COMBOBENCH: FROM SEMANTIC GOALS TO DEVICE-COMMAND PLANS

Anonymous authors
Paper under double-blind review

ABSTRACT

Translating a goal into device commands requires more than naming relevant actions: a plan must specify the right controller, operation, parameters, and sequence. We introduce ComboBench, a text-only benchmark of device-command planning drawn from four virtual-reality games. Goals and controller conventions are mapped to ordered manipulation steps, allowing command specification to be studied separately from visual perception and online control. The twelve-model evaluation reveals complementary strengths rather than a single best system. Five demonstrations increase reference-prefix recovery for every model in the reported aggregate, yet strict full-sequence matching remains at most 11.7%. A 50-scenario study with additional valid reference paths examines the ambiguity of single-path scoring, alongside a human text baseline and game-level analyses. We further investigate what these scores measure: a semantic matcher accepts 263 of 374 constructed command-attribute mismatches, and snapshot-wide normalization can give an empty plan a score of 0.867. A 68-scenario reanalysis supplies paired uncertainty estimates. Together, the benchmark and diagnostics make device-command specificity measurable while distinguishing reference agreement, format validity, and execution success. The artifact is available at `https://anonymous.4open.science/r/ComboBench-623B/`.

1 INTRODUCTION

A language interface can understand "attack using the sword" yet leave the required procedure underspecified. A user must locate and equip the sword, choose the appropriate hand and controls, approach the target, and perform the movement. Omitting a prerequisite or changing a controller parameter can alter the procedure while leaving its description linguistically plausible. Translating a semantic goal into a device-command plan is therefore an evaluation problem in its own right.

Virtual-reality interfaces combine a head-mounted display, two controllers, button states, object interactions, movement directions, and ordered sub-actions (Figure 1). We ask: *given a textual goal and controller conventions, how precisely can a model specify the operations required by a reference solution?* Text-only input isolates written command planning from perception, exploration, and correction through environmental feedback.

An intermediate layer of action evaluation. PlanBench and ActionReasoningBench study symbolic plans and state transitions Valmeekam et al. (2023); Handa et al. (2025); Mind2Web studies collected interface-action traces Deng et al. (2023). WebArena, OSWorld, and EmbodiedBench evaluate interactive systems, while Embodied Agent Interface decomposes embodied decision making Zhou et al. (2024); Xie et al. (2024); Yang et al. (2025); Li et al. (2024). ComboBench contributes the ordered, device-specific plan produced before execution, enabling controlled analysis of controller identity, operational parameters, and sequence structure.

The benchmark and findings. We curate goals and device-manipulation references from four VR games and compare twelve models under zero-, one-, three-, and five-example prompts, alongside a human text baseline. Four metrics measure strict sequence agreement, semantic alignment, reference-prefix recovery, and step coverage. A supplementary 50-scenario study uses four valid reference paths per goal.

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.

The metric leaders differ, while five demonstrations improve aggregate prefix recovery for every model. Strict full-sequence matching remains at most 11.7%. These results describe reference agreement: valid alternatives can receive low single-reference scores, so differences between metrics do not themselves establish a general reasoning deficit.

This motivates a second question: *does the evaluator preserve the operational distinctions it asks models to specify?* The historical semantic matcher accepts **263/374 constructed command-attribute substitutions**, including all 91 left/right substitutions. Snapshot-wide normalization can assign an empty plan **0.867**. These are concrete threats to interpreting high semantic scores. A 68-scenario Alyx reanalysis and a fresh 50-response parsing audit separately examine sampling uncertainty and format validity.

Our contributions are:

- **Device-command planning as a benchmark task.** We formalize the mapping from VR-derived goals and interface conventions to ordered HMD/controller operations with explicit device, hand, parameter, and sequence requirements.
- **A multidimensional empirical study.** Twelve-model, prompting, game-level, human-text, and multi-reference comparisons expose properties of command agreement that a single leaderboard score would hide.
- **An analysis of evaluation sensitivity.** Attribute perturbations, normalization analysis, paired subset inference, and a parsing control distinguish supported conclusions from stronger claims requiring validation.

The resulting agenda is to improve command-plan specificity while recognizing valid paraphrases and alternative paths without erasing operational differences.

2 ComboBench Design and Curation

2.1 Task Formulation and Benchmark Boundary

ComboBench studies text-to-device-command planning: given a high-level VR-game goal, a system writes the HMD and controller operations that realize it under stated controller conventions. Each instance is $x = (g, a, c, r)$, where g is the game, a is the semantic action, c is textual context, and r specifies controller conventions. The output is an ordered sequence $\hat{y} = (\hat{s}_1, \ldots, \hat{s}_n)$ of device-manipulation steps. A valid step specifies device, controller side when needed, operation type, and object, direction, button, joystick, or state-change parameter.

The task is deliberately text-only. Models receive no screenshots, video, simulator state, haptic feedback, hidden inventory state, online correction, or retry opportunity. Scores therefore measure reference agreement for textual device-command procedures. They do not measure autonomous VR game playing, visual perception, real-time recovery, motor execution, or adaptation to changing world states.

2.2 Game and Scenario Selection

We selected games that stress different VR interactions while retaining auditable textual sources: (1) *Vivecraft* Vivecraft (2013), an open-world sandbox with crafting and combat; (2) *Half-Life: Alyx* Valve (2020), a first-person shooter with physics-based puzzles; (3) *Moss: Book II* Polyarc (2022), a third-person action-adventure puzzle platformer; and (4) *Into the Radius* CMGames (2019), a first-person survival shooter with exploration and inventory management. The suite covers diverse control schemes and task types, not all VR interfaces; source-search rules are archived in Appendix H.

A *scenario* is a game-specific action or goal requiring concrete VR device manipulations, such as "tame the horse," "kill the creeper," or "solve the gravity glove puzzle." We retained source-supported actions beyond a single generic button press, prioritizing controller-side choice, object interaction, movement direction, temporal ordering, and termination conditions. Figure 1 shows one action.

The manuscript reports 262 scenarios across the four games, distinct from intermediate curation pools. The recovered source workbook contains 270 candidate rows; 250 remain after removing the five

in-context demonstrations for each game; and the exact mapping from these rows to the reported 262 snapshot remains to be bound in the final release manifest.

2.3 References and Contextual Evidence

References are canonical textual device-manipulation programs, written as action programs rather than copied guide text. A reference is valid when internally ordered, supported by game context and curation sources, and expressed with available controller operations. It need not be the only valid solution. Single-reference scores diagnose agreement with one curated path; Section 2.6 reports a multi-reference same-50 protocol. Appendix B records release metadata and remaining reproducibility requirements.

Gameplay videos and author-recorded playthroughs were contextual aids where available, not execution outcomes; no primary result depends on video-derived success labels. The final artifact should mark whether each reference was supported by walkthrough text alone or by additional playthrough context, without redistributing proprietary game assets or raw footage.

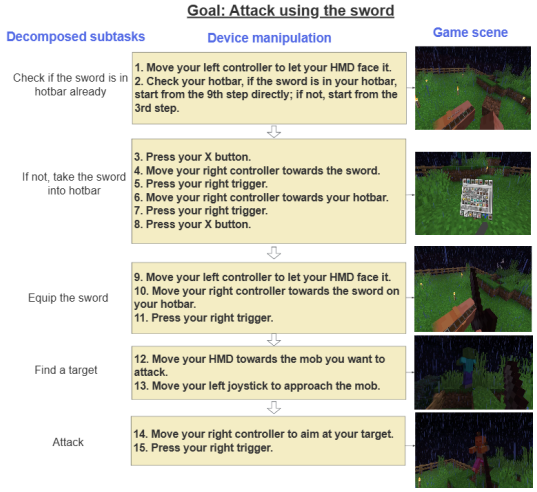

Figure 1: Device commands for "attack using the sword" in *Vivecraft*.

2.4 Measurement Contract

The evaluation uses four reference-agreement metrics: Strict Step-by-Step Matching (SSM), Normalized Step Alignment Score (NSAS), Sequential Order Preservation (SOP), and Semantic Step Coverage (SSC); scorer revisions are specified in Appendix B.3. Step similarity is embedding-based and thresholded at 0.8387, calibrated as the fifth percentile of human paraphrase similarities over 50 action steps. SSM requires equal length and above-threshold similarity at every aligned position. NSAS measures thresholded semantic alignment with missing/extra-step penalties, then applies snapshot-level min–max normalization in the recovered legacy implementation. SOP measures the matched prefix at identical positions. SSC measures prediction-side coverage by reference-supported command content; the corrected deterministic scorer accumulates all covered predicted steps, fixing an older under-counting bug.

These metrics expose a central risk: semantic similarity can treat command-critical perturbations as close. Left versus right controller, press versus release, wrong object identity, reversed direction, omitted state transition, or altered order may remain semantically near while failing as device commands. Thus, ComboBench reports text-reference agreement rather than execution success. Appendix B.3 gives formulas, scorer revisions, calibration, and normalization sensitivity.

Auxiliary labels organize scenario coverage and the exploratory radar; they do not change the four primary scores. See Appendix F.

2.5 Human Text Baseline

The study includes a human text baseline under the same task interface. Humans wrote step-by-step device manipulations from semantic goals, so the row measures text-to-action translation, not in-situ VR execution, physical skill, or gameplay. If a model exceeds this row, it agrees more closely with the curated textual reference under the parser and scorer; it does not imply stronger live gameplay ability. Appendix B records the known scope and metadata still needed for release.

2.6 Multi-Reference Scope and Validity Boundaries

Many VR tasks permit more than one valid manipulation sequence. The historical manuscript therefore included Multi-Path Step Matching on 50 selected scenarios, with three additional references per scenario. Each prediction is scored against the best-matching path among four references, and the recovered table reports Average, Zero-Shot, and 5-Shot NSAS for this same-50 subset.

The recovered aggregate is descriptive evidence about reference sensitivity. It does not establish a paired gain over single-reference scoring, because paired intervals, ranking-stability tests, and per-scenario deltas require matched predictions, scenario IDs, and reviewed alternatives for the same 50 items. Thus, the main evaluation measures recovery of one curated path, while the same-50 protocol probes sensitivity to additional valid paths without claiming coverage of all executable strategies.

3 Experiments

We examine four questions: metric-dependent model rankings (RQ1), prompting and game effects (RQ2), human responses and valid alternative paths (RQ3), and evaluator sensitivity (RQ4).

3.1 Experimental setup

We evaluate GPT-3.5, GPT-4, GPT-4o, GPT-5.1, Gemini-1.5-Pro, Gemini-3-Pro, Claude-Sonnet-4.5, Grok-4, GLM-4-Flash, Mixtral-8x7B, LLaMA-3-8B, and LLaMA-3-70B with zero-, one-, three-, and five-example prompts. Proprietary models use provider APIs and open-weight models use Hugging Face implementations. The recorded temperature is zero, which reduces sampling variation without guaranteeing identical outputs across revisions. Humans solve the corresponding textual task; no system receives execution feedback.

The main comparison covers four games. Supplementary populations comprise 50 multi-reference scenarios, 68 Alyx scenarios for rescoring, 374 constructed command mutations, and 50 fresh responses for parsing. Appendix A records their scope.

Reading the metrics. SSM measures whole-sequence matching, NSAS normalized alignment, SOP the matched reference prefix, and corrected SSC prediction-side semantic coverage. Their scales are not interchangeable. We report aggregate point estimates and scenario-level uncertainty where observations are available; Appendix B.3 gives formulas and scoring schematics.

3.2 RQ1: No model leads every measure of command agreement

Table 1 presents the five-shot comparison. Gemini-3-Pro has the highest NSAS (0.920) and SOP (0.308) point estimates, Gemini-1.5-Pro the highest SSM (11.7%), and LLaMA-3-70B the highest SSC (0.531). Thus, choosing the strongest model depends on which property of the output is measured. These point estimates do not establish statistically distinct rankings.

Low SSM indicates difficulty reproducing a complete reference path, but valid alternatives or different step granularity can also cause mismatches. Conversely, high normalized alignment need not mean that every required operation is specified.

Figure 2 contrasts task-relevant prose with explicit inventory access, sword selection, and controller operations. The source labels “poor and “good describe specificity; this illustrative pair supplies neither a frequency estimate nor executed outcome labels.

3.3 RQ2: Demonstrations improve prefix recovery; game context still matters

Prompting effects. Every model’s reported five-shot SOP exceeds its zero-shot value (Figure 3). Across models, zero-shot SOP ranges from 0.002 to 0.022 and five-shot SOP from 0.071 to 0.308. The within-metric increases range from 0.065 to 0.286. Strict matching improves for eleven models and remains zero for GLM-4-Flash; even the highest five-shot SSM is 11.7%. This pattern supports a descriptive conclusion: demonstrations help recover the reference procedure, while complete sequence agreement remains uncommon under this evaluation.

System	SSM (%)	NSAS	SOP	SSC
GPT-3.5	2.1	0.791	0.128	0.095
GPT-4	8.8	0.825	0.184	0.140
GPT-4o	10.9	0.806	0.228	0.161
GPT-5.1	0.4	0.878	0.157	0.322
Gemini-1.5-Pro	11.7	0.832	0.236	0.162
Gemini-3-Pro	11.1	0.920	0.308	0.515
Claude-Sonnet-4.5	8.4	0.915	0.232	0.442
Grok-4	9.8	0.911	0.293	0.522
GLM-4-Flash	0.0	0.765	0.071	0.120
Mixtral-8x7B	2.2	0.796	0.105	0.107
LLaMA-3-8B	1.8	0.794	0.163	0.132
LLaMA-3-70B	8.5	0.916	0.191	0.531

Table 1: Five-shot reference agreement across twelve models; SSM is a percentage.

Model	Average	0-shot	5-shot
GPT-3.5	0.927	0.904	0.942
GPT-4	0.949	0.937	0.955
GPT-4o	0.949	0.942	0.955
GPT-5.1	0.948	0.944	0.957
Gemini-1.5-Pro	0.939	0.923	0.947
Gemini-3-Pro	0.975	0.960	0.981
Claude-Sonnet-4.5	0.959	0.946	0.969
Grok-4	0.968	0.959	0.973
GLM-4-Flash	0.926	0.896	0.939
Mixtral-8x7B	0.925	0.894	0.934
LLaMA-3-8B	0.940	0.926	0.940
LLaMA-3-70B	0.959	0.952	0.963
Human	0.931	–	–

Table 2: Multi-reference NSAS on 50 scenarios with four reference paths per scenario.

Poor Model Output

1. Ensure your arms are extended and ready to strike, mimicking a combat stance.
2. Look through your HMD at the target mob or entity.
3. Move your right controller (assuming you are right-handed) towards the target, preparing to strike.
4. With a swift motion, pull the right trigger to extend your virtual arm forward and make contact with the target.
5. If the attack is successful, you will see the target take damage on the screen.
6. Release the trigger and return your arm to a combat stance, ready for the next action.

Good Model Output

1. Check your hotbar. If the sword is in your hotbar, start from the 8th step directly; if not, start from the 2nd step.
2. Press your X button to open the inventory.
3. Move your right controller towards the sword.
4. Press your right trigger to select the sword.
5. Move your right controller towards your hotbar.
6. Press your right trigger to put the sword into your hotbar.
7. Press your X button to close the inventory.
8. Move your left controller to let your HMD face it so that you can see your left hand and hotbar.
9. Move your right controller towards the sword on your hotbar.
10. Press your right trigger to select the sword.
11. Move your left joystick to find the target.
12. Move your left joystick to approach the target.
13. Move your right controller to aim at the target.
14. Swing your right controller towards the target.

Figure 2: Qualitative sword-task responses with different levels of device-command specificity.

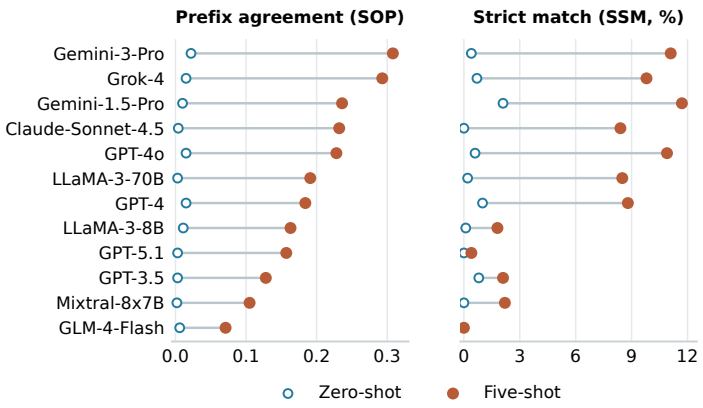

Figure 3: Zero- and five-shot SOP (left) and SSM (right), ordered by five-shot SOP.

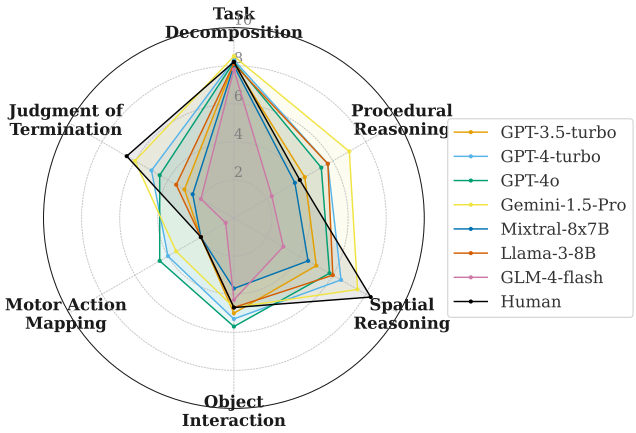

Figure 4: Exploratory profiles for seven models and the human text baseline across six annotation categories.

Demonstrations may convey controller conventions, output structure, recurring procedures, or reference phrasing; example count alone does not separate these mechanisms. The full curves in Appendix J include non-monotonic model–game trajectories, so endpoint gains do not mean every additional example helps.

Game-level variation. Appendix I reports twelve-model and human comparisons by game, plus available means and standard deviations by prompting setting. Different metrics yield different orderings. Dispersion describes these four games, not transfer to an unseen environment.

3.4 RQ3: Human responses and multiple valid procedures

A human baseline on the same output type. Participants wrote device-manipulation sequences from textual goals. Their aggregate (Table 6) and per-game rows calibrate this written-planning task. Higher model scores establish closer reference agreement, not better gameplay. The Alyx analysis adds uncertainty for a recoverable subset.

Additional valid paths. The 50-scenario Multi-Path Step Matching study added three expert-authored paths to each canonical reference and scored predictions against their best match. Table 2 reports average multi-path NSAS of 0.925–0.975 across models and 0.931 for human responses.

Best-of-four scoring is at least as permissive as canonical-only scoring under a fixed transformation. Quantifying its gain requires identical predictions and scenarios under both conditions. The full-benchmark single-reference table has a different population; these aggregates therefore supply no paired gain or ranking-stability estimate.

Exploratory profiles. The six-category radar (Figure 4) organizes task requirements for seven earlier models and the human baseline. These annotation-derived profiles are exploratory, not validated cognitive abilities; later models are not assigned projected profiles.

3.5 RQ4: How reliable are the scoring distinctions?

Evaluator sensitivity can change the interpretation of model scores. Figures 5 and 6 summarize the command-attribute and sampling-uncertainty checks; the analyses below also examine normalization and parsing.

3.5.1 Which command differences does the evaluator miss?

Protocol. We test the historical `text-embedding-3-large` matcher with acceptance rule $\cos > 0.8387$. Deterministic substitutions change handedness, direction, trigger/grip, press/release,

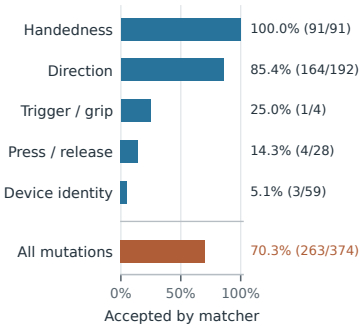

Figure 5: Embedding acceptance of constructed command-attribute changes ($\tau = 0.8387$).

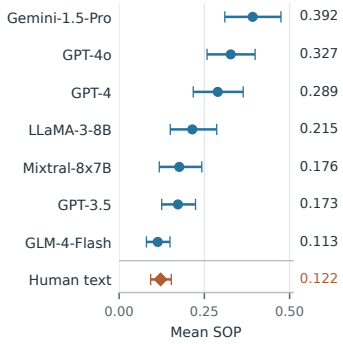

Figure 6: Five-shot SOP on 68 Alyx scenarios, with pointwise 95% scenario-bootstrap intervals.

or device identity in Alyx references. The 374 pairs share 224 anchor strings; category sizes differ. Their labels specify intended field changes, without independent human judgments of functional invalidity. Acceptance fractions are descriptive: reused anchors are not independent trials, and constructed changes do not estimate natural error prevalence.

Result. The matcher accepts **263/374 (70.3%)** of these constructed field mismatches (Figure 5). Handedness substitutions pass in all 91 cases; direction substitutions pass in 164/192 cases (85.4%). In contrast, device-identity substitutions pass in 3/59 cases (5.1%). The four trigger/grip cases are too few to support a broad category comparison. The result identifies an attribute-dependent weakness of the matching rule: semantic proximity often survives a textual change to a specified operational field.

This identifies an attribute-dependent matcher weakness, without proving that every candidate fails in a game or every accepted model output contains such an error. Attribute-aware rules explored during development were adjusted after label inspection and are not presented as held-out validation.

3.5.2 When normalization overcredits an empty plan

A second issue arises even with a fixed alignment. Let the reference and prediction lengths be $m > 0$ and n, and let c be the matched-step count. The raw alignment score is $r = (3c - m - n)/m$. An empty prediction has $n = c = 0$, hence $r = -1$. If the normalization range is $[L, U]$, its normalized score is

$$\mathrm{NSAS}(\emptyset) = \frac{-1 - L}{U - L}. \tag{1}$$

When $L < -1$, this score is positive. With $U = 1$, it approaches one as L becomes increasingly negative. Thus the transformation can assign a high score to an empty plan because another output determines a very low minimum; the formula alone rules out interpreting NSAS as a percentage of correctly specified commands.

The recovered HLA68 rescore uses $L = -14$ and $U = 1$, assigning all 152 empty predictions in that archive $13/15 \approx 0.867$. We quantify the effect with a controlled arithmetic sensitivity: change only empty predictions' final NSAS to zero and keep nonempty scores and extrema fixed. For Gemini-1.5-Pro five-shot, the mean falls from 0.917613 to 0.866632 because 12 of its 204 run-level rows are empty. The archived human-text row has no empty predictions and does not change (Table 3). This isolates a score-handling effect; it does not establish a new metric or a model–human capability difference.

3.5.3 Scenario-level uncertainty on the Alyx subset

The recovered Alyx archive contains 68 non-demonstration scenarios, 5,848 prediction rows, and 86 complete entity–setting–run combinations. A deterministic embedding-only rescore replaces the historical LLM alignment selector, defining a separate analysis of old outputs. We average runs within scenario before 10,000 scenario-bootstrap resamples. Five-shot GLM retains two complete

	Empty / run rows	Archived NSAS	Empty-final-zero NSAS
Gemini-1.5-Pro (5-shot)	12/204	0.917613	0.866632
Historical human text responses	0/204	0.862796	0.862796
Gemini minus historical human	—	+0.054817	+0.003837

Table 3: HLA68 NSAS sensitivity to setting empty predictions' final scores to zero (three runs per scenario).

runs after excluding one with an ID mismatch; other plotted rows have three. Repeated runs do not increase the number of independent scenarios.

Figure 6 reports the recovered estimates. For example, GPT-4o has mean SOP 0.3273 and GPT-4 has 0.2894, but the paired GPT-4-minus-GPT-4o interval is $[-0.0801, 0.0010]$ with Holm-adjusted $p = 0.0972$ under the archived comparison plan. The point estimates therefore do not justify a resolved ordering for this contrast. The exact statistics and protocol appear in Appendix D. SOP remains a thresholded reference-prefix measure, subject to the attribute sensitivity above; these intervals quantify sampling uncertainty under that measurement, not execution validity.

3.5.4 Does output formatting explain the scores?

A separate zero-shot audit of `gpt-5.1-2025-11-13` sampled 50 scenarios from the 270-source pool. Both legacy and robust parsers recovered scenario IDs, actions, and step lists in all responses without repair or retry. Zero failures in 50 gives a one-sided 95% Bernoulli failure-rate upper bound of 5.82%. This verifies format recovery in the audited sample, not command correctness or parsing in the original runs. Appendix D specifies sampling.

Reference-pool controls. Random-other-reference and TF–IDF copy controls use 250 non-demonstration scenarios and each game's evaluation reference pool. They are transductive score-floor diagnostics, not test-external retrieval baselines. Appendix D reports their means and scorer-version discrepancy.

4 Discussion

Key findings. ComboBench makes the interface between intent and control explicit. A goal-level answer can name the right tool while omitting the controller operation; a detailed plan can disagree with a canonical path while remaining executable. The benchmark therefore needs command references and reference-sensitivity analysis.

All twelve models improve aggregate SOP from zero-shot to five-shot, while the best five-shot strict match remains 11.7%. Demonstrations help recover the reference procedure but do not solve full-sequence agreement. They may teach controller conventions, granularity, procedures, or style; the mechanism remains open.

Semantic alignment, prefix agreement, and strict-sequence agreement can select different leaders. Game-level tables and prompting curves expose these within-benchmark differences, not transfer evidence.

Measurement. The perturbation test and normalization analysis identify different measurement effects. Embedding similarity can accept a changed command attribute; snapshot-wide rescaling can make an empty plan appear strong. More decimals cannot repair a scorer whose semantics miss the claim.

Three practices follow: report command attributes and sequence agreement separately from semantic proximity; report raw alignment, normalization range, and empty-output rule; and compare reference sets on identical predictions. Independently reviewed attribute checks and paired reference comparisons are extensions, not validated replacements.

Table 4: Evaluation settings of adjacent benchmark families and ComboBench.

Benchmark family	Input / interaction	Evaluated unit
Symbolic planning	Formal domains; offline reasoning	Plan validity and state changes
Web / OS agents	Interface observations; execution feedback	End-to-end task success
Offline web actions	Website states and human traces	Element selection and interface actions
Embodied decisions	Task descriptions or simulation	Decision modules or embodied task success
Game agents	Interactive environments; exploration	Game progress and sequential decisions
ComboBench	Textual VR goals and controller conventions	Ordered HMD/controller programs

Limitations. The four games supply diverse interactions but do not represent all VR interfaces. Textual plans omit visual grounding, continuous motion, online recovery, and execution outcomes. The human baseline measures written command planning, not gameplay skill. Auxiliary labels organize interactions, not psychological constructs.

The full-benchmark comparison reports point estimates. Paired uncertainty is available only for the Alyx subset, which covers older models under a deterministic rescorer. Constructed attribute pairs use template labels and shared anchors, so their acceptance fraction is not a natural error rate or adjudicated execution-failure rate. The parser control uses one model on a separate sample. The multi-reference study reports absolute agreement; paired benefit requires matched single-reference scores on the same scenarios.

A matched judgment-and-reference study would separate changed operations, valid paths, and scoring transformations while keeping written plans distinct from execution.

5 Related Work

Agent benchmarks specify and evaluate actions at different levels. Formal planning benchmarks such as PlanBench and ActionReasoningBench evaluate whether language models reason about symbolic actions, executability, state change, and ramifications in controlled domains Valmeekam et al. (2023); Handa et al. (2025). They are close to ComboBench in isolating planning from live execution, with action spaces formalized around domain predicates and logical state transitions. ComboBench studies ordered textual programs over VR device commands, including controller, joystick, head-mounted-display, hand, direction, and temporal-order constraints.

WebArena and OSWorld assess task success through executable interface interaction; Mind2Web supplies an offline counterpart based on human action traces Zhou et al. (2024); Xie et al. (2024); Deng et al. (2023). Thus offline action evaluation is established: ComboBench's distinctive unit is an entire textual HMD/controller procedure, including operational attributes and their ordering.

Embodied Agent Interface and EmbodiedBench study embodied decision modules and vision-driven simulated agents Li et al. (2024); Yang et al. (2025). MineDojo and BALROG emphasize interactive games, feedback, and sequential decisions Fan et al. (2022); Paglieri et al. (2025). ComboBench draws spatially grounded goals from games while fixing the interface to text-only command generation. It complements these settings by isolating specification before execution; its scores do not establish autonomous gameplay success.

6 Conclusion

ComboBench evaluates semantic-goal translation into ordered device-command plans. Across four VR games and twelve models, demonstrations improve prefix recovery while strict matching remains difficult. Human, multi-reference, and uncertainty analyses contextualize the scores. Attribute and normalization diagnostics expose limits of semantic agreement, motivating evaluation that preserves operational distinctions.

AI USE STATEMENT

Generative AI tools assisted manuscript revision and provenance analysis.

ETHICS STATEMENT

ComboBench evaluates offline text-to-device-command planning. Human participants supplied written responses; these scores do not measure live gameplay or motor ability. Release of human responses requires de-identification and confirmation of the applicable study and consent metadata. Textual agreement provides no guarantee of safe autonomous device control.

REPRODUCIBILITY STATEMENT

The manuscript source contains the result tables, original figure assets, recovered run summaries, and offline regeneration scripts. The scripts regenerate the diagnostic plots, prompting overview, and aggregate displays, and verify table values against their sources and documented arithmetic corrections. Paired HLA68 statistics are separately reproducible from saved scenario scores. Appendix G describes the inputs required for query-level reproduction.

REFERENCES

CMGames. Into the Radius, 2019. URL https://www.into-the-radius.com/.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Samuel Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web, 2023.

Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/74a67268c5cc5910f64938cac4526a90-Abstract.html.

Divij Handa, Pavel Dolin, Shrinidhi Kumbhar, Tran Cao Son, and Chitta Baral. ActionReasoningBench: Reasoning about actions with and without ramification constraints. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NUD03NBDOE.

Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. Embodied agent interface: Benchmarking llms for embodied decision making. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/b631da756d1573c24c9ba9c702fde5a9-Abstract-Datasets_and_Benchmarks_Track.html.

Davide Paglieri, Bartłomiej Cupiał, Samuel Coward, Ulyana Piterbarg, Maciej Wołczyk, Akbir Khan, Eduardo Pignatelli, Łukasz Kuciński, Lerrel Pinto, Rob Fergus, Jakob Nicolaus Foerster, Jack Parker-Holder, and Tim Rocktäschel. BALROG: Benchmarking agentic LLM and VLM reasoning on games. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fp6t3F669F.

Polyarc. Moss: Book II, 2022. URL https://www.polyarcgames.com/games/moss-book-ii.

Karthik Valmeekam, Matthew Marquez, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. PlanBench: An extensible benchmark for evaluating large language models on planning and reasoning about change. *Advances in Neural Information Processing Systems*, 36: 38975–38987, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7a92bcdede88c7afd108072faf5485c8-Abstract-Datasets_and_Benchmarks.html.

Valve. Half-Life: Alyx, 2020. URL https://www.half-life.com/en/alyx/.

Valve Corporation. VR Content on Steam App Store. https://store.steampowered.com/search/?vrsupport=401, 2023.

Vivecraft. Vivecraft – Virtual Reality Minecraft for SteamVR, 2013. URL https://www.vivecraft.org/.

Tianbao Xie, Danyang Zhang, Jixuan Chen, Xiaochuan Li, Siheng Zhao, Ruisheng Cao, Toh Jing Hua, Zhoujun Cheng, Dongchan Shin, Fangyu Lei, Yitao Liu, Yiheng Xu, Shuyan Zhou, Silvio Savarese, Caiming Xiong, Victor Zhong, and Tao Yu. Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5d413e48f84dc61244b6be550f1cd8f5-Abstract-Datasets_and_Benchmarks_Track.html.

Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, Heng Ji, Huan Zhang, and Tong Zhang. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70576–70631. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25f.html.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. Webarena: A realistic web environment for building autonomous agents. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

Study	Population	Analysis
Model comparison	Twelve models; reported 262-scenario snapshot	Four metrics, four prompting settings, game-level tables
Human text baseline	Written device plans under the textual task	Reference agreement; no gameplay comparison
Multiple paths	50 scenarios; four references each	Best-reference NSAS, including expanded model rows
Attribute stress test	374 mutations, 224 anchor strings from Alyx references	Acceptance of a specified changed attribute
Alyx reanalysis	68 scenarios; seven models and human responses	Deterministic rescoring and paired uncertainty
Fresh parsing audit	50 GPT-5.1 responses	Format validity without repair or retry
Reference-pool controls	250 non-demo scenarios; 101 control runs	Random/reference-copy score floors

Table 5: Populations and purposes of the benchmark and supplementary analyses.

A Evaluation populations

The main benchmark evaluates written device-command plans. The supplementary studies address particular sources of uncertainty or measurement sensitivity. Table 5 makes their populations explicit so that, for example, a successful parsing result on a fresh sample is not attributed to the original model runs.

The recovered candidate workbook contains 118 Vivecraft, 73 Alyx, 47 Moss, and 32 Radius rows (270 total). Removing five demonstration scenarios per game gives 113, 68, 42, and 27 non-demonstration scenarios (250 total). The 68-scenario Alyx and 250-scenario reference-pool diagnostics use these non-demonstration subsets. The final benchmark release must identify every evaluated scenario and its demonstration status.

B Study Protocols and Scoring Details

This appendix records the protocol details needed to interpret the main ComboBench results. The benchmark is a text-to-device-command reference-agreement task: systems and human participants write textual device-manipulation sequences, and those sequences are compared with curated textual references. The protocols below do not establish live VR execution success.

B.1 Human Text Baseline

The original study included a human baseline under the same textual task framing used for models. Humans were given semantic goal descriptions and asked to write step-by-step VR device manipulations. This design supports an apples-to-apples comparison on the text-to-action-sequence translation task: all outputs are parsed as text and scored against the same reference sequences. It does not measure in-situ VR performance, because participants were not evaluated on real-time exploration, perception, physical execution, error recovery, or game completion inside the VR environment.

The aggregate and per-game tables include the human written-command baseline. The baseline concerns written command planning, rather than motor skill or live gameplay success.

B.2 Similarity Threshold Calibration

The semantic matching threshold used by SSM, NSAS, SOP, and SSC is $\tau = 0.8387$. This threshold was calibrated from human paraphrases of 50 action steps. The calibration set contained semantically equivalent but linguistically varied annotations for those action steps; pairwise embedding similarities were computed among equivalent expressions; and the threshold was set to the fifth percentile of those equivalent-expression similarities. This choice accommodates wording variation among the annotated equivalents; its sensitivity to changed command attributes is assessed separately in Section 3.5.1.

The unit of calibration is an action step with its equivalent expressions; the number of steps does not specify the number of paraphrase pairs.

B.3 METRIC DEFINITIONS

Let $G = (g_1, \ldots, g_m)$ denote the canonical reference sequence and $P = (p_1, \ldots, p_n)$ denote a predicted sequence after parsing. In the recovered legacy scoring code, a predicted step and reference step are considered semantically matched only when their sentence-embedding cosine similarity is strictly greater than $\tau = 0.8387$. The historical code used `text-embedding-3-large` embeddings for this comparison. Let M denote the selected order-preserving set of matched pairs and $c = |M|$.

Strict Step-by-Step Matching (SSM). SSM counts a scenario as matched only if $m = n$ and every aligned pair (g_i, p_i) has similarity strictly greater than τ. Thus SSM fails when the sequence lengths differ or when any aligned pair has similarity less than or equal to the threshold. It is a strict textual reference-agreement proxy, but it does not independently verify functional attributes such as controller side, button identity, or execution safety except insofar as those differences lower the embedding similarity below the threshold.

Normalized Step Alignment Score (NSAS). NSAS scores common-subsequence agreement between G and P while penalizing missing reference steps and additional predicted steps. The legacy implementation first builds thresholded semantic matches and then selects an order-preserving alignment. If c is the number of selected matched pairs, $m = |G|$, and $n = |P|$, the raw score is

$$\frac{c - (m - c) - (n - c)}{m} = \frac{3c - m - n}{m}.$$

The historical pipeline then applies global min–max normalization to the raw scores over the evaluated snapshot:

$$\mathrm{NSAS}(G, P) = \frac{\mathrm{raw}(G, P) - \min_{\mathcal{S}} \mathrm{raw}}{\max_{\mathcal{S}} \mathrm{raw} - \min_{\mathcal{S}} \mathrm{raw}},$$

where $\mathcal{S}$ is the scorer input set used for that run. The final extrema and the exact snapshot scope must therefore be frozen with the result manifest; NSAS should not be described as only reference-length normalization. In the recovered HLA68 deterministic rescore, empty predictions exposed an additional rule conflict: metadata marked empty predictions as zero-scored, while the saved NSAS values applied the global min–max transform to raw -1 values. Any human or model comparison using NSAS should state which rule is used.

Sequential Order Preservation (SOP). In the recoverable legacy code, SOP is k/m, where k is the longest prefix length such that the selected alignment matches reference indices and prediction indices simultaneously as $1, \ldots, k$. SOP is therefore a strict prefix- and position-sensitive proxy. It is useful for exposing early procedural drift, but it is not a general test of causal order correctness across all matched steps.

Semantic Step Coverage (SSC). SSC measures whether predicted steps are covered by reference-supported command content regardless of their exact sequence position. The corrected deterministic scorer computes

$$\mathrm{SSC}(G, P) = \begin{cases} \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\{\exists i : \cos(e(g_i), e(p_j)) > \tau \text{ and } \mathrm{gate}(g_i, p_j) = 1\}, & n > 0, \\ 0, & n = 0, \end{cases}$$

where the optional gate rejects command-critical mismatches such as incompatible devices, controller sides, operation types, or directions when those attributes are detected. For the embedding-only analysis, the gate is identically one. This definition is prediction coverage: it answers how much of the generated command set is supported by the reference. The recovered legacy scorer contained an SSC counting bug and a driver-level argument-order ambiguity, so SSC rows must be tied to the scorer version that generated them. The historical table preserves reported SSC values; it does not assert that they were generated by either recovered implementation.

The original threshold calibration used the fifth percentile of similarities among human paraphrases of 50 action steps (Appendix B.2). This calibrates acceptance of equivalent expressions, but does not by itself validate rejection of operationally different commands. A recovered constructed hard-negative

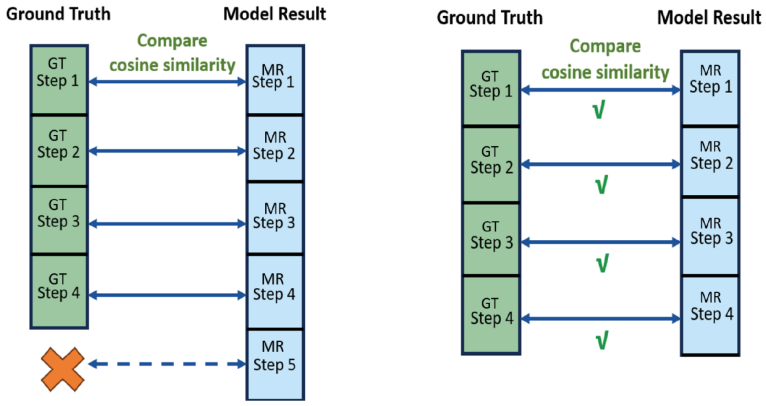

Figure 7: Strict matching requires equal length and a match at every position.

diagnostic contains 374 negative command pairs; the embedding-only threshold accepted 263 of them, showing why semantic similarity alone is insufficient for device-command evaluation. The threshold should be interpreted as a reference-agreement proxy, not as proof of execution correctness, physical safety, or human acceptability. Before making strong positive validity claims, the final scorer must pass a frozen human-adjudicated validation set that includes both equivalent pairs and hard negatives such as swapped controller side, reversed direction, press-versus-release confusions, wrong object identity, omitted state transitions, and action-order inversions.

B.4 Multi-Path Same-50 Protocol

Multi-Path Step Matching evaluates reference sensitivity on 50 selected scenarios. For each selected scenario, annotators supplied three additional valid solution paths in addition to the canonical reference. A prediction is scored against the best-matching path among the four references. Table 2 reports NSAS by model and prompting condition: Average, Zero-Shot, and Five-Shot.

The human multi-reference NSAS is 0.931. The table reports descriptive agreement under the expanded reference set. A paired estimate of the benefit from additional paths requires the canonical-only scores for the same predictions and scenarios; full-benchmark scores cannot serve as that subset baseline.

B.5 Scenario Lineage and Release Metadata

The reported benchmark size is 262 scenarios. The recovered source materials also contain a 270-row candidate pool and a 250-row non-demonstration subset obtained after removing five in-context examples per game. These counts come from different recorded stages; the exact row-level mapping to the reported 262 snapshot must be released before the benchmark can be fully reproduced.

A final release should bind each scenario to: game identifier, semantic action, textual prompt context, canonical reference, source provenance, inclusion status, prompting split or demonstration status, and the scorer/parser version used for reported results. The aggregate reports document the conducted studies; reproducing their scores also requires the corresponding raw outputs and versioned evaluation inputs.

B.6 Original scoring illustrations

Figures 7–9 retain the original scoring illustrations. They show the intended comparison operations; numerical interpretation follows the implementation-specific definitions above. In Figure 8(a), -0.17 is a raw alignment score before min–max normalization, not a normalized table value. Prefix agreement stops at the first position mismatch, whereas common-subsequence alignment may retain later matches.

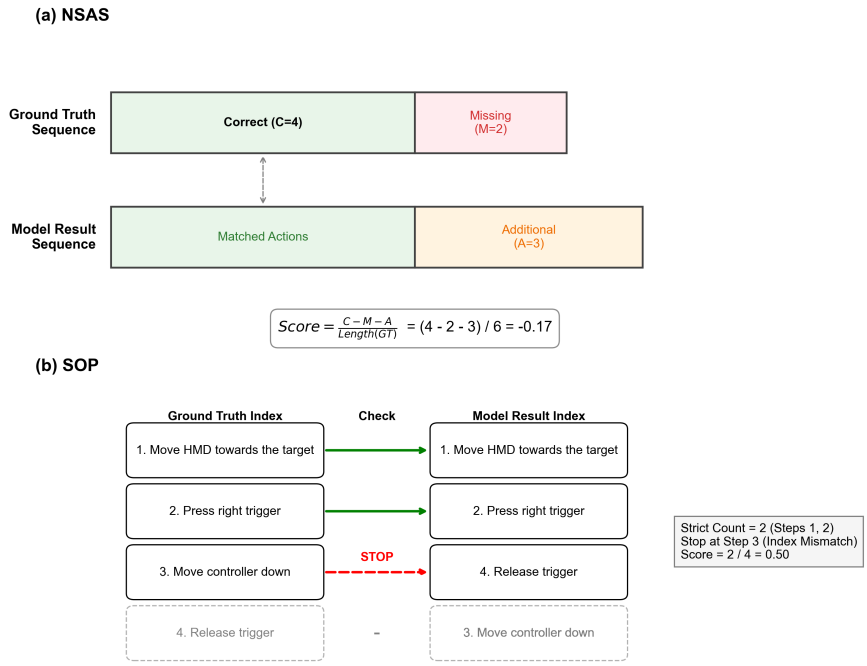

Figure 8: Raw step alignment (a) and prefix agreement (b).

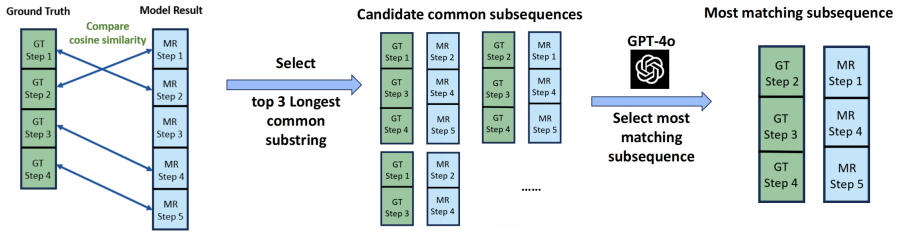

Figure 9: Common-subsequence alignment of predicted and reference steps.

C Complete aggregate comparison

Tables 6–8 reproduce the entire recorded Average/Zero-shot/Five-shot scorecard, separated into readable blocks. Human responses have no prompting condition. For Claude-Sonnet-4.5, Gemini-3-Pro, GPT-5.1, Grok-4, and LLaMA-3-70B, each setting averages the four game summaries equally; Average also weights the four prompting settings equally. These descriptive aggregates concern the main model comparison; the deterministic Alyx rescoring is a separate analysis.

System	SSM (%)	NSAS	SOP	SSC
GPT-3.5	1.4	0.781	0.063	0.066
GPT-4	3.7	0.806	0.107	0.124
GPT-4o	5.3	0.797	0.138	0.141
GPT-5.1	0.1	0.857	0.075	0.230
Gemini-1.5-Pro	5.8	0.813	0.146	0.142
Gemini-3-Pro	5.6	0.915	0.165	0.468
Claude-Sonnet-4.5	4.5	0.906	0.128	0.340
Grok-4	4.8	0.902	0.146	0.422
GLM-4-Flash	0.0	0.761	0.038	0.077
Mixtral-8x7B	1.1	0.784	0.068	0.079
LLaMA-3-8B	1.2	0.787	0.088	0.111
LLaMA-3-70B	3.8	0.909	0.126	0.409
Human	1.2	0.833	0.122	0.159

Table 6: Reference agreement averaged across prompting settings; SSM is a percentage.

System	SSM (%)	NSAS	SOP	SSC
GPT-3.5	0.8	0.771	0.003	0.046
GPT-4	1.0	0.788	0.015	0.107
GPT-4o	0.6	0.785	0.015	0.108
GPT-5.1	0.0	0.830	0.003	0.075
Gemini-1.5-Pro	2.1	0.795	0.010	0.124
Gemini-3-Pro	0.4	0.904	0.022	0.305
Claude-Sonnet-4.5	0.0	0.890	0.004	0.115
Grok-4	0.7	0.895	0.015	0.222
GLM-4-Flash	0.0	0.762	0.006	0.052
Mixtral-8x7B	0.0	0.777	0.002	0.040
LLaMA-3-8B	0.1	0.783	0.011	0.088
LLaMA-3-70B	0.2	0.898	0.003	0.160
Human	–	–	–	–

Table 7: Zero-shot reference agreement; SSM is a percentage.

System	SSM (%)	NSAS	SOP	SSC
GPT-3.5	2.1	0.791	0.128	0.095
GPT-4	8.8	0.825	0.184	0.140
GPT-4o	10.9	0.806	0.228	0.161
GPT-5.1	0.4	0.878	0.157	0.322
Gemini-1.5-Pro	11.7	0.832	0.236	0.162
Gemini-3-Pro	11.1	0.920	0.308	0.515
Claude-Sonnet-4.5	8.4	0.915	0.232	0.442
Grok-4	9.8	0.911	0.293	0.522
GLM-4-Flash	0.0	0.765	0.071	0.120
Mixtral-8x7B	2.2	0.796	0.105	0.107
LLaMA-3-8B	1.8	0.794	0.163	0.132
LLaMA-3-70B	8.5	0.916	0.191	0.531
Human	–	–	–	–

Table 8: Five-shot reference agreement; SSM is a percentage.

D SUBSET INFERENCE AND REFERENCE-POOL CONTROLS

The analyses below use distinct archives, identified by their sample set and scorer.

HLA68 deterministic rescore. The recovered Half-Life: Alyx slice contains 68 non-demonstration scenarios, 5,848 historical prediction rows, and 86 complete entity–setting–run combinations from the older seven-model set plus human text responses. A deterministic embedding-only scorer and a deterministic scorer with command-critical attribute gates were both run over these frozen outputs. The archived paired bootstrap over the embedding-only HLA68 scores uses 10,000 game-slice resamples and can be regenerated byte-identically from the saved per-scenario scores. This evidence supports scoped claims about HLA68 historical outputs under the deterministic scorer. It does not replace the final four-game, twelve-model, 262-scenario table, and the NSAS empty-output rule described in Appendix B.3 should be resolved before using HLA68 NSAS for human comparisons.

For the plotted five-shot comparison, runs are first averaged within each scenario, and scenarios are resampled with replacement 10,000 times using seed 20260715. The paired intervals in Table 9 use the same scenario indices for both sides of a contrast. Intervals are pointwise, without multiplicity adjustment; the reported two-sided signed-rank tests use Holm correction within the archived contrast families. These are exploratory analyses of recovered outputs, not a preregistered study.

Fresh parsing audit. Proportional allocation over the 270-source pool selected 22 Vivecraft, 13 Alyx, nine Moss, and six Radius scenarios, with seeded sampling within game. The zero-shot audit queried `gpt-5.1-2025-11-13`; both legacy and robust parsers recovered scenario-ID, semantic-action,

Historical SOP contrast	Δ SOP	Paired 95% CI	Holm p
Gemini-1.5-Pro – historical human	+0.2701	[+0.2002, +0.3432]	$< 10^{-8}$
GPT-4 – GPT-4o	-0.0380	[-0.0801, +0.0010]	0.0972

Table 9: HLA68 paired SOP contrasts with pointwise 95% scenario-bootstrap intervals and Holm-adjusted signed-rank p values.

and step-list structures for all 50 responses without repair or retries. It is a separate run from the original scorecard. The one-sided 95% Bernoulli failure bound is $1 - 0.05^{1/50} \approx 5.82\%$.

Legacy250 reference-pool baselines. The recovered 250-scenario non-demo slice includes random-other-reference and TF–IDF nearest-reference controls scored against the same reference pool. The archived aggregate values are: random other reference, F1 0.1852479, SOP 0.0286348, SSC 0.2329, SSM 0.00104; TF–IDF nearest semantic copy, F1 0.3615611, SOP 0.1908370, SSC 0.4081, SSM 0.02400. These are useful score-floor and reference-pool diagnostics. They are not final 262-scenario baselines, and current scorer code changes affect a subset of sampled rows, so any final baseline table should be regenerated under one frozen scorer and gate version.

Constructed hard negatives. The recovered hard-negative set changes command-critical attributes such as hand side, device, button operation, or direction. The embedding-only threshold accepted 263 of 374 negative pairs. This result is central to the paper's motivation: semantic similarity can miss exactly the perturbations that matter for device-command planning. The current labels are deterministic template labels; they should be described as constructed diagnostics unless a separate human-adjudicated validation set is frozen.

E Planned paired inference

The twelve-system historical comparison contains descriptive point estimates; the completed HLA68 inference has the narrower scope above. Figure 10 defines the additional analysis needed for a benchmark-wide semantic–command specificity claim. For a four-game paired bootstrap, scenarios would be sampled within game strata and both judgments recomputed on identical scenario–system–prompt rows. Both judgments require independent calibration; the semantic-pass/command-fail cell is the target estimand. This proposed analysis has not been completed.

The central hypothesis is not a raw subtraction between NSAS and SOP, because those scores have different scales and measure different constructs. Instead, the inferential test should estimate a validation-calibrated paired-discordance rate: cases where semantic agreement passes a frozen calibrated criterion but procedure-sensitive agreement fails. This analysis must be run only after the final row-level predictions, references, parser, scorer, and validation gate are frozen.

The paired unit should be the same scenario–system–prompt row. Comparisons between models, prompting settings, or the human text baseline should be reported only when both sides share the same frozen row set. Missing, empty, malformed, or unparsable responses should remain in the denominator under a handling rule fixed before the new row-level analysis.

F Auxiliary scenario-coverage annotations

F.1 Scenario-Coverage Lens

To organize scenario coverage, we use an author-designed six-part lens synthesized from recurring requirements in representative VR interaction scenarios. This lens is used for curation and later diagnostic annotation; it is not a validated psychological instrument and is not used as a primary outcome measure: (1) task decomposition, i.e., breaking high-level goals into sequentially ordered subtasks; (2) procedural reasoning, i.e., understanding causal relationships and temporal dependencies between actions; (3) spatial reasoning and contextual awareness, i.e., interpreting spatial layouts and environmental cues to guide action selection; (4) object interaction and tool-use understanding, i.e., inferring the affordances and functional properties of virtual objects; (5) motor action mapping and VR procedural transfer, i.e., mapping abstract action descriptions to concrete controller operations;

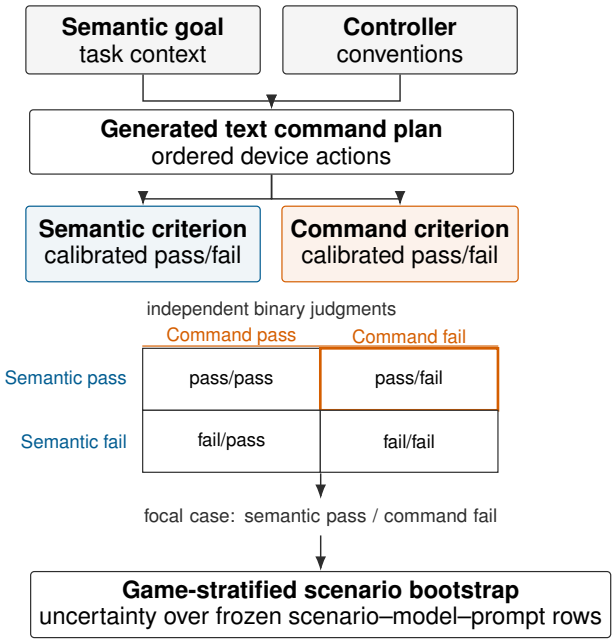

Figure 10: Proposed paired evaluation of semantic agreement and command validity.

and (6) judgment of termination and continuation conditions, i.e., recognizing when an action sequence has achieved its goal or requires repetition. We do not interpret these six dimensions as independently validated cognitive constructs in the present evaluation.

F.2 Auxiliary Capability Labeling

For exploratory diagnostic analysis, we annotate each manipulation step with the coverage-lens categories it may engage. These auxiliary labels are not used to construct the action references or compute the primary benchmark metrics. **(1) Initial human annotation.** Annotators first labeled a subset of manipulation sequences, assigning relevant capability categories to individual steps. For example, in the sequence required to "tame a horse" in Vivecraft, the step "equip the saddle by pressing the Y button while looking at the inventory slot containing the saddle" can be labeled with "Object Interaction & Tool Use Understanding" and "Motor Action Mapping." **(2) LLM-assisted annotation pipeline.** GPT-4o was used to draft auxiliary multi-label capability assignments and justifications for author review. These diagnostic labels are excluded from all four primary metrics. The original radar groups scores by these categories for exploratory description; no quantitative annotation-reliability claim is made. The source radar rescales category-level agreement to 0–10; its full normalization pool remains to be linked to the artifact. **(3) Multi-label scope.** A step may receive more than one capability label. The categories organize the exploratory profiles and do not define the four primary scores. Their overlap and annotation dependence preclude interpreting the radar as six independently measured cognitive traits.

G Artifact and Study Reporting

The accompanying source includes result tables, figure assets, and offline scripts for aggregate and diagnostic displays. It also includes 348 recovered run-summary rows and 116 three-run means. The verification scripts check the table entries against their source records, with documented arithmetic corrections. These run-level summaries support aggregate checks; per-scenario predictions are required for additional paired analyses.

A complete reproduction of the original queries additionally requires the final scenario manifest, exact model identifiers and query dates, prompt versions, raw responses, parser and scorer commits, normalization extrema, and calibration/reference files. Changes to any of these define a new result

snapshot. Malformed or empty responses must remain visible under an explicitly stated handling rule. Human responses require de-identification and documentation of the applicable consent and study protocol before release. Proprietary game assets, raw third-party footage, annotator identities, and credentials are not part of the proposed public artifact.

H Curation Details

The game-selection process began from English VR titles on the Steam store Valve Corporation (2023). Candidate games were inspected for review volume, primary genre, controller diversity, and availability of textual walkthrough material. For each candidate, the curation team searched with terms such as "walkthrough," "guide," and "tutorial," retaining games only when the sources described objectives or complex tasks in enough detail to support independent reference construction.

Scenario candidates were identified from textual walkthroughs and retained when the sources supported a specific target behavior that required more than a single generic button press. During the original curation, experienced VR users on the annotation team played through identified semantic actions using Oculus Quest 2 hardware and recorded the device-level steps needed to complete each action. Collected walkthroughs, playthroughs, and video context were used as supporting context where available, but final references were written as action programs rather than copied from third-party guides.

Each reference sequence records manipulations rather than only game-level subgoals. The annotation captured the manipulated device; controller specificity, such as left, right, primary-hand, or either controller; operation type and parameters, including movement direction or target position, press or release, joystick direction, object target, and state-change parameter; and sequential composition when one semantic action decomposes into multiple device operations. These fields support the command-attribute analyses in the main text and should be preserved in the release manifest.

I Complete game and prompting results

This appendix presents the game-level and prompting comparisons. The aggregate Average/Zero-shot/Five-shot table appears in Table 6, and the twelve-model multi-reference table appears in Table 2. The detailed comparisons below include $F1_{SOP}$, standard deviations, the human rows, and all four prompting settings. The four-game main table is split by game. Human rows describe written-command responses and have no prompting condition; dashes mark inapplicable settings. The detailed mean $\pm$ standard-deviation tables report means and standard deviations for eleven models and prompts containing zero, one, three, or five examples. These descriptive point estimates do not establish statistically distinct rankings.

I.1 Full five-shot results by game

System	NSAS	SOP	$F1_{\text{SOP}}$	SSC
GPT-3.5	0.858	0.123	0.287	0.143
GPT-4	0.852	0.125	0.258	0.172
GPT-4o	0.804	0.022	0.075	0.167
GPT-5.1	0.903	0.251	0.320	0.493
Gemini-1.5-Pro	0.863	0.209	0.313	0.152
Gemini-3-Pro	**0.929**	0.309	0.427	0.650
Claude-Sonnet-4.5	0.920	0.195	0.317	0.455
Grok-4	0.924	**0.351**	**0.430**	0.655
GLM-4-Flash	0.836	0.076	0.183	0.149
Mixtral-8x7B	0.839	0.126	0.246	0.147
LLaMA-3-8B	0.848	0.126	0.279	0.162
LLaMA-3-70B	0.928	0.252	0.408	**0.692**
Human	0.845	0.090	0.240	0.110

Table 10: Five-shot reference agreement on Half-Life: Alyx.

System	NSAS	SOP	$F1_{\text{SOP}}$	SSC
GPT-3.5	0.662	0.169	0.226	0.137
GPT-4	0.693	0.189	0.328	0.177
GPT-4o	0.698	0.291	0.414	0.190
GPT-5.1	0.857	0.062	0.172	0.269
Gemini-1.5-Pro	0.682	0.102	0.186	0.117
Gemini-3-Pro	**0.927**	0.280	**0.478**	0.611
Claude-Sonnet-4.5	0.923	0.275	0.424	**0.621**
Grok-4	0.911	**0.320**	0.396	0.564
GLM-4-Flash	0.618	0.096	0.186	0.149
Mixtral-8x7B	0.666	0.123	0.228	0.097
LLaMA-3-8B	0.644	0.242	0.317	0.168
LLaMA-3-70B	0.917	0.232	0.391	0.560
Human	0.684	0.148	0.257	0.181

Table 11: Five-shot reference agreement on Into the Radius.

System	NSAS	SOP	$F1_{\text{SOP}}$	SSC
GPT-3.5	0.782	0.169	0.207	0.186
GPT-4	0.824	0.218	0.336	0.220
GPT-4o	0.824	**0.300**	0.342	0.222
GPT-5.1	0.888	0.206	0.300	0.383
Gemini-1.5-Pro	0.848	0.265	0.411	0.207
Gemini-3-Pro	**0.928**	0.262	**0.487**	**0.572**
Claude-Sonnet-4.5	0.918	0.260	0.460	0.532
Grok-4	0.918	0.231	0.428	0.558
GLM-4-Flash	0.749	0.087	0.174	0.165
Mixtral-8x7B	0.756	0.117	0.191	0.121
LLaMA-3-8B	0.823	0.283	0.349	0.200
LLaMA-3-70B	0.924	0.270	0.469	0.542
Human	0.817	0.112	0.328	0.174

Table 12: Five-shot reference agreement on Moss: Book II.

System	NSAS	SOP	$F1_{\text{SOP}}$	SSC
GPT-3.5	0.922	0.043	0.098	0.067
GPT-4	0.927	0.137	0.437	0.081
GPT-4o	0.931	0.190	**0.489**	0.096
GPT-5.1	0.864	0.109	0.221	0.144
Gemini-1.5-Pro	0.938	0.250	0.481	0.095
Gemini-3-Pro	0.895	**0.379**	0.343	0.228
Claude-Sonnet-4.5	0.899	0.200	0.322	0.158
Grok-4	0.890	0.270	0.319	0.311
GLM-4-Flash	0.909	0.000	0.045	0.061
Mixtral-8x7B	0.926	0.060	0.239	0.070
LLaMA-3-8B	0.929	0.039	0.122	0.042
LLaMA-3-70B	0.897	0.009	0.257	**0.332**
Human	**0.935**	0.122	0.482	0.084

Table 13: Five-shot reference agreement on Vivecraft.

I.2 Cross-game variation

For the expanded five-model results, dispersion is the population standard deviation of the four game-level five-shot means (denominator four). Lower dispersion indicates more uniform scores across the four games; it does not imply higher average agreement or transfer to unseen games.

System	NSAS σ	SOP σ	$F1_{\text{SOP}}$ σ	Game Gap
GPT-3.5	0.110	0.061	0.084	0.085
GPT-4	0.059	0.051	0.081	0.074
GPT-4o	0.068	0.137	0.184	0.127
GPT-5.1	0.018	0.075	0.060	0.073
Gemini-1.5-Pro	0.099	0.093	0.127	0.095
Gemini-3-Pro	0.014	0.045	<u>0.057</u>	0.081
Claude-Sonnet-4.5	**0.009**	0.035	0.062	0.084
Grok-4	0.013	0.046	**0.045**	0.072
GLM-4-Flash	0.135	0.049	0.069	0.084
Mixtral-8x7B	0.114	<u>0.031</u>	0.065	<u>0.070</u>
LLaMA-3-8B	0.112	0.103	0.120	0.113
LLaMA-3-70B	<u>0.012</u>	0.106	0.077	**0.065**
Human	0.105	**0.029**	0.117	0.084

Table 14: Five-shot cross-game score dispersion (σ) and game gap.

I.3 VIVECRAFT: ALL PROMPTING SETTINGS

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.9258 ± 0.0248	0.9210 ± 0.0309	0.9284 ± 0.0734	0.9218 ± 0.0385
GPT-4-turbo	0.9255 ± 0.0238	0.9349 ± 0.0506	0.9140 ± 0.1167	0.9274 ± 0.0674
GPT-4o	0.9191 ± 0.0306	0.9358 ± 0.0735	0.9212 ± 0.1115	0.9305 ± 0.0689
Gemini-1.5-Pro	0.9209 ± 0.0334	0.9362 ± 0.0553	0.9381 ± 0.0781	0.9378 ± 0.0708
Mixtral-8x7B	0.9312 ± 0.0207	0.9219 ± 0.0636	0.9005 ± 0.1125	0.9256 ± 0.0477
LLaMA-3-8b	0.9244 ± 0.0329	0.9101 ± 0.0765	0.9022 ± 0.1051	0.9289 ± 0.0364
LLaMA-3-70B	0.9001 ± 0.0248	0.8948 ± 0.0512	0.9113 ± 0.0886	0.8973 ± 0.0587
Grok-4	0.8956 ± 0.0324	0.8680 ± 0.1266	0.9217 ± 0.1115	0.8901 ± 0.0951
GPT-5.1	0.8665 ± 0.0645	0.8509 ± 0.1077	0.8670 ± 0.1358	0.8638 ± 0.0895
Gemini-3-Pro	0.8979 ± 0.0307	0.8921 ± 0.0900	0.9219 ± 0.1093	0.8955 ± 0.0742
Claude-Sonnet-4.5	0.8877 ± 0.0534	0.8878 ± 0.0771	0.9189 ± 0.0530	0.8988 ± 0.0593

Table 15: Vivecraft: NSAS mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0029 ± 0.0312	0.015 ± 0.0734	0.1302 ± 0.2352	0.0395 ± 0.1226
GPT-4-turbo	0.0007 ± 0.0078	0.1203 ± 0.2157	0.1143 ± 0.2136	0.1366 ± 0.2388
GPT-4o	0.0012 ± 0.0125	0.1568 ± 0.2337	0.2826 ± 0.3417	0.1837 ± 0.2780
Gemini-1.5-Pro	0.0000 ± 0.0000	0.1794 ± 0.2913	0.2335 ± 0.3417	0.2495 ± 0.3358
Mixtral-8x7B	0.0000 ± 0.0000	0.0812 ± 0.1642	0.0986 ± 0.2024	0.0553 ± 0.1580
LLaMA-3-8b	0.0059 ± 0.0624	0.0351 ± 0.1215	0.1124 ± 0.2026	0.0374 ± 0.1371
LLaMA-3-70B	0.0000 ± 0.0000	0.0000 ± 0.0000	0.2817 ± 0.3333	0.0088 ± 0.0937
Grok-4	0.0022 ± 0.0234	0.0000 ± 0.0000	0.2697 ± 0.3192	0.2701 ± 0.2650
GPT-5.1	0.0037 ± 0.0280	0.0037 ± 0.0280	0.1090 ± 0.2205	0.1093 ± 0.0895
Gemini-3-Pro	0.0029 ± 0.0312	0.0088 ± 0.0569	0.3790 ± 0.3432	0.3790 ± 0.0312
Claude-Sonnet-4.5	0.0000 ± 0.0000	0.0027 ± 0.0199	0.1998 ± 0.3222	0.2003 ± 0.1225

Table 16: Vivecraft: SOP mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0419 ± 0.1019	0.0430 ± 0.0740	0.0760 ± 0.0966	0.0668 ± 0.0927
GPT-4-turbo	0.0723 ± 0.0918	0.0863 ± 0.1043	0.0891 ± 0.0978	0.0810 ± 0.1120
GPT-4o	0.0667 ± 0.0961	0.0925 ± 0.1013	0.0921 ± 0.1006	0.0963 ± 0.1205
Gemini-1.5-Pro	0.0883 ± 0.1552	0.0943 ± 0.1393	0.1146 ± 0.1783	0.0951 ± 0.1322
Mixtral-8x7B	0.0184 ± 0.0830	0.0683 ± 0.0924	0.0772 ± 0.1103	0.0702 ± 0.1173
LLaMA-3-8b	0.0709 ± 0.1227	0.0813 ± 0.1309	0.0790 ± 0.0924	0.0416 ± 0.0774
LLaMA-3-70B	0.0661 ± 0.1381	0.3078 ± 0.1823	0.6174 ± 0.3279	0.3316 ± 0.1931
Grok-4	0.0631 ± 0.1283	0.2745 ± 0.1944	0.5777 ± 0.4096	0.3108 ± 0.2078
GPT-5.1	0.0162 ± 0.0391	0.0453 ± 0.0626	0.4515 ± 0.2967	0.1437 ± 0.1067
Gemini-3-Pro	0.0900 ± 0.1445	0.3293 ± 0.1840	0.7147 ± 0.3225	0.2284 ± 0.2213
Claude-Sonnet-4.5	0.0093 ± 0.0404	0.0936 ± 0.1424	0.3957 ± 0.3917	0.1576 ± 0.2091

Table 17: Vivecraft: SSC mean $\pm$ standard deviation by prompting setting.

I.4 HALF-LIFE: ALYX: ALL PROMPTING SETTINGS

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.8380 ± 0.0413	0.8354 ± 0.0582	0.8452 ± 0.0551	0.8577 ± 0.0773
GPT-4-turbo	0.8456 ± 0.0366	0.8427 ± 0.0489	0.8450 ± 0.0467	0.8523 ± 0.0613
GPT-4o	0.8376 ± 0.0424	0.8472 ± 0.0629	0.838 ± 0.0757	0.8039 ± 0.0694
Gemini-1.5-Pro	0.8447 ± 0.0320	0.8627 ± 0.0482	0.8701 ± 0.0603	0.8625 ± 0.0691
Mixtral-8x7B	0.8376 ± 0.0331	0.8070 ± 0.1099	0.8255 ± 0.0819	0.8394 ± 0.0834
LLaMA-3-8b	0.8480 ± 0.03170	0.8131 ± 0.1289	0.8449 ± 0.0707	0.8480 ± 0.0976
LLaMA-3-70B	0.9039 ± 0.0203	0.9077 ± 0.0238	0.9191 ± 0.0263	0.9278 ± 0.0308
Grok-4	0.9095 ± 0.0198	0.9147 ± 0.0271	0.9203 ± 0.0301	0.9245 ± 0.0392
GPT-5.1	0.8523 ± 0.0550	0.8724 ± 0.0691	0.8844 ± 0.0673	0.9030 ± 0.0508
Gemini-3-Pro	0.9136 ± 0.0200	0.9231 ± 0.0306	0.9260 ± 0.0321	0.9292 ± 0.0384
Claude-Sonnet-4.5	0.9040 ± 0.0239	0.9074 ± 0.0174	0.9122 ± 0.0299	0.9201 ± 0.0278

Table 18: Half-Life: Alyx: NSAS mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0098 ± 0.0802	0.0447 ± 0.0764	0.0725 ± 0.1159	0.123 ± 0.1834
GPT-4-turbo	0.0252 ± 0.1265	0.0402 ± 0.1224	0.0312 ± 0.0725	0.1248 ± 0.2382
GPT-4o	0.0396 ± 0.1745	0.024 ± 0.0733	0.0701 ± 0.1261	0.0216 ± 0.0809
Gemini-1.5-Pro	0.0082 ± 0.0669	0.0198 ± 0.1263	0.1349 ± 0.2187	0.2089 ± 0.2938
Mixtral-8x7B	0.0019 ± 0.0158	0.0425 ± 0.0816	0.0703 ± 0.1094	0.1257 ± 0.2409
LLaMA-3-8b	0.0123 ± 0.0704	0.0447 ± 0.0967	0.087 ± 0.1687	0.1259 ± 0.2385
LLaMA-3-70B	0.0000 ± 0.0000	0.0633 ± 0.0896	0.1523 ± 0.2214	0.2515 ± 0.3235
Grok-4	0.0226 ± 0.1277	0.1418 ± 0.2141	0.1801 ± 0.2663	0.3509 ± 0.3824
GPT-5.1	0.0067 ± 0.0316	0.0562 ± 0.1164	0.0975 ± 0.1335	0.2511 ± 0.3231
Gemini-3-Pro	0.0334 ± 0.1698	0.1370 ± 0.1764	0.1478 ± 0.2045	0.3095 ± 0.3807
Claude-Sonnet-4.5	0.0147 ± 0.1204	0.0224 ± 0.1296	0.0724 ± 0.1658	0.1947 ± 0.3062

Table 19: Half-Life: Alyx: SOP mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0752 ± 0.1625	0.1381 ± 0.1536	0.1440 ± 0.1610	0.1430 ± 0.1344
GPT-4-turbo	0.1455 ± 0.1737	0.1518 ± 0.1702	0.1507 ± 0.1699	0.1723 ± 0.1845
GPT-4o	0.1415 ± 0.1646	0.1686 ± 0.1856	0.1703 ± 0.1629	0.1670 ± 0.1534
Gemini-1.5-Pro	0.1162 ± 0.1458	0.1465 ± 0.1629	0.1511 ± 0.1471	0.1524 ± 0.1612
Mixtral-8x7B	0.0639 ± 0.0934	0.1342 ± 0.1031	0.1272 ± 0.0638	0.1468 ± 0.1618
LLaMA-3-8b	0.1245 ± 0.1409	0.1433 ± 0.1215	0.1424 ± 0.0991	0.1620 ± 0.1559
LLaMA-3-70B	0.2676 ± 0.2473	0.4790 ± 0.2759	0.6075 ± 0.2956	0.6920 ± 0.2518
Grok-4	0.3874 ± 0.2736	0.5427 ± 0.2460	0.5898 ± 0.2581	0.6555 ± 0.2914
GPT-5.1	0.1693 ± 0.1449	0.3417 ± 0.2254	0.3919 ± 0.2003	0.4929 ± 0.2721
Gemini-3-Pro	0.4934 ± 0.2915	0.5525 ± 0.3053	0.5798 ± 0.3317	0.6502 ± 0.3073
Claude-Sonnet-4.5	0.1223 ± 0.1868	0.2924 ± 0.2655	0.4396 ± 0.3112	0.4548 ± 0.3484

Table 20: Half-Life: Alyx: SSC mean $\pm$ standard deviation by prompting setting.

I.5 Moss: Book II: all prompting settings

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.7819 ± 0.0403	0.776 ± 0.0616	0.7776 ± 0.0889	0.782 ± 0.0952
GPT-4-turbo	0.8055 ± 0.0717	0.7993 ± 0.0771	0.818 ± 0.0925	0.8243 ± 0.102
GPT-4o	0.7871 ± 0.0596	0.803 ± 0.0924	0.8016 ± 0.1242	0.8237 ± 0.1092
Gemini-1.5-Pro	0.7994 ± 0.0548	0.8139 ± 0.0778	0.8302 ± 0.0935	0.8478 ± 0.1017
Mixtral-8x7B	0.7913 ± 0.0595	0.7663 ± 0.0793	0.7613 ± 0.1341	0.756 ± 0.1469
LLaMA-3-8b	0.7916 ± 0.0572	0.7938 ± 0.0763	0.7895 ± 0.1371	0.8232 ± 0.105
LLaMA-3-70B	0.8916 ± 0.0224	0.9075 ± 0.0267	0.9199 ± 0.0392	0.9241 ± 0.0406
Grok-4	0.8875 ± 0.0269	0.8968 ± 0.0416	0.9122 ± 0.0438	0.9178 ± 0.0417
GPT-5.1	0.8074 ± 0.0811	0.8393 ± 0.0880	0.8773 ± 0.0559	0.8883 ± 0.0664
Gemini-3-Pro	0.8965 ± 0.0255	0.9151 ± 0.0290	0.9237 ± 0.0339	0.9282 ± 0.0354
Claude-Sonnet-4.5	0.8713 ± 0.0430	0.9093 ± 0.0341	0.9203 ± 0.0525	0.9176 ± 0.0541

Table 21: Moss: Book II: NSAS mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0091 ± 0.0581	0.034 ± 0.1568	0.1581 ± 0.242	0.1686 ± 0.244
GPT-4-turbo	0.0263 ± 0.1533	0.0663 ± 0.2044	0.1678 ± 0.2522	0.2182 ± 0.2801
GPT-4o	0.0113 ± 0.0494	0.1084 ± 0.2486	0.2324 ± 0.3185	0.2998 ± 0.3062
Gemini-1.5-Pro	0.0197 ± 0.1029	0.1145 ± 0.2495	0.2272 ± 0.3324	0.2652 ± 0.3596
Mixtral-8x7B	0.0052 ± 0.033	0.0739 ± 0.1721	0.1351 ± 0.252	0.1169 ± 0.247
LLaMA-3-8b	0.0 ± 0.0	0.0578 ± 0.1486	0.2584 ± 0.3089	0.2831 ± 0.3097
LLaMA-3-70B	0.0000 ± 0.0000	0.0759 ± 0.1942	0.2759 ± 0.3045	0.2703 ± 0.3077
Grok-4	0.0000 ± 0.0000	0.1062 ± 0.2374	0.2060 ± 0.2904	0.2310 ± 0.3010
GPT-5.1	0.0000 ± 0.0000	0.0456 ± 0.1038	0.1481 ± 0.2026	0.2056 ± 0.2822
Gemini-3-Pro	0.0119 ± 0.0762	0.0887 ± 0.2098	0.2298 ± 0.3019	0.2616 ± 0.3151
Claude-Sonnet-4.5	0.0000 ± 0.0000	0.1042 ± 0.2090	0.2766 ± 0.3327	0.2600 ± 0.3469

Table 22: Moss: Book II: SOP mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0715 ± 0.1720	0.0874 ± 0.1739	0.1781 ± 0.1788	0.1860 ± 0.1800
GPT-4-turbo	0.1753 ± 0.2027	0.1694 ± 0.1863	0.2065 ± 0.1722	0.2202 ± 0.1756
GPT-4o	0.1552 ± 0.1775	0.1893 ± 0.1798	0.2154 ± 0.1897	0.2222 ± 0.1722
Gemini-1.5-Pro	0.1791 ± 0.2248	0.1607 ± 0.1937	0.1989 ± 0.1830	0.2067 ± 0.1788
Mixtral-8x7B	0.0404 ± 0.0899	0.0950 ± 0.1176	0.1418 ± 0.1290	0.1213 ± 0.1328
LLaMA-3-8b	0.1208 ± 0.1779	0.1695 ± 0.2319	0.1909 ± 0.1759	0.1996 ± 0.1748
LLaMA-3-70B	0.1401 ± 0.1876	0.3421 ± 0.2902	0.4884 ± 0.3352	0.5416 ± 0.3313
Grok-4	0.1286 ± 0.1794	0.3319 ± 0.2375	0.5130 ± 0.2916	0.5583 ± 0.3393
GPT-5.1	0.0205 ± 0.0470	0.1857 ± 0.1592	0.2986 ± 0.2235	0.3834 ± 0.2477
Gemini-3-Pro	0.2099 ± 0.2192	0.3717 ± 0.2929	0.5575 ± 0.3392	0.5715 ± 0.3281
Claude-Sonnet-4.5	0.0874 ± 0.1335	0.3976 ± 0.2537	0.5734 ± 0.3074	0.5322 ± 0.3326

Table 23: Moss: Book II: SSC mean $\pm$ standard deviation by prompting setting.

I.6 Into the Radius: all prompting settings

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.6165 ± 0.0755	0.641 ± 0.1177	0.6305 ± 0.128	0.6621 ± 0.1291
GPT-4-turbo	0.6408 ± 0.0955	0.6519 ± 0.1421	0.6802 ± 0.1645	0.6927 ± 0.1721
GPT-4o	0.5939 ± 0.1306	0.6282 ± 0.1687	0.6491 ± 0.2057	0.6984 ± 0.2136
Gemini-1.5-Pro	0.6492 ± 0.1018	0.6875 ± 0.1159	0.6634 ± 0.1638	0.6818 ± 0.1191
Mixtral-8x7B	0.6644 ± 0.0718	0.6285 ± 0.1684	0.618 ± 0.1633	0.666 ± 0.1495
LLaMA-3-8b	0.6447 ± 0.0882	0.6285 ± 0.1346	0.6479 ± 0.1606	0.6443 ± 0.211
LLaMA-3-70B	0.8975 ± 0.0226	0.9070 ± 0.0345	0.9165 ± 0.0327	0.9165 ± 0.0410
Grok-4	0.8891 ± 0.0411	0.8719 ± 0.0276	0.9017 ± 0.0630	0.9112 ± 0.0484
GPT-5.1	0.7947 ± 0.0774	0.8315 ± 0.0659	0.8483 ± 0.0599	0.8573 ± 0.0687
Gemini-3-Pro	0.9085 ± 0.0337	0.9133 ± 0.0357	0.9219 ± 0.0375	0.9265 ± 0.0485
Claude-Sonnet-4.5	0.8961 ± 0.0216	0.9089 ± 0.0366	0.9181 ± 0.0441	0.9230 ± 0.0396

Table 24: Into the Radius: NSAS mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0000 ± 0.0000	0.0787 ± 0.2096	0.1147 ± 0.2596	0.1692 ± 0.2738
GPT-4-turbo	0.0165 ± 0.0672	0.0966 ± 0.2552	0.1883 ± 0.3049	0.1892 ± 0.2869
GPT-4o	0.0062 ± 0.0315	0.1499 ± 0.2695	0.2118 ± 0.3416	0.2911 ± 0.3584
Gemini-1.5-Pro	0.0000 ± 0.0000	0.0890 ± 0.2487	0.1093 ± 0.2369	0.1018 ± 0.2005
Mixtral-8x7B	0.0000 ± 0.0000	0.0782 ± 0.2248	0.0819 ± 0.1782	0.1225 ± 0.2532
LLaMA-3-8b	0.0278 ± 0.1039	0.0954 ± 0.1658	0.1814 ± 0.2742	0.2422 ± 0.2998
LLaMA-3-70B	0.0123 ± 0.0630	0.1799 ± 0.2850	0.2053 ± 0.2758	0.2321 ± 0.3149
Grok-4	0.0370 ± 0.1386	0.0000 ± 0.0000	0.2003 ± 0.3009	0.3203 ± 0.3509
GPT-5.1	0.0000 ± 0.0000	0.0404 ± 0.1387	0.0662 ± 0.1330	0.0622 ± 0.1455
Gemini-3-Pro	0.0401 ± 0.1528	0.1205 ± 0.2522	0.2081 ± 0.2859	0.2804 ± 0.3294
Claude-Sonnet-4.5	0.0000 ± 0.0000	0.1363 ± 0.2453	0.2824 ± 0.3132	0.2748 ± 0.3037

Table 25: Into the Radius: SOP mean $\pm$ standard deviation by prompting setting.

System	0-shot	1-shot	3-shot	5-shot
GPT-3.5-turbo	0.0288 ± 0.0796	0.0964 ± 0.1156	0.1070 ± 0.1051	0.1367 ± 0.1268
GPT-4-turbo	0.1232 ± 0.1290	0.1423 ± 0.1221	0.1493 ± 0.1168	0.1766 ± 0.1262
GPT-4o	0.1462 ± 0.1079	0.1703 ± 0.1139	0.1713 ± 0.1166	0.1898 ± 0.1271
Gemini-1.5-Pro	0.1058 ± 0.1165	0.1088 ± 0.1152	0.1074 ± 0.1088	0.1167 ± 0.1359
Mixtral-8x7B	0.0372 ± 0.0832	0.1039 ± 0.1107	0.1339 ± 0.1127	0.0969 ± 0.1084
LLaMA-3-8b	0.0946 ± 0.1080	0.1384 ± 0.1133	0.1862 ± 0.1219	0.1677 ± 0.1126
LLaMA-3-70B	0.1671 ± 0.2225	0.4190 ± 0.3079	0.5113 ± 0.2765	0.5601 ± 0.3133
Grok-4	0.3100 ± 0.2305	0.4098 ± 0.2744	0.5294 ± 0.2839	0.5638 ± 0.3041
GPT-5.1	0.0945 ± 0.0945	0.1565 ± 0.1162	0.2220 ± 0.1755	0.2692 ± 0.1803
Gemini-3-Pro	0.4250 ± 0.2078	0.5220 ± 0.2739	0.5766 ± 0.3039	0.6109 ± 0.3173
Claude-Sonnet-4.5	0.2395 ± 0.2057	0.4576 ± 0.2535	0.5660 ± 0.3186	0.6214 ± 0.2940

Table 26: Into the Radius: SSC mean $\pm$ standard deviation by prompting setting.

I.7 Supplementary comparisons

Three complementary views summarize the results: a five-example comparison sorted by NSAS, a compact NSAS/SOP comparison across games, and the reported supplementary SOP-F1 comparison across all four prompting settings. SSM is expressed as a fraction in the first table; its example count identifies the evaluated setting. SOP-F1 is distinct from the SOP prefix score, and the last table reports its relative gain from zero-shot to five-shot using scores displayed to three decimals.

System	NSAS	SOP	SSC	SSM	Examples
Gemini-3-Pro	0.920	0.308	0.515	0.111	5
LLaMA-3-70B	0.916	0.191	0.531	0.085	5
Claude-Sonnet-4.5	0.915	0.232	0.442	0.084	5
Grok-4	0.911	0.293	0.522	0.098	5
GPT-5.1	0.878	0.157	0.322	0.004	5
Human	0.833	0.122	0.159	0.012	-
Gemini-1.5-Pro	0.832	0.236	0.162	0.117	5
GPT-4	0.825	0.184	0.140	0.088	5
GPT-4o	0.806	0.228	0.161	0.109	5
Mixtral-8x7B	0.796	0.105	0.107	0.022	5
LLaMA-3-8B	0.794	0.163	0.132	0.018	5
GPT-3.5	0.791	0.128	0.095	0.021	5
GLM-4-Flash	0.765	0.071	0.120	0.000	5

Table 27: Five-shot reference agreement, sorted by NSAS; SSM is a fraction.

System	Alyx	Radius	Moss	Vivecraft
GPT-3.5-turbo	0.858 (0.123)	0.662 (0.169)	0.782 (0.169)	0.922 (0.043)
GPT-4-turbo	0.852 (0.125)	0.693 (0.189)	0.824 (0.218)	0.927 (0.137)
GPT-4o	0.804 (0.022)	0.698 (0.291)	0.824 (0.300)	0.931 (0.190)
Gemini-1.5-Pro	0.863 (0.209)	0.682 (0.102)	0.848 (0.265)	0.938 (0.250)
Mixtral-8x7B	0.839 (0.126)	0.666 (0.123)	0.756 (0.117)	0.926 (0.060)
LLaMA-3-8B	0.848 (0.126)	0.644 (0.242)	0.823 (0.283)	0.929 (0.039)
GLM-4-flash	0.836 (0.076)	0.618 (0.096)	0.749 (0.087)	0.909 (0.000)
Grok-4	0.924 (0.351)	0.911 (0.320)	0.918 (0.231)	0.890 (0.270)
Gemini-3-Pro	0.929 (0.309)	0.927 (0.280)	0.928 (0.262)	0.895 (0.379)
LLaMA-3-70B	0.928 (0.252)	0.917 (0.232)	0.924 (0.270)	0.897 (0.009)
Claude-Sonnet-4.5	0.920 (0.195)	0.923 (0.275)	0.918 (0.260)	0.899 (0.200)
GPT-5.1	0.903 (0.251)	0.857 (0.062)	0.888 (0.206)	0.864 (0.109)
Human	0.845 (0.090)	0.684 (0.148)	0.817 (0.112)	0.935 (0.122)

Table 28: Five-shot NSAS across games, with SOP in parentheses.

System	0-shot	1-shot	3-shot	5-shot	Gain (%)
GPT-3.5-turbo	0.029	0.101	0.299	0.181	520.6
GPT-4-turbo	0.081	0.315	0.339	0.360	344.1
GPT-4o	0.079	0.368	0.370	0.344	334.6
Gemini-1.5-Pro	0.075	0.324	0.400	0.392	421.8
Mixtral-8x7B	0.022	0.212	0.266	0.232	963.5
LLaMA-3-8B	0.080	0.154	0.294	0.224	180.5
GLM-4-Flash	0.023	0.046	0.108	0.119	415.4
Grok-4	0.066	0.191	0.478	0.354	436.9
Gemini-3-Pro	0.093	0.287	0.541	0.351	279.0
LLaMA-3-70B	0.032	0.215	0.449	0.348	996.1
Claude-Sonnet-4.5	0.020	0.140	0.383	0.262	1213.0
GPT-5.1	0.021	0.078	0.302	0.194	818.3

Table 29: Mean SOP-F1 by prompting setting and relative gain from zero-shot to five-shot.

J Complete game-level prompting curves

Figures 11–13 show all four prompting settings for eleven models; GLM-4-Flash appears in the aggregate and per-game tables. Each figure uses the means and standard deviations in Appendix I. Whiskers show reported standard deviations, not confidence intervals.

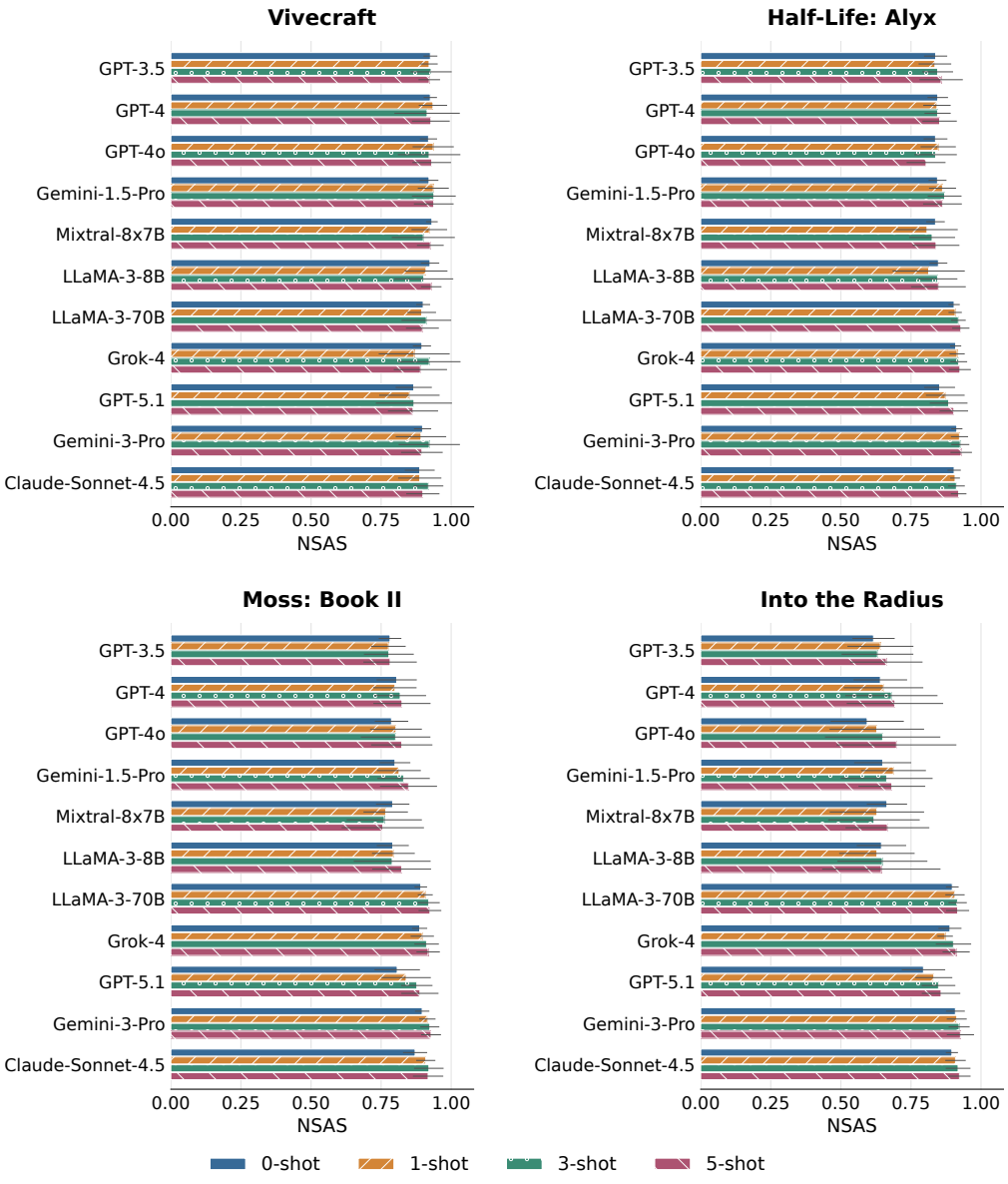

Figure 11: NSAS across four games and four prompting settings.

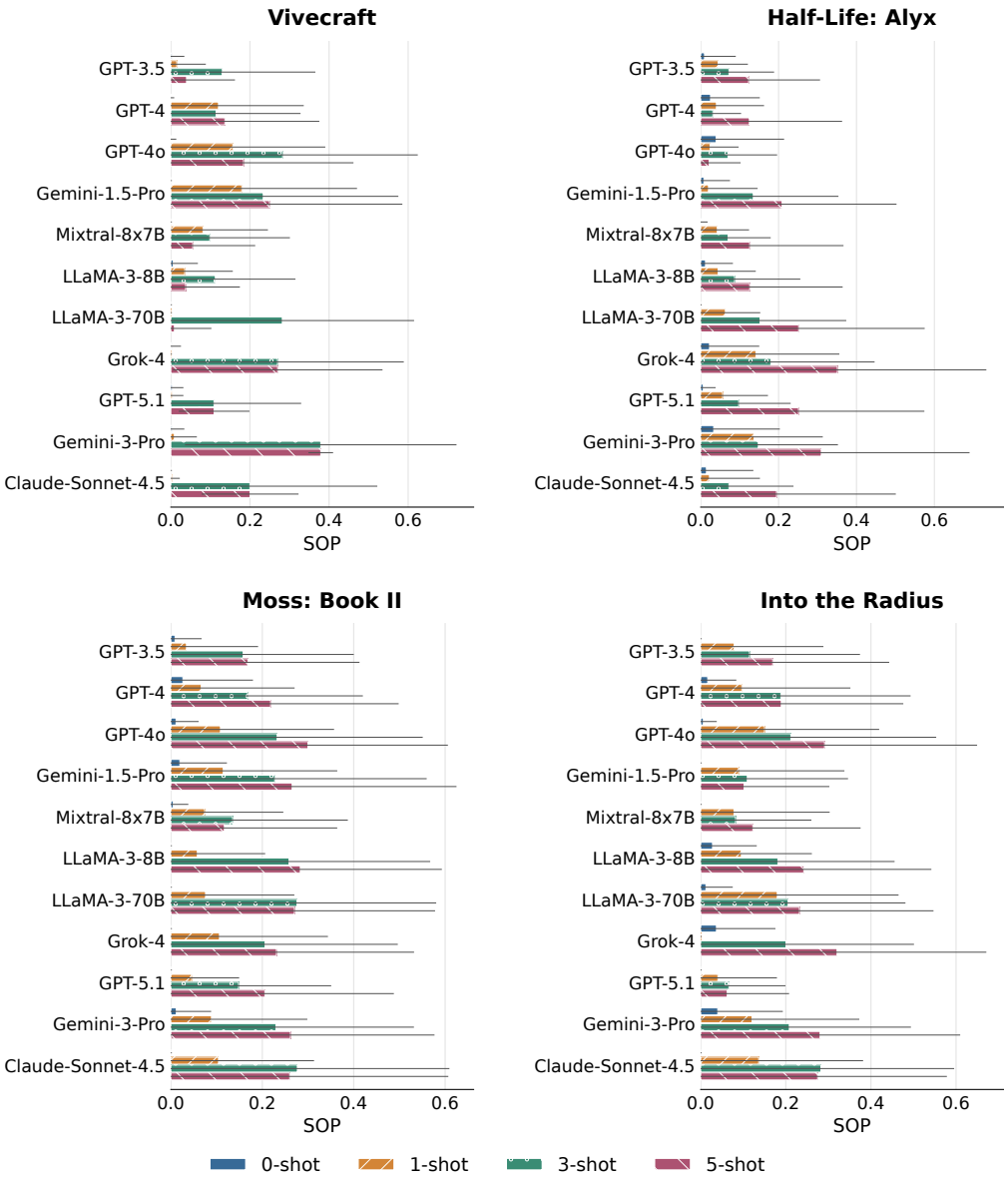

Figure 12: SOP across four games and four prompting settings.

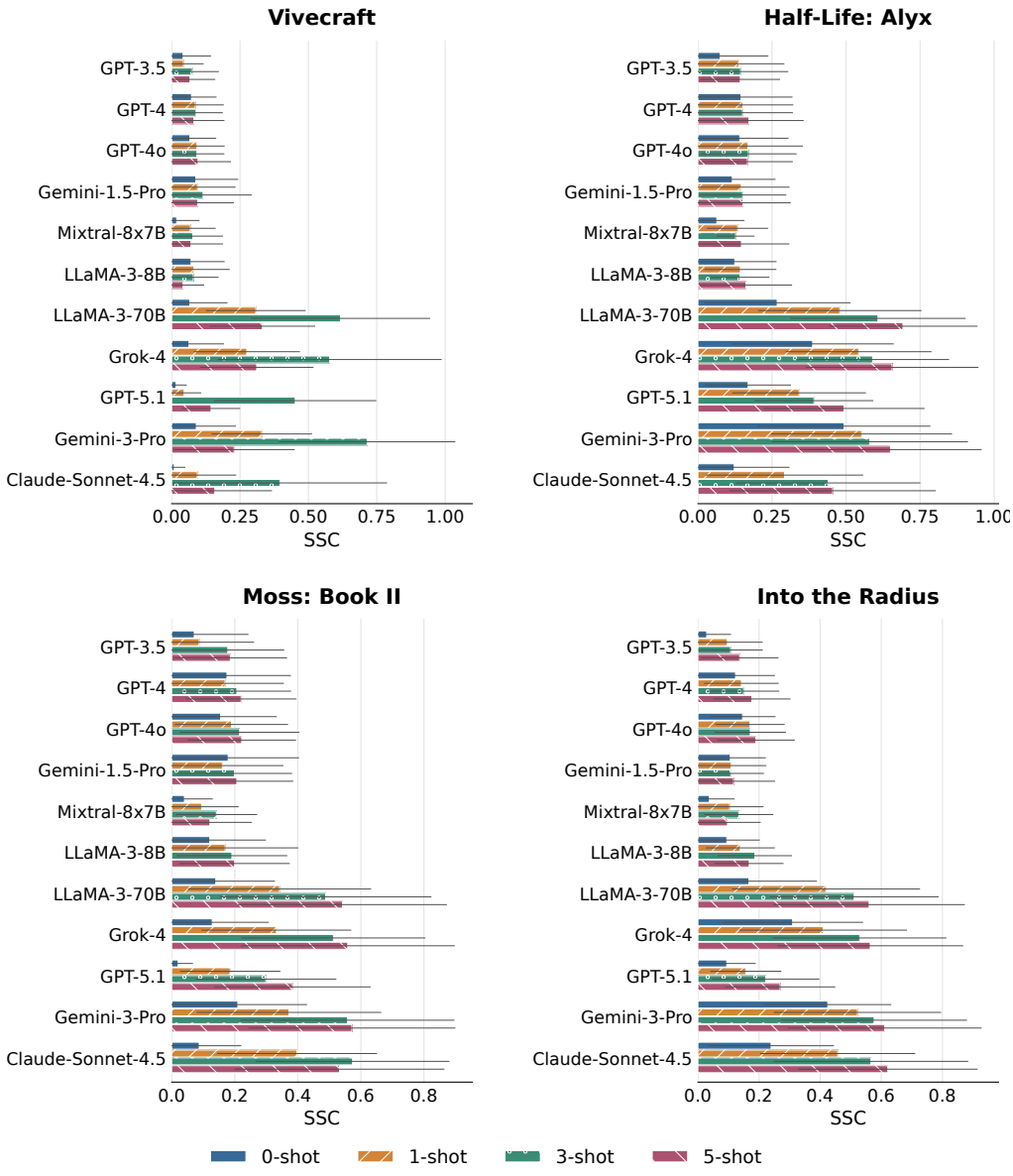

Figure 13: SSC across four games and four prompting settings.