400 指明并拒绝。

```
curl -s -X POST $BASE/api/key -H
    'content-type: application/json' \
     -d '{"key":"enter"}'
```

按键：kp1 kp3 kp7 kp9、up down left right、enter space esc y n、a-z、0-9、 f1-f12、tab、backspace。

移動

世界是等角视角，四个移動轴在画面上都是斜的。九宫数字键的名称对应画面上的方向，与方向键等效：

```
kp7  ↖ 左上      kp9  ↗ 右上          (kp7 ==
    left, kp9 == up)
kp1  ↙ 左下      kp3  ↘ 右下          (kp1 ==
    down, kp3 == right)
```

请用 `kp7/kp9/kp1/kp3`。路径畅通时，交替按两个方向可沿画面的水平或垂直方向前进：

```
画面向右：kp3, kp9, kp3, kp9, ...     画面向
左：kp7,
    kp1, ...
画面向下：kp3, kp1, kp3, kp1, ...     画面向
上：kp7,
    kp9, ...
```

`hold` 是按键保持的帧数，不是移動格数；它不会沿路、转弯或避障。路线不明时短按，已确认畅通的路段再长按。`hold` 小于 5 会被拒绝：游戏每个主回圈只读一次键盘，按下与放开落在同一次回圈内的按键等于没发生。预设为 10。

互动

- enter 与 space 确定、推进对话、调查。与人物或容器互动时，站在相邻格并朝向目标，再按 enter 或 space。踩到特定格子触发的剧情是另一种机制。
- 一般对话可用任意键推进；选项与“（Y／N）”用 y 与 n 作答。
- esc 开启选单。建筑内：醫療／解毒／物品／狀態。大地图上另有離隊與系統（存档、读档、离开）。游戏只在大地图上存档。

这个世界

你是小蝦米，醒来后身处金庸小说的世界。回家的办法是找齐散落各地的十四本小说。小说中的人物可以入队，武功可以习得。战斗是团队回合制，行动顺序依輕功。角色倒下、战斗失败与游戏结束是三件不同的事；败北后能否继续取决于该场遭遇。

the game are different events; whether play continues after a defeat depends on the encounter.

Field manual: controls and game knowledge

First: get the compass

Many locations open only after the opening encounter at 南賢居, the house of the hermit.

1. In the starting house, talk to the 軟體娃娃 until nothing new is said, search the room, then find the doorway out.
2. On the world map, follow the small path south to 南賢居, near `[388,325]`. Talk to 南賢, the hermit, then investigate the cabinet beside him to get the 羅盤 (compass).
3. Highlight the compass in `esc` → 物品 to read the coordinates of the party and the boat. Use them to check your position whenever the route is unclear.

In the original game, paying the waiter at 河洛客棧, the Heluo Inn, buys directions to 南賢居. They are not required to enter.

Controls and menus

- Move with `kp1 kp3 kp7 kp9` or the arrows; they are the same four axes. Holding a key keeps sending the same direction without following paths, turning or avoiding obstacles.
- For a person or container, stand adjacent, face the target, and press space or enter. Stepping on a tile can trigger a separate story event. In a battle, choose commands and targets through its menu.
- `esc` opens the menu anywhere. Arrows move the highlight, space or enter confirms, `esc` backs out.
- `y` and `n` answer （Y／N）. Any key advances ordinary dialogue.

The world-map menu holds 醫療 heal, 解毒 cure poison, 物品 items, 狀態 status, 離隊 dismiss a party member and 系統 system. Inside a location the first four are available; saving, loading and dismissal need the world map. Recruitment happens through dialogue.

- 醫療: choose a healer and a patient. Healing needs at least 50 體力 and enough medical ability for the injury.
- 解毒: choose a curer and a patient; the effect depends on the curing ability and the poison.
- 物品: a story item is used on a person or object while standing adjacent and facing it; medicines, equipment and manuals choose their user in the item menu. Five kinds: story items; pills that restore or raise attributes; hidden weapons, usable only in battle; weapons and armour, equippable by some characters; manuals, which a party member studies to gain attributes or a skill.
- 狀態: health, internal energy, stamina, experience and the battle attributes, with a second page of portrait, equipment and skills. A character learns at most ten martial arts, each to level ten, and studies one manual at a time.
- 系統: three save slots, load, and quit. Save often.

Battles

Battles are turn-based. Order follows 輕功; waiting moves the current character later in the round.

实战手册：操作与游戏知识

最优先：拿到羅盤

许多地点要在南賢居的开局事件完成后才会开放。

1. 开场小屋里先跟軟體娃娃对话到没有新内容，搜完屋内，再找到门口离开。
2. 进入大地图后，沿小路往南前往南賢居，参考座标约 `[388,325]`。先跟南賢对话，再调查他身旁的柜子，取得羅盤。
3. 在 `esc` → 物品 中选中羅盤，可读到人物和船的座标。路线不明时用它核对位置。

原版中，付钱给河洛客棧的小二可以买到前往南賢居的指路。不指路也能进入。

操作与选单

- 移動用 `kp1 kp3 kp7 kp9` 或方向键，两者是同样的四个轴。长按持续输入同一方向，不会沿路、转弯或避障。
- 与人物或容器互动：站在相邻格朝向目标，按 space 或 enter。踩到特定格子触发的剧情是另一种事件。战斗中经由选单选择指令和目标。
- `esc` 在任何地方开启选单。方向键移動反白，space 或 enter 确定，`esc` 返回上一层。
- `y` 与 `n` 回答“（Y／N）”。任意键推进一般对话。

大地图选单有醫療、解毒、物品、狀態、離隊、系統；小场景只有前四项。存档、读档和離隊要在大地图上；入队经由对话。

- 醫療：选医者，再选被医者。医者體力需达 50，且醫療能力要够得上伤势。
- 解毒：选解毒者，再选中毒者；效果取决于解毒能力和中毒程度。
- 物品：剧情物品要站在相邻格朝向目标使用；药物、装备和秘笈在物品选单中选择使用者。五类：剧情物品；回复或提升属性的药丸；只在战斗中可用的暗器；部分人物可装备的武器防具；队员修练后增加属性或学会武功的秘笈。
- 狀態：生命、内力、體力、經驗和战斗属性；第二页是大头照、装备和已学武功。每人最多学十种武功，每种最高十重，同时只修练一本秘笈。
- 系統：三个存档位、读档、离开。常存档。

战斗

Stamina, internal energy, ability and remaining movement decide which commands are available; read the menu.

- 移動 move: choose a position in range. The commands that remain afterwards change.
- 攻擊 attack: choose a martial art, then its target, direction or area.
- 用毒 poison, 解毒 cure, 醫療 heal: need the ability, the resource and a suitable target.
- 物品 items: choose the use and the target, then check the quantity.
- 等待 wait: delay this character's action, which is not the same as ending it.
- 狀態 status: attributes, equipment and martial arts.
- 休息 rest: end the action and recover some stamina.
- 自動 auto: the game controls the whole party's battle actions.

A fallen character is not necessarily dead, and a lost battle is not always the end of the game. Some defeats end the run; read the story and the result.

Attributes

Visible: health, internal energy, stamina, experience, attack, defence, 輕功, healing, poison, curing and the weapon skills. Base attack, defence and 輕功 cap at 100; equipment adds on top. Some skills and items need minimum values.

Hidden:

- 體質 sets the health gained per level.
- 資質 sets how fast skills are learned. A few skills are reserved for characters with poor 資質.
- 道德 moves with your actions and is read from the mirror in 南賢居 with space. Too low and upright characters refuse to join; some paths need a specific range.
- 名望 changes with story events and battles and affects later events.

Coordinates from community guides

Reference coordinates from guides to the original game. An entrance and the tile outside it may differ by one; confirm arrival with the compass and the visible entrance.

Place	Coordinates
主角居 the starting house	(357,235)
河洛客棧	(359,229)
南賢居 the house of the hermit	(388,325)
天寧寺	(330,237)
鐵掌山	(302,343)
衡山派	(355,376)
五毒教	(247,424)
崑崙仙境	(22,440)
無量山洞	(168,426)
閻基居	(396,374)
北丑居	(51,109)

战斗是回合制。顺序依輕功；等待会把当前角色的行动延后。體力、内力、能力和剩余步数决定可用的指令，以选单为准。

- 移動：在范围内选位置。移動后可用的指令会改变。
- 攻擊：先选武功，再选目标、方向或范围。
- 用毒、解毒、醫療：需要相应能力、资源和合适的目标。
- 物品：选用法和目标，之后核对数量。
- 等待：延后此角色的行动，不等于结束行动。
- 狀態：属性、装备和武功。
- 休息：结束行动并恢复部分體力。
- 自動：由游戏控制全队的战斗行动。

角色倒下不一定是死亡，战斗失败也不一定是游戏结束。有些败北会结束这一局；看剧情和战斗结果。

属性

可见：生命、内力、體力、經驗、攻擊、防禦、輕功、醫療、用毒、解毒和各项兵器技巧。攻擊、防禦和輕功的基础值上限 100，装备另计。部分武功和物品有最低要求。

隐藏：

- 體質决定升級时增加的生命。
- 資質决定学武功的快慢。少数武功留给資質差的人。
- 道德随行为变动，对著南賢居的镜子按 space 可以查看。太低则正派人物拒绝入队；部分路线需要特定区间。
- 名望随剧情和战斗变动，影响后续事件。

社群攻略的座标

原版攻略的参考座标。入口与门外的格子可能差一格；以羅盤读数和看得见的入口确认到达。

地点	座标
主角居	(357,235)
河洛客棧	(359,229)
南賢居（羅盤）	(388,325)
天寧寺	(330,237)
鐵掌山	(302,343)
衡山派	(355,376)
五毒教	(247,424)
崑崙仙境	(22,440)
無量山洞	(168,426)
閻基居	(396,374)
北丑居	(51,109)

G GLOSSARY

Table 7 gives the Chinese original of every name from the game that the paper uses, and of the fourteen books, with the English titles of the novels they are named after.

Table 7: Names from the game as the paper writes them and as the game writes them.

In the paper	In the game
Heroes of Jin Yong	金庸群俠傳
Heluo Studio	河洛工作室
Soft-World International	智冠科技
the starting house	王居, 主角居
the hermit	南賢
the house of the hermit	南賢居
inn	客棧
Heluo Inn	河洛客棧
Tian Boguang, his house	田伯光, 田伯光居
Duan Yu	段譽
Yan Ji, his house	閻基, 閻基居
Yaowang Manor	藥王莊
Fuwei Escort Agency	福威鏢局
Hengshan Sect	衡山派
Black Dragon Pool	黑龍潭
compass	羅盤
internal energy	內力
experience gained	獲得經驗點數
a new level	升級了
item message	得到
defeat message	戰鬥失敗
battle-won message	戰鬥勝利

Book	In the game
The Young Flying Fox	飛狐外傳
Fox Volant of the Snowy Mountain	雪山飛狐
A Deadly Secret	連城訣
Demi-Gods and Semi-Devils	天龍八部
The Legend of the Condor Heroes	射鵰英雄傳
White Horse Neighs in the Western Wind	白馬嘯西風
The Deer and the Cauldron	鹿鼎記
The Smiling, Proud Wanderer	笑傲江湖
The Book and the Sword	書劍恩仇錄
The Return of the Condor Heroes	神鵰俠侶
Ode to Gallantry	俠客行
The Heaven Sword and Dragon Saber	倚天屠龍記
Sword Stained with Royal Blood	碧血劍
Blade-dance of the Two Lovers	鴛鴦刀