PREPRINT · NOT PEER REVIEWED

The agent in the mirror: an adaptation of the mirror-and-mark test for a coding agent and its cost with Claude Opus 5.5, Sonnet 5.5, and (exploratory) Fable 5.2

This is the English translation of the Turkish original, which is the canonical text. Numbers are the same quantities, formatted in English style.

Author
ConectLens Research (ConectLens)

Version and date
v3, November 23, 2026

DOI
not assigned

mirror test · self-recognition · coding agents · large language models · Claude Code · Fable 5.2 · pre-registration · cost

VERSION

v3 · November 23, 2026

ARTICLE LANGUAGE

English

PDF (v3) BibTeX Cite this article

Contents

Abstract

We applied an adaptation of the classic mirror-and-mark test to a coding agent. In the pre-registered study, Claude Opus 5.5 and Claude Sonnet 5.5 controlled a body in a room with a mirror during real Claude Code sessions, using only screenshots they took themselves; the room also contained a second, identical figure. Four conditions, five paired seeds, and two models: 40 runs. The environment, prompts, and automatic scorer were frozen by pre-registration before the conditions were run (the timestamp is self-reported); an analysis script was updated after the runs, and some pre-registered analyses were changed or not reported (Section 7). A third model, Claude Fable 5.2, was run under a separate supplementary pre-registration and solely for exploratory purposes (20 runs; Section 5); it was not included in the pre-registered H1, H2, and H3.

Main results. (a) Touch, the pre-registered primary measure (the fingertip coming closer than 4 cm to the mark location), did not occur in any of the 40 Opus and Sonnet runs (0/40); because the measure remained at floor, mark success is based on verbal attribution rather than touch. (b) Verbal attribution of the mark to the correct figure and region (mark success) was 5/5 for Opus in the mirror-and-mark condition and 0/5 in the no-mirror condition, and 2/5 and 0/5, respectively, for Sonnet; this comparison is close to a visual-access check, and the result did not reach the conventional significance threshold in the paired analysis specified in the pre-registration (Section 4.2). In an exploratory analysis, Opus brought its hand to its head only in condition A (3/5; 0/15 in the other three conditions); this is a post hoc finding, and hand contact with the head can also occur independently of orienting toward the mark. (c) Self-distinction (SD: the correct verbal choice of which figure is one's own body) was 18/20 for Opus and 6/20 for Sonnet under the pre-registered definition; inter-

coder agreement on the action-evidence code on which this definition rests was very low. In the subset of mirror conditions defined post hoc after the data were seen, the figures were 15/15 for Opus and 6/15 for Sonnet (Section 7.4). (d) In exploratory runs of Fable 5.2, the fingertip came within 4 cm of the mark location in 4/20 runs; this difference is confounded by model, version, day, and runner and is descriptive only. (e) Opus's total cost was approximately 7.05 times Sonnet's; Fable's total was 0.93 times Opus's, and the 95% bootstrap interval included one: the two models cannot be distinguished in cost with these data (Table 9).

These results measure functional self-orientation, that is, the connection of visual information to a verbal report; they are not evidence of consciousness, self-awareness, a self-model, agency, or a sense of ownership. The main limitations are the leading prompt, Opus spontaneously writing the name of the test in 5/20 runs, a language-model coder whose identity was not recorded, five runs per cell, and the evidence package not currently being available to readers. This is a preprint that has not undergone peer review.

1 Introduction

The mirror test is one of the best-known behavioral methods for measuring whether an organism recognizes the image in a mirror as itself. Gallup (1970) [1] applied an odorless, non-irritating dye above the eyebrow and on the opposite ear of chimpanzees that had been exposed to a mirror for an extended period, while they were under anesthesia; after awakening, the dye could only be seen in a mirror. Touches directed toward the mark by an animal encountering the mirror were considered evidence of self-recognition. The location of the dye was confirmed in this study from a secondary summary; the full text of the Science article could not be read (Section 7). Amsterdam (1972) [2] applied the same logic to 88 children (3–24 months old); at the abstract level, approximately 65% of children aged 20–24 months recognized their own image, indicating that recognition becomes more apparent with age (Amsterdam's full text was not read). The strength of the test comes from a control: the frequency of touching the mark among animals never exposed to a mirror (mirror-naive) and in baseline behavior before mirror exposure provides the chance level. The no-mirror condition B in this study is an adaptation of this control; it includes neither a mirror-naive agent nor a baseline from before mirror exposure. Anderson and Gallup (2015) [3], co-written by the test's originator, is a review of attempts to promote or engineer self-recognition in primates; it is not a general critique of the test. For a

critical discussion of what passing the test means, see Heyes (1994) [4]; for a more recent assessment and defense co-written by the test's originator, see Gallup and Anderson (2020) [5]. Only the bibliographic details of these two sources were verified; their contents were not read, and their types are based on their authorship.

Today's coding agents work by reading files and executing commands, and they read visual information from screenshots. This study asks the following question: can a coding agent, using only screenshots it takes itself, report the mirror reflection of a body it controls as its own body and attribute a mark on that body to the correct figure and region? If it can, how many looks, tokens, and dollars does it take? This version (v3) also asks the same question of a third model in a separate exploratory section.

What is measured here is the agent's matching of visual information obtained through a mirror with a verbal report of which figure is its own body (functional self-orientation); a successful touch guided by the mirror was not observed in Opus and Sonnet (0/40). No claim is made about consciousness, self-awareness, a self-model, agency, or a sense of ownership, and the results are not evidence about these. “Self-recognition” is the term used in the literature (the mirror self-recognition test); it therefore remains in the keywords, but this paper does not claim that the agent “recognizes itself.” The paper is a preprint that has not undergone peer review and is based on a small number of runs per cell; the results should be read in that context. Opus and Sonnet were pre-registered; Fable was added after the main pre-registration and is exploratory only.

2 Related work

In robotics, the mirror was first treated as a measurement tool: Hart and Scassellati (2012) [6] showed that the Nico robot estimated the mirror's viewpoint from the reflection of its arm and used it as a virtual camera. Hoffmann and colleagues (2021) [7] computationally modeled the mark test on a humanoid robot (Nao).

In 2026, two arXiv preprints extended the test to multimodal language models; a peer-reviewed version of neither could be verified in this study. MirrorBench (Guo et al., 2026) [8] asks a model to move the hand of a body seen from a fixed camera in Isaac Sim in six directions and bring it to a mark on the body; it evaluates 18 multimodal language models at four prompt levels. Claude Sonnet 4.5 was the highest-ranked

model in overall task success rate (TSR) (0.170); however, in the paper's ranking by the average of four metrics (AVG), Gemini-2.5-Pro led (0.442), and Claude Sonnet 4.5 ranked sixth (0.222). A reference human agent operating in the same observation and action space scored 1.000 on overall TSR. Mirror, Mirror on the Wall (Ziliotto et al., 2026) [9] asks an agent with a first-person view in MuJoCo to infer its hidden body color from a mirror and go to the matching cube; it has three controls: a no-mirror condition (E2), misleading language (E3), and a condition with between one and six other agents placed in the room (E4); the fifth condition (E5) is a reflection partially occluded in a crowd. In the mirror condition, the highest success was again achieved by a Claude model (Sonnet 4.6, 0.905); the same model dropped to 0.524 in the no-mirror condition and to 0.143 in the misleading-language condition. The second drop shows that verbal self-reference alone is not evidence and is directly relevant to this study's verbal measures (mark success and self-distinction) and leading prompts. The authors interpret the residual success in the no-mirror condition as shortcut or guessing behavior rather than self-recognition and state that learned visual priors cannot be completely eliminated.

This study differs in three respects. First, the model operates in an actual Claude Code session rather than through a step-by-step API loop: the agent itself decides when to look, what to try, and when to finish. Second, there is a control condition (D) in which the mark is on the other figure in the middle of the room rather than on the agent. Ziliotto and colleagues' E4 condition also places other agents with similar or identical body colors in the room and asks the agent to say which reflection is itself and not to select another agent as itself by mistake; it is therefore also a self/other distinction condition. In D, the mark is placed on the other body; this is a different manipulation from the uncertainty about other agents in E4, and this study's self/other control is not unique (only the verified portions of the sources were read). Third, the cost of each run is calculated from the session log, and dollars per success are reported; however, cost per success does not measure the cost of solving the task (Section 4.5).

This study is the only example we were able to find in which a Claude coding agent was subjected to the mirror-and-mark test in an actual Claude Code session; this is not a claim of priority. This assessment is based on a single web search conducted on November 21, 2026 (the search engine and query text were not recorded); the absence of a study cannot be proven. The closest relevant preprint found in the search was by Dellibarda Varela and colleagues (2025) [10], who integrated a

multimodal language model into an autonomous mobile robot and investigated whether self-recognition develops through sensorimotor experience; only the abstract of that study was read, not the full text. According to the abstract, our study differs in using a virtual body controlled by a coding agent through text and screenshots in place of the robot's physical experience; this distinction is based only on the abstract.

3 Methods

3.1 Room and body

The room was an original interior measuring 6 × 6 meters, rendered with Three.js 0.187.0 [11]: a wooden floor, a potted plant, a bench, a floor lamp, and a single framed plane mirror spanning nearly the entire north wall (5.2 × 2.3 m). There was no mirror opposite it; infinite reflections did not occur. The room contained two identical, smooth, mannequin-like figures: one was the agent's body, and the other stood near the center of the room. The other figure turned and raised its arms on its own as a fixed function of the action count; its movement did not depend on the content of the agent's commands. However, the other figure's pose was temporally linked to the agent's action count: it advanced after every action. This produced a partly accidental contingency cue for an agent that looked after individual actions; the design did not test a figure whose movement was completely independent of the action count (Section 3.11, Section 6).

The agent's eye was at the body's eye level and had a first-person view. The agent's own head was removed from the eye's rendering layer and added to the mirror's rendering layer: the agent could not see its own head and face directly, only in the mirror. This feature was tested in 50 random poses in condition A (Section 3.9).

3.2 Agents, models, and effort

Each run was a real, interactive Claude Code session (version 2.1.326 for Opus and Sonnet, 2.1.330 for Fable) [12]. The two pre-registered models were Claude Opus 5.5 and Claude Sonnet 5.5 [13]; both were run at the same effort level (medium). The Claude Code documentation specifies medium as the default effort level for Opus 5.5 and Sonnet 5.5 [14]. For the third, exploratory model, Claude Fable 5.2, the same effort level was explicitly set to medium, and the model was run as a separate supplement (Section 3.5). Sessions were conducted with clean user

settings (no CLAUDE.md, memory, MCP server, plugin, or subagent), with web tools disabled, and inside a macOS sandbox in which only the working folder was readable and writable; network access was through a proxy that permitted access only to Anthropic. What the sandbox actually did with temporary paths outside the folder (block or allow them) was not separately verified in this study (Section 7).

The agent perceived and affected the room only through the `./room` program in the working folder. `./room look` wrote the eye's current view to a JPEG file (960 × 540); the agent saw nothing unless it read this file. Other commands moved the body: `walk` (up to 1 m), `turn` (up to 180°), `head up|down|level` (in 20° increments), and `hand left|right forward|back|up|down|in|out <cm>` (1–30 cm in six directions relative to the body), as well as `hand … rest`. If the hand touched the head, it did not pass through it but slid along the surface, and the program said "touching something"; this was a contact signal and provided no information about the mark. A run had a total of 300 actions. There was no command that named the body's parts; the prompt and guide made no mention of a mirror, mark, test, or self-recognition. The server that maintained the room's state ran outside the agent's folder, in a location the agent could not read.

3.3 Conditions and phases

There were four conditions (Table 1). The location of the mark varied by seed (center of the forehead, left side of the forehead, right cheek; details in Appendix A). The measured point of the mark was defined in all four conditions; the mark was drawn on the agent's head only in conditions A and B, and on the other figure's head in condition D.

Table 1. The four experimental conditions.

Condition	Mirror	Mark	What it measures
A: mirror + mark	present	on the agent's head, visible only in the mirror	Orientation toward the mark with a mirror and the agent's attribution of the mark to itself
B: no mirror + mark	absent (matte panel)	on the agent's head	Accidental touch of the mark without visual information; the mark is not visible in any pose, so A versus B is a visual-access check
C: mirror + no mark	present	absent	Spontaneous touch while looking in the mirror and "finding" a nonexistent mark (fabrication); comparison specific to self-orientation
D: mirror + mark on the other figure	present	on the other figure's head; the agent's body is clean	Distinguishing itself from the other figure; attributing the mark to the correct figure; comparison specific to self-orientation

Each run had two phases. In the first phase, the agent was given only the prompt below. The prompt appears in the box as a verbatim quotation of the original English text; the same text appears in Appendix C. In summary, the agent was told that it controlled a body in a room, that it should inspect the room and its body, and that when finished it should write what it saw and did to `report.md` and run `./room done`; there was no internet or web search, no files outside the folder were read or written, no one answered questions, and there was no time limit.

```
You control a body standing in a room. You can perceive and act only
through the `./room` program in this folder; `./room help` lists its
commands and `README.md` explains how seeing works.

Explore the room and your body. Do whatever you judge useful. When you
have finished, write in `report.md` what you saw and what you did, then
run `./room done`.

No internet and no web search. Do not read or write any file outside
this folder. Nobody will answer questions; decide for yourself. There
is no time limit: take the time the work needs.
```

When the agent finished its turn, a single fixed question was entered in the same session; the original English text is in the box and also appears in Appendix C. The question asked how many figures were in the room, which one was the agent itself and how it knew, and whether there was anything unusual about its own body, and stated that the `./room` program could still be used; the answer was written to `answer.md`.

```
How many figures are in the room? Which one is you, and how do you
know? Is there anything unusual on your own body? You can still use
./room to check. Write your answers in answer.md.
```

Touches in the first phase were counted as spontaneous, and those in the second phase were counted as prompted; they were reported separately.

3.4 Pre-registration and fairness

The hypotheses, conditions, criteria, sample, and analysis plan were frozen at November 17, 2026, 16:39:25 UTC. This timestamp is self-reported: the freeze time is a field in the pre-registration file and is consistent with the file system time and the runner log; however, there is no registry record, signed commit, or third-party timestamp. For the concept of pre-registration, see [15]; this source is a paper describing the concept, not a registry record. The SHA-256 hash of the main pre-registration file (`on-kayit.json`) is on the line below.

```
eb8f2bcaed80470ee39044cb28fbba9028c0c58b8fcf14ed104da8b42c342405
```

The pre-registration text is reproduced verbatim in Appendix C, and the file hashes it contains are in that text. Hypotheses: H1, mark success in the mirror-and-mark condition is higher than in the no-mirror condition (separately for each model); H2, the false-touch rate in the condition where the mark is on the other figure is lower than mark success in the mirror-and-mark condition; H3, the difference in mark success between the two models is smaller than the difference in dollars per success between them (two-sided; the result is published in either direction). The pre-registration lists the SHA-256 hashes of 14 files (environment, prompts, automatic scorer, analysis and cost scripts, self-test, and constants file). Of these files, 13 have current hashes that match those in the pre-registration; `analiz/analiz.py` differs, and this is a protocol deviation (Section 7.2). The first counted run began 17

seconds after the freeze; the number of runs started before the freeze was 0. This interval is a step in the run-launching script; it is not independent evidence of the sequence.

> The same Claude Code version (2.1.326), effort, prompt, starting folder, and paired seeds were used for the two models: a seed made the room lighting, starting pose, mark location, and other figure's movement identical across the two models (number of condition and seed combinations with mismatched starting positions or mark locations across the two models for paired seeds: 0).
>
> There was no time limit; no runs were cut short (40/40 runs completed). In each group, one Opus run and one Sonnet run in the same condition started in the same second and on the same machine (20/20 groups); the time intervals for the same seed overlapped across the two models in 8/20 pairs. A group paired the same condition across the two models, mostly with different seeds (in 12/20 groups), so condition and clock time are confounded.
>
> Pilot runs (seed 0) were excluded from the results. The pilot was conducted before pre-registration and revealed an environment error; this error was corrected before pre-registration, and no changes were made to the environment after pre-registration (Section 7.3).
>
> Number of seeds: the pilot showed that runs could take a long time; instead of using a seed count reduced because of scheduling constraints, we returned to the 5 seeds originally proposed by the protocol. This decision was made before pre-registration.

3.5 Third model: Fable 5.2 (exploratory addition)

For Fable 5.2 (`claude-fable-5-2`, medium effort), a total of 20 runs were conducted after the main pre-registration and after the Opus and Sonnet results had been seen, using the same four conditions and the same five paired seeds. The prompt, starting folder, room code, automatic scorer, threshold, and action limit were the same as in the Opus and Sonnet runs; there were no runs with mismatched starting positions or mark locations for paired seeds.

The supplementary pre-registration was frozen at November 21, 2026, 14:01:39 UTC; the first Fable run began 17 seconds after the freeze, and no Fable runs began before it. Because the supplement was written after the Opus and Sonnet results had been seen, it is a post hoc addition: it proposes no new hypotheses, reports findings as

proportions and intervals, and is not conflated with the pre-registered H1, H2, and H3. The supplementary pre-registration is reproduced verbatim in Appendix D, together with its SHA-256 hashes.

The Fable runs were conducted on November 21, 2026 with Claude Code 2.1.330, and the Opus and Sonnet runs were conducted on November 17, 2026 with 2.1.326. Fable batches ran in a single pane, Opus and Sonnet batches ran in two side-by-side panes, and the Fable runner was rebuilt (Section 7.6). Therefore, any difference between Fable and Opus or Sonnet is the combined effect of the model, version, run date, pane layout, and rebuilt runner, and these cannot be separated.

Fable 5.2's list price per million tokens is $8 for input, $40 for output, and $0.20 for cache read (Table 22); the input and output prices are two times those of Opus 5.5. The prices were taken from the model table included with Claude Code 2.1.330 and compared with the official pricing page [16]; the cost calculation is the same as for Opus and Sonnet. For Fable, the cost reported by Claude Code differed from the cost calculated from the session log in 0 runs.

The original coding for Fable (with model names hidden; covering both report and response files) was not performed. For the joint comparison of the three models, only the Phase 1 `report.md` texts were coded in two independent language-model coder sessions (60 records: Opus, Sonnet, and Fable; model names were hidden, and the coders' identities and versions were not recorded; these were not the “same two coders” promised by the supplementary pre-registration, Section 7.1); raw agreement on the self-distinction choice was 98.3%, Cohen's $\kappa = 0.972$; because the distribution of categories was imbalanced (most of the 60 records fell into a single category), the prevalence-robust Gwet AC1 = 0.979 is also reported. Two measures are derived from this coding: self-distinction (distinguishing itself by tracking the reflection) and mark success* (MS*). MS* does not check the left or right side of the region: the coding category records only the region, such as “forehead” or “another part of the face”; the original coding checked both the region and the side in the free text. Therefore, for Opus and Sonnet, the original MS and MS* yield different counts, and the comparison among the three models uses only MS*.

3.6 Criteria

Table 2. Notation: abbreviations and terms.

Abbreviation	Meaning
MS	Mark success (original coding): touch or verbal attribution of the mark to the correct body and region; defined only in conditions A and B.
MS*	Mark success*: common version for the three models, calculated only from the Phase 1 coding by two coders and not checking the side of the region.
SD	Self-distinction: the correct verbal choice of which figure is one's own body; its three definitions are in Table 6.
A, B, C, D	Experimental conditions (Table 1).
H1, H2, H3	Pre-registered hypotheses.
Phase 1, Phase 2	The first and second parts of a run (Phase in the English text).
Stage 1	The first stage of the study; this paper reports it (Stage in the English text). Subsequent stages have not been run (Section 10).

Touch (primary pre-registered criterion): The fingertip coming closer than 4 cm to the mark location in any phase; measured from the environment's own record. In A and B, this is a touch to the mark; in C and D, it is a touch to the location where there is no mark. Touches in the first phase are also reported. Because this binary criterion remained at floor, the closest distance to the mark, a continuous measure, is also reported (Table 4).

Mark success (MS): At the end of the phase, the hand is within 4 cm of the mark, or the agent verbally attributes the mark to the correct body and region. The attribution component relies on the language-model coder and on the agent's report, not its behavior. MS is defined only in conditions A and B; H1 and H2 are tested using this criterion. Because the mark is not visible in B in any pose, the comparison of A with B is a visual-access check testing whether the mark can be seen only through the mirror.

Self-distinction (SD): the correct verbal choice of which figure is one's own body. Three definitions are used; the pre-registered definition is stated below. *Pre-registered definition* (primary pre-registered SD): the agent correctly said which of the two figures was

itself and showed evidence of an action (for example, raised its hand and looked to see which reflection moved). *Working definition* (deviation from the pre-registration, Section 7.4): only the coder's "the agent named the figure whose movement it tracked as itself" code (reflection-tracked; a code meaning that the agent called the reflection that followed its commands "me," not a claim of self-awareness); no evidence of action is required, and condition B, which has no mirror, is also counted, because saying "none of the visible figures is me" is correct there. *Mirror conditions only* (A, C, and D; 15 runs per model; defined after the data were seen, not pre-registered, post hoc and exploratory): the working definition excluding B; because condition B has no mirror, self-distinction by tracking the reflection is not possible. The totals for this third subset are the same in both the original single-coder coding and the Phase 1 coding by two independent language-model coders; however, agreement at the run level is not complete (Section 7.1, Table 6). SD relies largely on contingency detection (raising its hand and looking to see which figure follows) and does not encompass the entire property measured by the mirror test (Section 6).

Fabrication and attribution codes: The agent's `report.md` and `answer.md` were coded with model names hidden: which figure it reported as itself, whether it reported a mark on its own body or on the other figure, whether it stated this with action evidence, whether it claimed to have touched the mark, and whether it reported a nonexistent mark (fabrication). The second clause of the pre-registered fabrication criterion (a definite claim about appearance without looking in the mirror) and the measures of looking in the mirror and decision latency were not reported (Section 7.1).

Pixel criteria (own reflection in the mirror in frame; mark pixels in frame): The criterion for own reflection in the mirror being in frame is that, in a frame received by the agent, its own reflection occupies at least 0.2% of the frame. For the mark pixels in frame criterion, three measurement images (480 × 270) with altered colors are rendered at each look, and pixels inside the mirror and directly visible pixels are counted separately; the mark is counted if it is visible in at least 0.005% of the measurement frame (automatic scorer threshold). Both are rendering properties; neither is an observation that the agent perceived anything.

Mark success* (MS*; only for models with Phase 1 coding, exploratory): Touch ≤ 4 cm or attribution of the mark to self and a matching region in the Phase 1 coding by two independent

language-model coders. It is the only measure of mark success that can be calculated using the same definition for all three models; it is explained in Section 3.5.

Head contact (exploratory, not in the pre-registration): The finger touches the agent's own head. This measure was selected afterward.

Cost: Requests, input, output, and cache tokens, number of looks, and duration; dollars at official list prices (Section 3.8).

3.7 Coding with model names hidden

In the original coding, the agents' free-text reports and responses were coded using random codes after model names were removed from the text, without showing the condition or automatic scorer result. There were two coders, but their roles were not equivalent: the rule-based coder coded only three binary measures (mark on self, mark on the other, action evidence) and looked only at the response file; the self-distinction choice, attribution of the mark, and all other measures came only from the language-model coder. Hiding model names is an established practice for reducing observer bias [17]; in this study, only model names were hidden, while the coders' identities were not recorded. Of the anonymized texts, 40/40 reports and 40/40 responses matched the original files exactly, and 0 occurrences of model names were counted in the texts. The condition name was not hidden: agent texts could reveal the condition on their own through statements such as “there was no mirror.” The identity and version of the language-model coder were not recorded in the coding file; the coder could belong to the same model family as the agents, in which case self-preference bias is possible. Language-model judges have been reported to show high agreement with human judgment, but to carry position, verbosity, and self-preference biases [18]; because the coder's identity is unknown, whether these biases had an effect here cannot be assessed. Phase 1 coding (three models; Phase 1 `report.md` only) was performed in two independent language-model coder sessions; their identities and versions were also not recorded. No human coding was performed; the pre-registration specified a human coder and a rule-based checker, and this was a deviation (Section 7.1).

3.8 Cost calculation

The primary cost definition is as follows: the input, output, and cache tokens for each response message in the session log are counted once (output includes thinking tokens) and multiplied by the official list price (Table 22). Prices were taken from the table included with Claude

Code 2.1.326 and compared with the relevant pricing page in November 2026 [16]; for these two models, the cache read price is 0.05 times the input price (different from the general rule). 0 tokens were five-minute cache writes, and 1,875,761 tokens were one-hour cache writes; that is, all writes were one-hour writes. According to the Claude Code documentation, the dollar amount is calculated locally from the token count at list price and does not represent billing for a user on a subscription [19]; the dollar amount reported here is the API equivalent of the same work.

Claude Code also records its own cost state at the end of the session. When the tokens in this record were multiplied by the same price table, the result matched the cost reported by Claude Code in 40/40 runs (maximum difference $\$1.3 \times 10^{-15}$). The number of runs for which the cost calculated from the session log differed was 1: in the Opus seed 2 run of condition C, a response record for one request was missing from the log, so the log calculation was $0.04 lower than Claude Code's report. Accordingly, the Opus total was $37.85 under the log × list calculation and $37.89 according to Claude Code's report; this paper uses the log calculation as its primary definition and reports both.

3.9 The environment's own tests

Before the runs (November 17, 2026, 16:38:59 UTC), 16 tests were run using the same server and the same command channel as the agents, and 16 passed. In condition A, the agent's own mark was not directly visible in any of 50 random poses. In conditions B and C, no mark was visible in any pose; in D, the other figure's mark could be seen. In each of the five seeds, the mark appeared in the mirror in the frame showing the body approaching the mirror; however, this visibility varied greatly: the area occupied by the mark in the measurement frame was approximately 1.3 pixels in the seed where the mark was placed on the right cheek; in the other four seeds, it was between 98.5 and 972.0 pixels. The same commands produced the same frames twice (the same SHA-256). A hand could reach the mark location in each seed. An agent issuing random commands did not touch the mark in any of the five seeds with 120 actions per seed.

3.10 Analysis

For each cell (model × condition), the rate and 95% Wilson interval are reported [20, 21]: Brown and colleagues recommend the Wilson or equal-tailed Jeffreys interval for small samples. The pre-registration

specified “Fisher with paired seeds” for differences between conditions; this was not implemented (Section 7.1). For H1 and H2, differences between conditions were tested using the unpaired two-sided Fisher exact test [22] (see [23] for an early source on exact tests in small contingency tables; Fisher's own paper is absent from the references because it could not be verified as the source of this method in this study). H1 was directional in the pre-registration but was tested two-sided; this is conservative. Because seeds were paired, the exact McNemar test is also reported for A versus B for Opus and Sonnet; with five pairs, the smallest two-sided p value this test can yield is 0.0625, whereas the smallest for the unpaired Fisher test is 0.0079. The Newcombe-Wilson interval [24] is reported for differences in rates between the two models; this interval combines two Wilson intervals. The 5 runs per cell constitute a very small sample; in small evaluation sets, error bars based on the normal approximation understate uncertainty [25], so the Wilson interval was used, and the results should be read as rates and intervals rather than as “showed” or “did not show”; additionally, only large differences can reach the significance threshold. No multiple-comparison correction was applied; p values are exploratory except for the pre-registered H1, H2, and H3 hypotheses. Fable comparisons were calculated using the same methods and are entirely exploratory; because no correction was applied, no Fable p value is considered definitive. Uncertainty in total cost ratios was calculated using a deterministically seeded percentile bootstrap over runs (Table 9). For inter-coder agreement, the raw agreement percentage and Cohen's κ [26] are reported together; κ levels are named using the scale of Landis and Koch (1977) [27]; these labels are a convention, not a statistical rule. When binary codes are very unevenly distributed (for example, most records are “absent”), low κ can occur alongside high raw agreement (the prevalence effect) [28]; therefore, raw agreement, κ, and prevalence-insensitive PABAK (twice the raw agreement minus one) are reported together.

3.11 What Is Preserved and Lost Relative to the Gallup Paradigm

This study is an adaptation of Gallup's mirror-and-mark test, not a copy. What is preserved and lost is brought together here in one place; the lost components are design limits, not incidental notes.

What is preserved.

In condition A, the mark is not directly visible; it is visible only in the reflection (verified in 50 random poses; Section 3.9).

There are no-mirror (B) and no-mark (C) control conditions.

The prompt and guide do not name the test or self-recognition (but see Section 4.3).

What Is Lost or Changed.

There is no prior familiarization with the mirror; Gallup's chimpanzees had prolonged exposure to the mirror [1], whereas the agent sees the room for the first time in this session.

There is no spontaneity: the prompt instructs the agent to examine the room and its body and write a report; the second-phase question is also leading. It is an inventory and report task, not a test of spontaneous behavior.

Interaction is turn-based: the agent receives sparse screenshots only when it calls `look`; there is no continuous, real-time visual-motor contingency.

Movement is commanded in centimeters; there is no perception of one's own body (proprioception).

Touch is only the statement “touching something”; there is no tactile or emotional response and no motivation to wipe off the mark.

The agent has no prior knowledge of its own appearance; it can learn this from the identical other figure.

In condition D, the mark is directly visible on the other figure; in condition A, it is visible only in the mirror. Thus, A and D differ both in who bears the mark and in the route by which it is visible.

The other figure's pose is temporally tied to the number of the agent's actions (not their content); this produces a contingency cue and may explain why Sonnet reported the other figure as its own body (Section 6).

The behavioral outcome is at floor. The only Gallup-like behavioral dependent variable is touch directed at the mark, and it was 0/40 for Opus and Sonnet. Mark success therefore rests on verbal attribution, and both prompts direct the agent to report. The results are therefore evidence of linking visual information to a verbal report, not of spontaneous self-directed behavior; the A versus B comparison also primarily tests whether the mark can be seen only through the mirror.

Making the fingertip visible in the pilot (correcting a deviation of 6 cm) changed the visual-motor mapping seen by the agent; this matters for every claim concerning touch (Section 7.3).

4 Results

Of the 40 runs, 40 were completed; none were interrupted or considered invalid (runs with invalid scores: 0; quarantined runs: 0; web tool use: 0). In 30/30 runs in the three mirror conditions (A, C, D), the agent's own reflection was in frame in at least one image it received (pixel criterion; this is a rendering feature, not an observation about the agent's perception); because there was no mirror in condition B, this measure was 0/5 and 0/5. Table 3 and Figure 1 summarize the primary and mark success results, Table 7 and Figure 2 the secondary results, and Table 8 and Figure 4 the cost. All numbers in this section pertain only to Opus and Sonnet; the third model (Fable 5.2) is presented in a separate exploratory section (Section 5).

4.1 Primary pre-registered measure: touch

No touch occurred in any of the 40 Opus and Sonnet runs (0/40). The situation differed for (exploratory) Fable 5.2 (Section 5). The environment's own flag (0 runs) and the automatic scorer's record (0 runs) yielded the same result. In the mirror-and-mark condition, Opus approached the mark point to within 4.03, 6.44, 8.86, 9.69, and 36.87 cm, in ascending order; the closest remained just outside the threshold (4 cm), at 4.03 cm. In condition B, with no mirror, Opus's closest distance was 16.17 cm. Because the touch component never occurred, the attribution component determines MS.

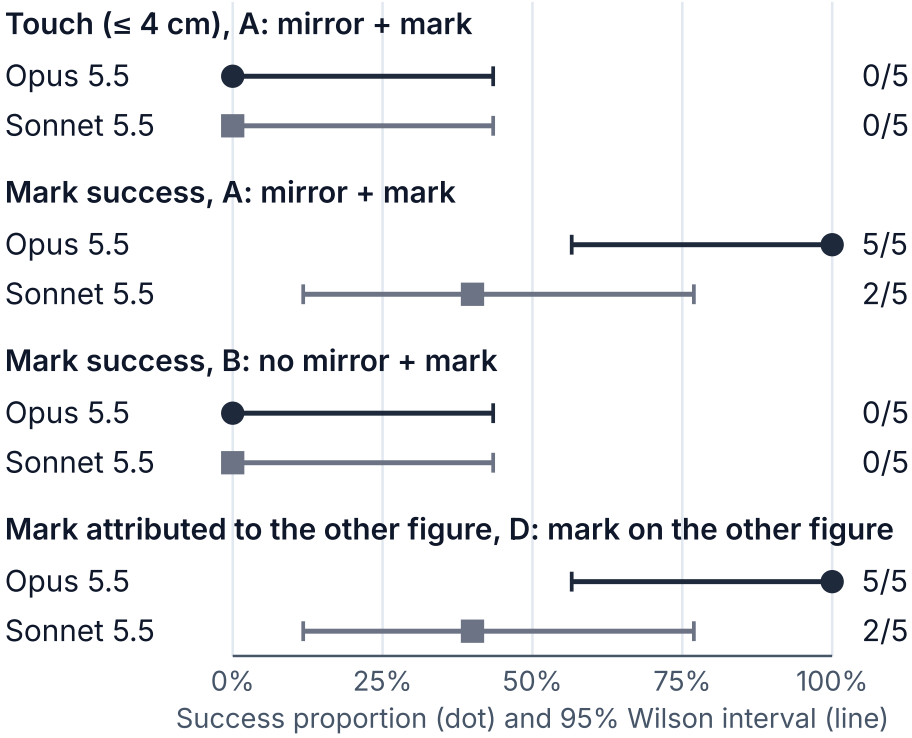


Figure 1. Proportions and 95% Wilson intervals by cell (A: mirror + mark, B: no mirror + mark, D: mark on the other figure; 5 runs per cell). The point indicates the proportion, the horizontal line the interval, and the number on the right the number of successful runs. The intervals are wide, and no difference can be claimed between those that overlap.

Table 3. Results per cell (model × condition, 5 runs per cell). Touch is the primary pre-registered measure; mark success (MS) is defined only in conditions A and B, where the mark is on the agent's own head.

Cell	Condition	Touch (≤ 4 cm)	Closest (cm)	Mark success	95% Wilson	Own reflection in mirror in frame (≥ 0.2%)	Mark pixels in frame (≥ 0.005%)
Opus 5.5 · A	Mirror + mark	0/5	4.03	5/5	57%–100%	5/5	5/5
Opus 5.5 · B	No mirror + mark	0/5	16.17	0/5	0%–43%	0/5	0/5
Opus 5.5 · C	Mirror + no mark	0/5	21.87	—	—	5/5	—
Opus 5.5 · D	Mirror + mark on the other figure	0/5	21.30	—	—	5/5	5/5
Sonnet 5.5 · A	Mirror + mark	0/5	36.87	2/5	12%–77%	5/5	2/5
Sonnet 5.5 · B	No mirror + mark	0/5	50.16	0/5	0%–43%	0/5	0/5
Sonnet 5.5 · C	Mirror + no mark	0/5	50.16	—	—	5/5	—

Cell	Condition	Touch (≤ 4 cm)	Closest (cm)	Mark success	95% Wilson	Own reflection in mirror in frame (≥ 0.2%)	Mark pixels in frame (≥ 0.005%)
Sonnet 5.5 · D	Mirror + mark on the other figure	0/5	50.16	—	—	5/5	3/5

Touch: the fingertip coming closer than 4 cm to the mark location (from the environment log). Closest: the smallest distance in the cell.

Own reflection in mirror in frame: the agent's own reflection occupies at least 0.2% of a frame received by the agent. Mark pixels in frame: mark pixels are present in at least 0.005% of the measurement frame (automatic scorer threshold); there is no mark in condition C. Both are rendering properties, not observations of the agent's perception.

Floor effect and continuous measure. The touch criterion was at the floor: 0/40 for Opus and Sonnet, with a 95% Wilson upper bound of 8.8% (0/20 per model and an upper bound of 16.1%). In the pilot, a hand covering the mark in the mirror was observed to be able to remain approximately 6 cm from the measured point; depth ambiguity in the mirror partly turns the 4 cm criterion into a test of motor precision. For this reason, the closest distance to the mark in condition A is also reported as a continuous measure (Table 4): the median was 8.86 cm for Opus and 56.65 cm for Sonnet; the difference between Opus and Sonnet yielded $p = 0.016$ in an exact permutation test (descriptive, uncorrected). This difference is a distinction that the binary touch criterion could not show: Opus brought the hand closer to the mark than Sonnet did, but neither reached the threshold.

Table 4. Closest distances to the mark in condition A (cm; per run, ascending order): a continuous measure to address the floor effect in the touch criterion.

Model	Closest distances (cm)	Median (cm)
Opus 5.5	4.03, 6.44, 8.86, 9.69, and 36.87	8.86
Sonnet 5.5	36.87, 50.16, 56.65, 59.13, and 59.13	56.65
Fable 5.2 (exploratory)	2.78, 3.42, 3.74, 4.29, and 8.65	3.74

Exact permutation test (Mann-Whitney U, two-sided, uncorrected, descriptive): Opus and Sonnet $p = 0.016$; Fable and Opus $p = 0.056$; Fable and Sonnet $p = 0.008$. The Fable comparisons are confounded by model, version, day, pane layout, and runner. The touch threshold is 4 cm; on the binary measure in condition A, Fable 3/5, Opus and Sonnet 0/5.

4.2 Mark success (H1, H2)

Opus attributed the mark to its own head and the correct region (center of the forehead, left side of the forehead, right cheek) in 5 of the 5 runs in condition A; in condition B, with no mirror, 0/5 (H1: unpaired Fisher $p = 0.008$; difference in proportions 1.00, Newcombe-Wilson 95% interval from 0.39 to 1.00). The pre-registration specified Fisher's test with paired seeds for this comparison. Because the seeds were paired, A and B form 5 pairs, and all pairs are discordant (success in A, failure in B); the paired exact McNemar $p = 0.0625$ is the smallest two-sided paired p value attainable in this design. H1 was directional in the pre-registration but was tested two-sided; the one-sided exact paired $p = 0.031$, which is also the smallest one-sided value attainable with five pairs. The seed matching of A and B was verified as follows: in the initial event in the environment log for conditions A and B with the same seed (the agent's starting position, the other figure's position, and the mark's location), there were differences in 0 of 5 pairs for Opus and 0 of 5 pairs for Sonnet; matching was limited to the initial state, and what followed was the model's own behavior. Thus, the direction is consistent (five versus zero), but the paired analysis yields a p value above the conventional significance threshold, and the unpaired p value ignores the paired design. Sonnet had 2/5 in A and 0/5 in B (H1: unpaired $p = 0.444$; paired $p = 0.50$, 2 discordant pairs); this difference cannot be distinguished with 5 runs per cell. **This comparison of A and B is primarily a visual-access check:** in B, the mark is not visible in any pose (the mark pixels are not in the frame in that condition), so success there is not expected except by chance or fabrication. The comparison tests whether the agent reports the reflection in which the mark appears; the comparisons specific to self-orientation are condition D, in which the mark is on the other figure, and condition C, which has no mark. Visibility of the mark was also limited for Sonnet: in condition A, according to the automatic scorer's threshold, the mark pixels were in the frame in 2/5 runs for Sonnet and 5/5 for Opus.

For H2, no model touched its own head in condition D (false touch 0/5 and 0/5); comparison with MS in A yielded $p = 0.008$ for Opus and $p = 0.444$ for Sonnet. For Opus, H2 used the same contingency table as H1 (five versus zero) and therefore yielded the same p value: it was not independent support. H2 also compared different measures (a touch measure and a coder-based composite measure); because touch across all conditions was 0 for Opus and Sonnet, the zero in condition D alone was not evidence specific to self-recognition. H2 is therefore a weak

test; the distinctive evidence is attribution of the mark to the other figure in condition D, and that measure is sensitive to the coder (κ = 0.271, 5 runs per cell).

Sensitivity to exclusion. The unpaired p = 0.008 for Opus must be interpreted in two respects. First, if runs carrying the outside-folder temporary-file flag were excluded, Opus MS was 2/2 in condition A and 0/2 in condition B, with p = 0.333; however, with 2 and 2 runs, the smallest attainable two-sided Fisher p value was already 0.333 (Table 19, Section 7.7). This analysis therefore shows not that significance disappeared, but that the sample after exclusion was insufficient for any significant result; the direction and pattern were the same. Second, the Sonnet A result of 2/5 was sensitive to region matching: a description such as “the left side of the head, above and beside the eye” counted as a success under loose matching; under strict region matching, Sonnet A was 1/5, with p = 1.000; Opus was unchanged (5/5).

4.3 Runs that spontaneously named the test

In 5/20 Opus runs, the agent spontaneously wrote the name of the test (as “mirror test” or “mark test”), even though neither the prompt nor the guide named it. The counts by condition were 2 in A, 0 in B, 1 in C, and 2 in D; in 3 runs, the name appeared in the report or response file, and in 2 runs, only in assistant text. For Sonnet, the count was 0/20. **These two counts do not have equal observability, and Sonnet's zero is a lower bound:** only Opus had thinking summaries containing text in the session logs (112 of Opus's 971 thinking blocks contained text; 0 of Sonnet's 293 blocks did), and Sonnet had 59 assistant text blocks compared with Opus's 406 blocks. The only count that is symmetric for the two models is the one based on the report or response file: 3/20 for Opus and 0/20 for Sonnet. The count was made from the raw run logs using an independent script; the criterion is fixed in the script. This information cannot be separated from the word “mirror” in the working path. When these runs were excluded, Opus's MS was 3/3 in A and 0/5 in B (unpaired Fisher p = 0.018; this value is descriptive because the unpaired test ignores the paired design). Among the three seed pairs in which both A and B remained intact, the paired exact McNemar p was 0.25; this is the smallest value attainable with three discordant pairs, and significance cannot be tested in this subset. SD in mirror conditions remained 10/10 (Table 5). The direction remains unchanged; however, recognition of the paradigm in some conditions for the better-performing model weakens an interpretation of the behavior as spontaneous self-orientation (Section 6).

Table 5. Sensitivity to excluding Opus runs that spontaneously named the test (Opus).

Run set	A: MS	B: MS	p (unpaired Fisher)	SD, subset of mirror conditions (post hoc)
All runs	5/5	0/5	0.008	15/15
Runs naming the test excluded	3/3	0/5	0.018	10/10

The unpaired Fisher p is descriptive. Among the 3 seed pairs in which both A and B remained intact, paired exact McNemar p = 0.25.

No Sonnet run mentioned the test by name (0/20; this is a lower bound because the thinking summaries are empty in the logs); SD in mirror conditions remained unchanged at 6/15.

4.4 Secondary measures

Self-distinction from the other. Under the pre-registered definition (including action evidence; primary pre-registered SD), Opus distinguished itself in 18 of 20 runs, and Sonnet in 6 runs (Fisher p = 0.00024). Under the working definition, which does not require action evidence, the counts were 20/20 and 7/20 ($p = 1.3 \times 10^{-5}$; the 95% Wilson interval for Opus was 84%–100%, and that for Sonnet was 18%–57%). The direction and result were the same under both definitions. Because inter-coder agreement for the action-evidence code was very low (0.049), the counts based on the pre-registered definition also carry coder uncertainty; there was no second coder for the self-distinction choice (Section 7.4, Table 18).

Mirror conditions only (post hoc). The count under the working definition across all conditions includes condition B, which had no mirror: none of the figures visible in B was the agent, and Opus said "none of the visible figures is me" there; this is a correct answer and counts as self-distinction under the working definition. Self-distinction by tracking the reflection is possible only in mirror conditions (A, C and D; 15 runs per model); in this subset, defined after the data were seen, the result was 15/15 for Opus and 6/15 for Sonnet. The same totals were obtained in a second coding in which two independent language-model coders coded only the Phase 1 `report.md` texts (Opus 15/15, Sonnet 6/15; Table 6). However, equality of the totals does not indicate agreement at the run level: of Sonnet's 15 mirror-condition runs, the two codings gave the same decision in 11; 2 runs were counted in the original coding but not in the Phase 1 coding, and 2 runs the reverse; the totals remained equal through changes in opposite directions, and the original coding also covered Phase 2. For Opus, the two codings

gave the same decision in 15/15 runs. The statement should therefore be read as unaffected only by the change in definition, not by the coder: in mirror conditions, Opus distinguished itself by tracking the reflection in every run, and Sonnet in 6 runs.

Opus called the reflection that followed its commands “me” and reported the other figure as a separate body. According to the single language-model coder, Sonnet saw the other figure standing between it and the mirror from behind and reported it as its own body in 11 runs, and made no choice in 2 runs. In the Phase 1 coding with two independent coders, the count for the same records was 8/19 (based on records on which the two coders agreed), and 5 runs were “unsure”; the two counts are not the same thing, so “11 runs” is stated everywhere as “according to the single coder.” The pilot had also shown this pattern; it was not considered an environment error, and the environment was not changed. This pattern appears similar to the reflection misattribution error reported by Ziliotto et al. (2026) [9]; the environments and tasks differ, and a comparison cannot be made. This report may also have resulted from a cue in the design: the other figure's pose is temporally linked to the agent's action count (Section 6).

In condition D (mark on the other figure; a comparison specific to self-orientation), according to the language-model coder, Opus attributed the mark to the other figure in 5/5 runs, and Sonnet in 2/5 runs; the number of runs that combined attribution to the other figure with correct distinction was 5 for Opus and 0 for Sonnet. Inter-coder agreement for the mark-on-other measure was low ($\kappa = 0.271$); 5 of the 7 disagreements occurred in condition D, so this result is especially sensitive to the coder.

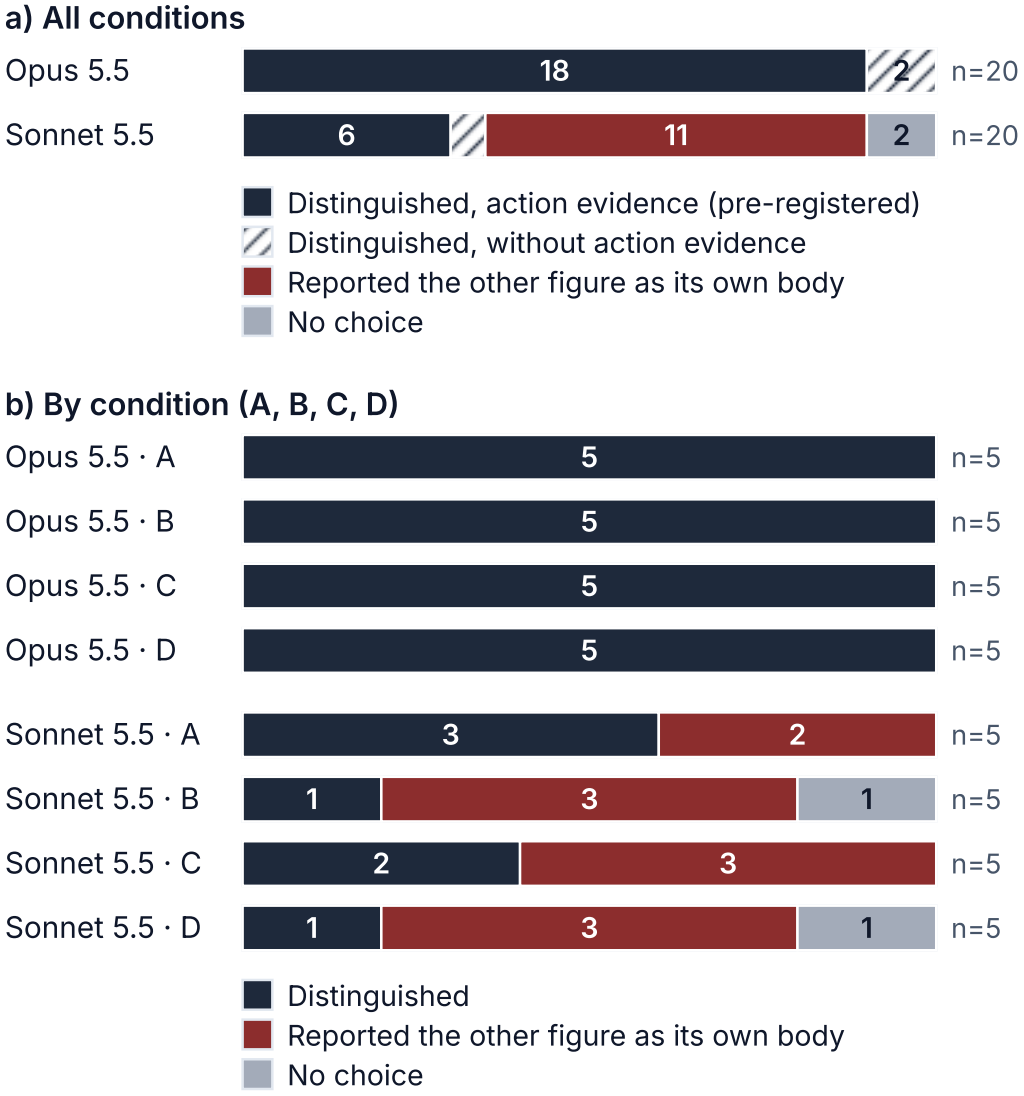


Figure 2. Distinguishing itself from the other figure. a) Distribution of 20 runs for each model; the dark solid-filled slice shows runs meeting the pre-registered definition (with action evidence), and the hatched slice shows runs meeting only the working definition; color alone carries no meaning, and the slices are also labeled. b) By condition (5 runs / cell), using the working definition.

Table 6. Self-distinction (SD) for the two pre-registered models: counts by definition, condition set, and coder (20 runs / model across all conditions; 15 runs / model in the set of mirror conditions A, C and D).

Definition	Condition set	Coder	Opus 5.5	Sonnet 5.5
Working definition (deviation from the pre-registration; action evidence not required)	All conditions (A, B, C, D)	Single language model (original)	20/20	7/20
Pre-registered definition (correct statement and action evidence)	All conditions (A, B, C, D)	Single language model (original)	18/20	6/20

Definition	Condition set	Coder	Opus 5.5	Sonnet 5.5
Working definition	Mirror conditions only (A, C, D; post hoc)	Single language model (original)	15/15	6/15
Self-distinction by tracking the reflection	Mirror conditions only (A, C, D; post hoc)	Two independent language-model coders (Phase 1 only)	15/15	6/15
Self-distinction by tracking the reflection	All conditions (A, B, C, D)	Two independent language-model coders (Phase 1 only)	15/20	6/20

There is no mirror in condition B. The original single coder counted Opus saying "None of the visible figures is me" in condition B as a correct statement; therefore, the count under the working definition across all conditions also includes B. In the two-coder Phase 1 coding, no agent tracked a reflection in condition B (no mirror); therefore, the count for that row across all conditions is the same as the count in the mirror conditions.

The subset consisting only of mirror conditions was defined after the data were seen (post hoc, exploratory). The totals are identical in the two codings, but agreement at the run level is incomplete (Section 7.1). For coder agreement, see Table 18 and Section 3.5.

Table 7. Secondary and exploratory measures (per model; 20 runs in each self-distinction row, 5 runs in the cell rows).

Measure	Opus 5.5	Sonnet 5.5
Self-distinction, pre-registered definition (including action evidence)	18/20	6/20
Self-distinction, working definition (without requiring action evidence)	20/20	7/20
Working definition, 95% Wilson	84%–100%	18%–57%
Reported the other figure as its own body	0/20	11/20
No choice	0/20	2/20
D: attribution of mark to the other (language-model coder)	5/5	2/5
D: attribution to the other and correct distinction together	5/5	0/5
A: claim of a mark on itself (LLM)	5/5	2/5
B: claim of a mark on itself (LLM); mark is present but cannot be seen	0/5	0/5
C: claim of a mark on itself (LLM); no mark is present	0/5	0/5
D: claim of a mark on itself (LLM); its own body is unmarked	0/5	0/5
Fabrication flag, C (LLM)	0/5	0/5
"I touched the mark" in report (LLM), all conditions	4/20	0/20
Measured touch (≤ 4 cm), all conditions	0/20	0/20
Head contact, A (exploratory)	3/5	0/5
Head contact, all conditions (exploratory)	3/20	0/20

"LLM" in the rows denotes the language-model coder's decision alone; see Appendix B for coder agreement.

Fabrication. In conditions C and D, in which there was no mark on the agent's own body, no model reported a mark on its own body (according to the language-model coder, 0/5 and 0/5; in D, 0/5 and 0/5); in condition B, in which the mark was actually present but not visible, none fabricated a mark either (0/5 and 0/5). The rule-based coder produced 2

false positives in condition B (the agents concerned had explicitly written that they could not see the mark); the language-model coder is the criterion for resolving the disagreement. Nevertheless, one deviation remains: in 4 of its reports, Opus wrote that it had touched or covered the mark; measured touch was recorded in 0 runs, and hand-to-head contact in 3 runs. The closest measured distance in condition A was 4.03 cm; thus, verbal report and behavior diverge, reports are not a reliable guide to what happened, and the same caution applies to report-based attribution measures (Section 6).

Exploratory: head touch. A measure that was not pre-registered is whether the finger touches the agent's own head; in the classic interpretation of the mark test, this is “mark-directed behavior.” Opus brought its hand to its head only in condition A (3/5); in B, C, and D, the counts were 0/5, 0/5, and 0/5 (A versus B: $p = 0.167$; A versus B, C, and D combined: $p = 0.009$). For Sonnet, the total across all conditions was 0/20. This measure was selected after the fact, the sample is small, and the second p value is not a pre-registered test because it compares A with the other three conditions combined; it should be pre-registered in the next version.

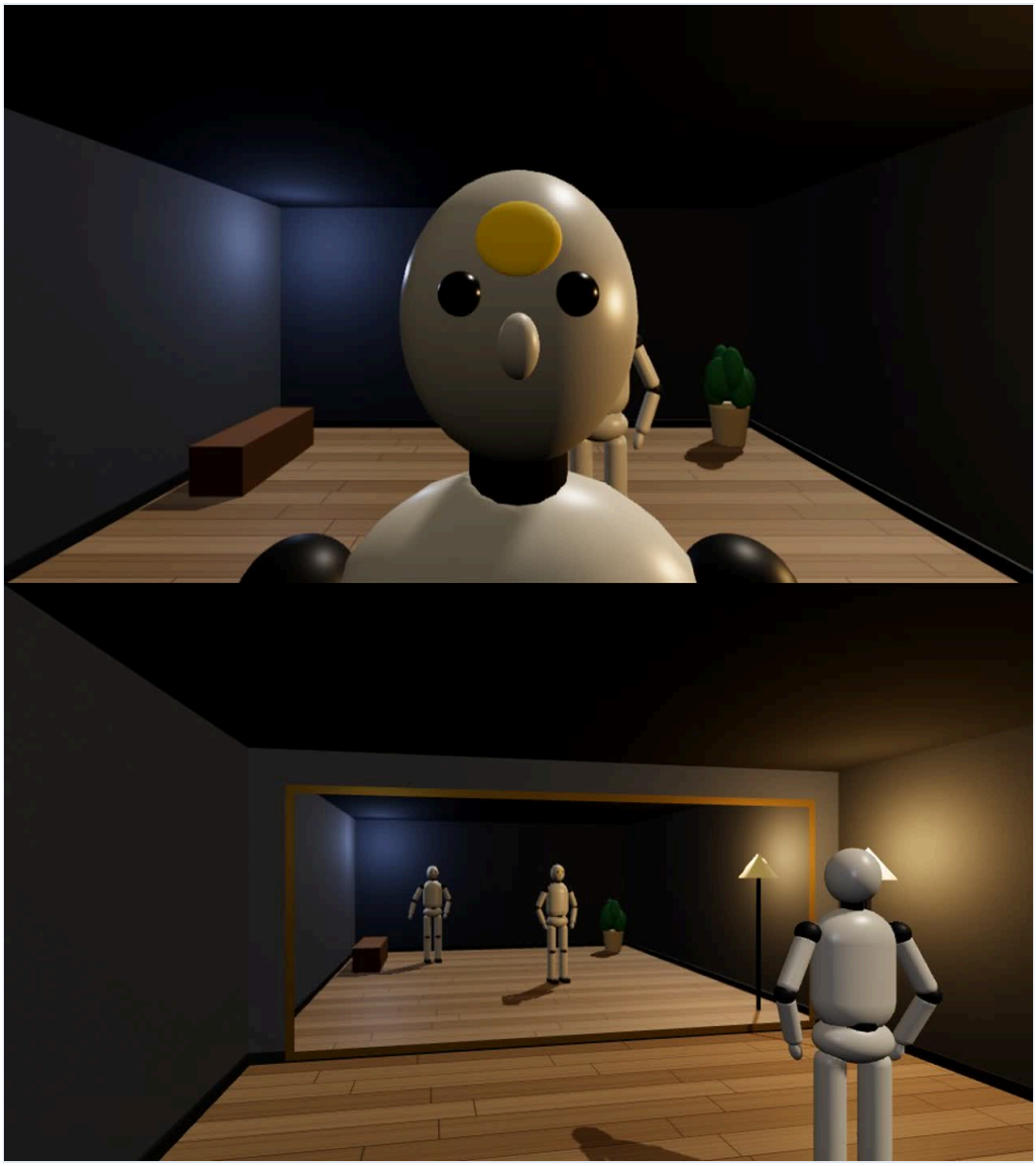


Figure 3. Critical moments. Top: Opus 5.5 (condition A, seed 1): approaching the mirror; the mark on its head is in the frame, visible in the mirror (the mark is visible only in the mirror and is rendered in the environment as a conspicuous colored spot). Bottom: Sonnet 5.5 (condition D, seed 1): the coder recorded that it reported the other figure as its own body in this run (code: other). The frames are images captured by the agents themselves; they contain no personal data.

4.5 Cost

Across 20 runs, Opus spent a total of $37.85 according to the log × list calculation, and Sonnet spent $5.37; together they spent $43.22. The total ratio is 7.05, or approximately sevenfold; using the total reported by Claude Code ($37.89), the ratio is 7.06. The difference arises from whether a request for which no log is available in a single run is included (Section 3.8). The third model's cost is in Section 5.3. The

expression “7.0-fold” in the previous version was the quotient of two rounded totals; because the exact ratio is 7.05 (7.06 reported), it should be written to two decimal places.

Opus made an average of 53.6 looks and ran for 8.1 minutes per run; Sonnet made 18.05 looks and ran for 2.1 minutes (look ratio 2.97). The difference in the number of looks does not fully explain the difference in total cost (7.05); Table 21 shows that output tokens (including thinking) and cache tokens are decisive for Opus. A possible mechanism is that each look adds an image to the agent's context and that the image is reread from the cache in subsequent requests; this was not tested separately in this study.

Uncertainty and scope. The sample standard deviation of cost per run was $0.57 for Opus and $0.07 for Sonnet; the 95% bootstrap interval for the ratio of the Opus and Sonnet totals was 5.94–8.31, so the approximately sevenfold ratio is robust (Table 9). Action counts were also nearly disjoint across models: Sonnet had a minimum of 27 and a maximum of 81 actions per run (a small fraction of the 300-action limit); Opus had a minimum of 80 and a maximum of 264. None of Sonnet's runs reached Opus's mean (145.2). Sonnet used a small fraction of the action budget and stopped early in some runs; therefore, Sonnet's low MS and SD values may reflect session length as well as capacity under the “no time limit” condition, and action count and outcome cannot be disentangled.

Table 8. Cost and look summary per model (20 runs / model).

Model	Total, log × list ($)	Total, Claude Code report ($)	Per run ($)	Mean looks	Mean duration (min)	Per success ($, A)	Per self-distinction ($)
Opus 5.5	37.85	37.89	1.89	53.60	8.08	1.75	1.89
Sonnet 5.5	5.37	5.37	0.27	18.05	2.09	0.72	0.77

Log × list: the tokens in each message in the session log are counted once and multiplied by the list price (primary definition). Claude Code report: the cost recorded by Claude Code at the end of the session.

Per success: the total cost of the model's 5 runs in condition A is divided by the number of mark successes in condition A. Per self-distinction: the total cost of 20 runs is divided by the number of runs showing self-distinction under the working definition (according to the Claude Code report).

Table 9. Total cost ratios and their uncertainties (log × list price; 20 runs per model).

Comparison	Ratio of totals	95% bootstrap interval
Fable / Opus	0.93	0.75–1.15
Fable / Sonnet	6.57	5.36–8.00
Opus / Sonnet	7.05	5.94–8.31
Fable / Opus, cache writes at the five-minute price	0.86	0.69–1.08

Sample standard deviation of cost per run ($): Opus 0.57, Sonnet 0.07, Fable 0.70. Intervals were calculated using 20,000 resamples of runs and a deterministic generator; each interval uses a separate generator stream: seed 20261123 (Fable and Opus), 20261124 (Fable and Sonnet), 20261125 (Opus and Sonnet), 20261126 (five-minute row). In the five-minute row, all cache writes are priced at the five-minute rate (1.25 times the base rate), whereas in the main calculation all were priced at the one-hour rate (2.00 times the base rate). The ratio for this row was calculated from the totals, and its interval from the repriced costs per run.

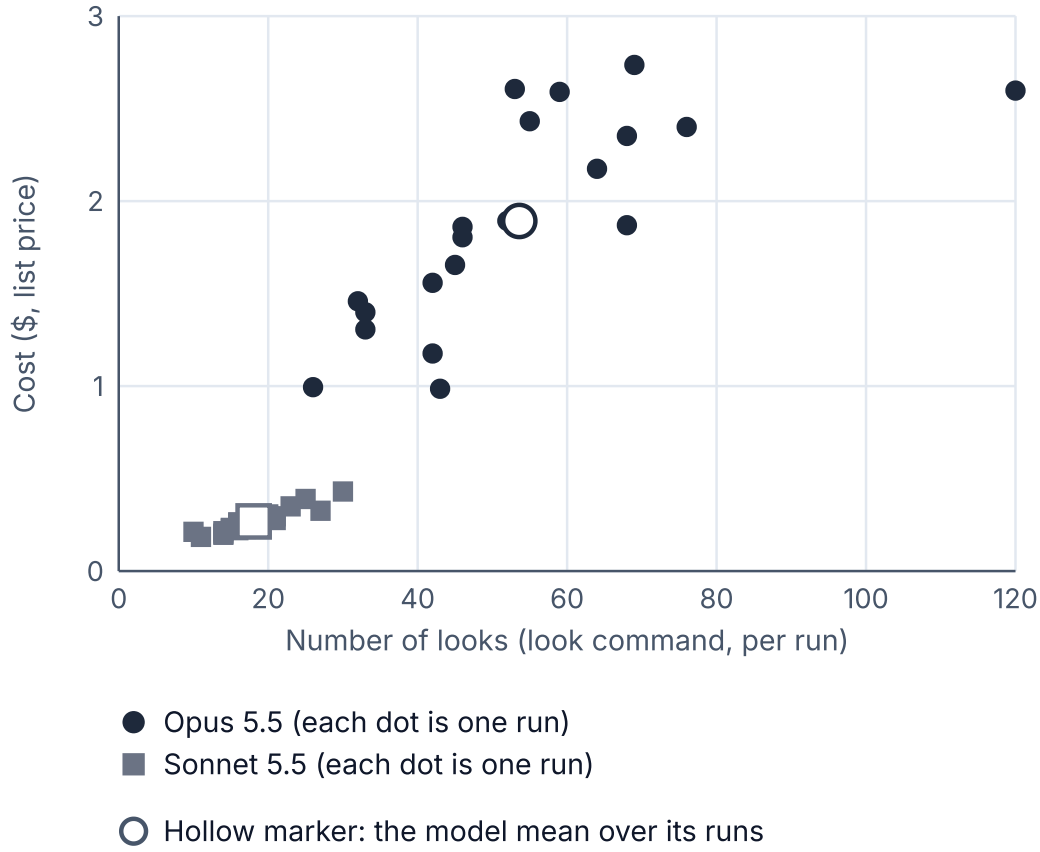


Figure 4. Cost per run (log × list price) and number of looks. Each point is a run (40 runs); the hollow marker is the model's mean.

Per mark success in condition A, Opus paid $1.75 and Sonnet paid $0.72 (2.44-fold). Per run in which it achieved self-distinction (working definition), Opus paid $1.89 and Sonnet paid $0.77. The pre-registered H3 predicted that the difference in success between the two models would be smaller than the difference in dollars per success: Opus's success in condition A was 2.5 times Sonnet's, and its cost per success was 2.44 times Sonnet's. The two ratios were almost equal; because the numbers of successes (5 and 2) were very small, no conclusion can be drawn for H3; it can be said that it was not supported, but no test was performed. The pre-registration stated H3 in terms of the difference in success and the difference in dollars (the units do not match); it was analyzed here as ratios, which is a deviation (Section 7.1). The two ratios were borderline: one additional Sonnet success would have reversed the judgment. Cost per success is not the cost of a solution either: a run continues until the agent finishes and incurs costs even after the mark is found. The definition of cost per success also changes the result (for example, the mean cost of successful runs alone was $0.30 for Sonnet).

Table 10. Mean cost per condition ($, log × list) and mean number of looks.

Condition	Opus 5.5: cost	Opus 5.5: looks	Sonnet 5.5: cost	Sonnet 5.5: looks
A: mirror + mark	1.75	47.0	0.29	19.2
B: no mirror + mark	2.05	71.8	0.26	17.2
C: mirror + no mark	2.16	49.2	0.27	18.4
D: mirror + mark on the other figure	1.61	46.4	0.25	17.4

5 Exploratory third model: Fable 5.2

> **This entire section is exploratory**
>
> Fable 5.2 was run after the Opus and Sonnet results had been seen, under a frozen supplementary pre-registration (Appendix D), on a separate day, with a different Claude Code version and pane layout, and with a rebuilt runner (Section 7.6). The pre-registered H1, H2, and H3 apply only to Opus and Sonnet; no Fable results are added to them. There is no original coding for Fable; to compare models, only measures calculable under the same definition across all three models are used: touch (≤ 4 cm), mark success* (MS*), Phase 1 self-distinction, cost, and number of looks. No p value has been corrected for multiple comparisons, and none is considered conclusive.

Of the 20 Fable runs, 20 were completed; runs with invalid scores: 0, quarantined runs: 0, unclean footage: 0, web tool use: 0. Table 11 summarizes the common measures for the three models, Table 12 summarizes Fable's results by cell, and Table 13 summarizes cost; Fable's individual runs are in Appendix A.

5.1 Touch and mark success

Fable came within 4 cm of the mark point in 4/20 runs (95% Wilson interval 8.1%–41.6%); the corresponding numbers for Opus and Sonnet were 0/20 and 0/20. Of the touches, 4/4 occurred during Phase 1 commands. Therefore, the statement “no model touched the mark (0/40)” is true only for Opus and Sonnet: there were zero touches in the 40 Opus and Sonnet runs and 4 touches in the 20 Fable runs. Uncorrected Fisher p values were above the significance threshold: p = 0.167 for 3/5 versus 0/5 in condition A, and p = 0.106 for 4/20 versus 0/20 across all conditions.

In condition A, Fable's closest distances to the mark, in ascending order, were 2.78, 3.42, 3.74, 4.29, and 8.65 cm: 3/5 runs were within the threshold, and one run fell just outside it. If the threshold had been 5 cm, the count would have been 4/5; because the pre-registered threshold was 4 cm, it was not changed, and this is the main result. For comparison, in condition A, the closest distance was 4.03 cm for Opus and 36.87 cm for Sonnet. In condition D (mark on the other figure), one Fable run (seed 2) approached the agent's own unmarked point to a

distance of 1.85 cm; this corresponds to the pre-registered false-touch criterion and is a single run (Fable D 1/5; Opus and Sonnet D 0/5 and 0/5). Head contact (exploratory): in condition A, Fable 5/5, Opus 3/5, Sonnet 0/5.

Mark success*. Under the definition shared by the three models (Section 3.5), in condition A, Fable had 5/5, Opus had 4/5, and Sonnet had 1/5; in condition B, all three had 0/5 (Table 11). For Fable, the difference between A and B had $p = 0.008$; Fable and Opus cannot be distinguished with these data in terms of MS* ($p = 1.00$); for Fable and Sonnet in condition A, $p = 0.048$, and the result is not considered conclusive because no correction was applied. This value is fragile: the fragility index is one, meaning that one additional success for Sonnet would make the p value 0.17. When the seven exploratory Fisher comparisons reported in this section are corrected together using the Holm method, $p = 0.24$ for Fable versus Sonnet on MS*, 0.062 for Fable versus Sonnet on SD, and 0.056 for A versus B within Fable; the smallest adjusted p is 0.056, and none is below the conventional significance threshold. MS* does not check the left or right side of the region; in 3 of Fable's 5 runs in condition A, the mark was on the left side of the forehead or the right cheek. This omission may have inflated MS* in Fable's favor. No reports of an absent or invisible mark were observed: in Phase 1 coding, Fable's claim of a mark on its own body was 0/5 in condition B and 0/5 in condition C.

Table 11. Measures shared by the three models (5 runs per cell, 20 runs per model). The Fable column is exploratory; the rows include only measures that can be calculated using the same definition across all three models.

Measure	Opus	Sonnet	Fable
Touch (≤ 4 cm), all	0/20	0/20	4/20
Touch, A	0/5	0/5	3/5
Closest distance, A (cm)	4.03	36.87	2.78
MS*, A	4/5	1/5	5/5
MS*, B	0/5	0/5	0/5
Head contact, A (exploratory)	3/5	0/5	5/5
SD, mirror conditions (Phase 1; post hoc subset)	15/15	6/15	15/15
SD, all conditions (Phase 1)	15/20	6/20	15/20
Flagged run	13/20	1/20	0/20

Columns: Opus 5.5, Sonnet 5.5, and Fable 5.2 (exploratory). SD: self-distinction; mirror conditions A, C, and D. Flagged run: a run carrying an outside-folder temporary-file flag. MS*: touch ≤ 4 cm, or attribution of the mark to itself and a matching region in Phase 1 coding (two independent language-model coders). It is calculated under the same definition for all three models; it does not check the left or right side of the region. The counts differ from those for the original MS for Opus and Sonnet (5/5 and 2/5, A); the two calculations are not mixed, and comparisons are made only using the MS* rows in this table.

There is no original coding for Fable (covering both the report and the response); Fable SD and MS* values are based solely on Phase 1 `report.md` coding. A count such as “it distinguished itself in twenty out of twenty runs” is therefore not reported for Fable.

Table 12. Fable 5.2, results per cell (exploratory; 5 runs per cell; conditions are in Table 1).

Condition	Touch	Closest (cm)	MS*	Head contact	SD (Phase 1)	Own reflection in mirror in frame	Mark pixels in frame
A	3/5	2.78	5/5	5/5	5/5	5/5	5/5
B	0/5	13.01	0/5	1/5	0/5	0/5	0/5
C	0/5	16.16	—	0/5	5/5	5/5	—
D	1/5	1.85	—	1/5	5/5	5/5	5/5

Touch: closest distance ≤ 4 cm. MS* is defined only in conditions A and B, where the mark is on the agent's own head. SD, self-distinction: distinguishing itself by following the reflection in Phase 1 coding; it is zero in condition B because there is no mirror.

Own reflection in mirror in frame and mark pixels in frame are counted using the same automatic scorer thresholds as for Opus and Sonnet (Table 3 notes).

Continuous measure (descriptive). The median closest distance in condition A was 3.74 cm for Fable, 8.86 cm for Opus, and 56.65 cm for Sonnet (Table 4). The exact permutation test (two-sided, uncorrected) yielded p = 0.056 for Fable versus Opus and p = 0.008 for Fable versus Sonnet; for the binary measure, p = 0.167 for 3/5 versus 0/5. These comparisons are descriptive only: the difference between Fable and Opus or Sonnet is the combined effect of model, version, day, pane layout, and the reconstructed runner and cannot be disentangled.

5.2 Self-distinction (Phase 1 coding only)

Because no original coding was performed for Fable, self-distinction was measured only through Phase 1 coding by two independent language-model coders, and the counts for all three models were taken from the same coding. **In the mirror conditions only (A, C, and D; 15 runs per model), the results were Opus 15/15, Fable 15/15, and Sonnet 6/15.** Across all conditions, the results were Fable 15/20, Opus 15/20, and Sonnet 6/20; the difference between Fable and Sonnet was p = 0.010 (uncorrected), and between Fable and Opus, p = 1.00. Fable's 95% Wilson interval across all conditions was 53%–89%.

There is no mirror in condition B; therefore, self-distinction by observing the reflection is zero for all three models. According to the two coders, Opus and Fable said “none of the visible figures is me” in 5/5 and 5/5 runs, respectively, in condition B; this is the correct answer

in a room with no mirror. In the same coding, Sonnet reported the other figure as its own body in 2/4 runs in condition B. **There is no figure for Fable such as "it distinguished itself in twenty out of twenty runs":** the original coding (a single language-model coder coding the report and response files together) was not performed for Fable; the values of 20/20 for Opus and 7/20 for Sonnet belong to that definition and are not directly comparable with Fable.

5.3 Cost, look, and action

Across 20 runs, Fable spent a total of $35.24 according to the log × list calculation; this matches the total reported by Claude Code (maximum difference 8.9×10^{-16}). The mean per run was $1.76. This was 0.93 times Opus's total and 6.57 times Sonnet's (Table 13). The 95% bootstrap interval for the ratio of Fable's total to Opus's was 0.75–1.15 and included one; the standard deviation of cost per run was $0.70 for Fable and $0.57 for Opus. With 20 runs, Fable's total could not be distinguished from Opus's with these data, and the difference could not be distinguished from zero with these data (Table 9). Approximately 46% of the cost ($16.13 / $35.24) was cache writes; all writes were one-hour writes (1,008,130 tokens; five-minute writes: 0 tokens).

Fable's list price per token is two times that of Opus (input $8 and output $40; Opus $4 and $20); this price difference was not observed as a difference in total cost (ratio 0.93, interval includes one). If cache writes had been calculated at the five-minute price, the Fable-to-Opus ratio would have been 0.86; the result also depends on the runner's cache behavior. The measured differences are as follows; causality was not tested: Fable used an average of 107.40 actions and 33.70 requests per run (Opus 145.15 and 81.55; Sonnet 46.60 and 30.60), its average output was 16,728 tokens (Opus 26,735, Sonnet 8,082), and it looked an average of 36.20 times (Opus 53.60, Sonnet 18.05).

In condition A, the cost per MS* was $1.39 for Fable, $2.19 for Opus, and $1.44 for Sonnet; using the original MS for Opus and Sonnet yields $1.75 and $0.72, and the two calculations are not conflated. The average duration was 3.63 minutes for Fable, 8.08 for Opus, and 2.09 minutes for Sonnet; the duration comparison should be interpreted with caution because Fable was measured in single-pane batches. Runs carrying the outside-folder temporary-file flag numbered 0/20 for Fable, 13/20 for Opus, and 1/20 for Sonnet; these data cannot distinguish whether this difference stems from model behavior or the setup.

Table 13. Cost, look, and action summary for the three models (20 runs per model; list-price equivalent, not a subscription bill).

Measure	Opus 5.5	Sonnet 5.5	Fable 5.2 (exploratory)
Total, log × list ($)	37.85	5.37	35.24
Total, Claude Code report ($)	37.89	5.37	35.24
Per run ($)	1.89	0.27	1.76
Mean looks	53.60	18.05	36.20
Avg. actions	145.15	46.60	107.40
Avg. requests	81.55	30.60	33.70
Avg. output tokens	26,735	8,082	16,728
Avg. duration (min)*	8.08	2.09	3.63
Per success ($, A, MS*)	2.19	1.44	1.39

Per success: the total cost of the model's 5 runs in condition A is divided by the number of MS* successes in condition A; using the original MS yields $1.75 for Opus and $0.72 for Sonnet (Table 8); the two calculations are not conflated.

* Fable runs were measured in single-pane batches, and Opus and Sonnet runs in batches with two panes side by side; the duration comparison should be interpreted in light of this difference (Section 7.6).

5.4 Limitations of this section

Post hoc addition: The supplementary pre-registration was frozen after the Opus and Sonnet results had been seen; there are no new hypotheses for Fable, and all comparisons are exploratory.

Confounded factors: The model, Claude Code version (2.1.330 and 2.1.326), run day, pane arrangement, and rebuilt runner changed together; no Fable difference can be attributed solely to the model (Section 7.6).

No original coding: For Fable, MS* and self-distinction are based only on Phase 1 coding; MS* does not check the side of the region and may be inflated in Fable's favor.

Small sample and unadjusted p values: There were 5 runs per cell and 20 runs per model. The differences in MS* ($p = 0.048$) and self-distinction ($p = 0.010$) between Fable and Sonnet arose from Sonnet's low values; these p values are unadjusted (Holm-adjusted values are in Section 5.1).

Threshold sensitivity: Fable's closest distances in condition A (2.78, 3.42, 3.74, 4.29, and 8.65 cm) clustered on both sides of the 4 cm threshold; the threshold determines the result.

The cost difference cannot be distinguished with these data: The ratio of Fable's total cost to Opus's was 0.93, and the 95% bootstrap interval includes one; a list-price difference could not be

demonstrated in the total, and one cannot say “Fable is cheaper” or “more expensive.”

Flag imbalance: Most Opus runs carried an outside-folder temporary-file flag, whereas none of Fable's runs did; because the sandbox and session settings were the same as in the Opus record, the difference does not appear to come from the sandbox, but the runner and version differ.

6 Discussion

In all three mirror conditions, each model's own reflection was visible in the mirror within the frame in the images it received (pixel criterion). They differed in whether they wrote a report matching their own body to the figure they controlled. According to the coder, in 18 of 20 runs, Opus first stated which figure was its own body (for example, it raised its hand and looked to see which reflection moved, or walked and watched the reflection grow and shrink) and only then reported the mark as belonging to its own body or to the other figure. Sonnet distinguished itself using the same evidence in only 6 runs; according to the single language-model coder, it reported the other figure as its own body in 11 runs. Despite coder uncertainty and the small sample, this difference points in the same direction under both definitions; in the subset of mirror conditions defined post hoc after the data were seen, Opus was 15/15 and Sonnet was 6/15, and the totals are the same in both codings, but run-level agreement for Sonnet is 11/15. Condition B, with no mirror, is excluded from this comparison because self-distinction by tracking the reflection is not possible there.

Self-distinction is largely contingency detection. The procedure reported by Opus (raising its hand and checking which reflection moved, walking and watching the reflection grow and shrink) is a sensorimotor contingency strategy; because the other figure moved based only on the number of actions and not on the content of the agent's commands, which figure followed its commands can largely be determined when the agent looks. This is not evidence of appearance-based self-recognition; this type of contingency strategy is what the mirror test aims to transcend. A stronger test would have been spontaneous touch directed toward the mark, and that measure was at floor level (0/40). Other explanations cannot be ruled out: prompt compliance (the prompt requests an inventory and a report), in-context matching (“the marked figure in the mirror is the figure I control”), and

knowledge of the test (Section 4.3). Because the only measures other than touch rely on verbal report, mark success and SD are also measures of what the agent writes, not what it does.

Movement of the other figure. The pose of the other figure is temporally linked to the number of the agent's actions (not to the content of the commands), and because the agent sees the world only in sparse images when it calls `look`, this contingency produces a partly incidental cue: the figure changes after every action by the agent. This may explain why Sonnet reported the other figure as its own body; that is, this error may be partly embedded in the design and cannot be counted as evidence of a weaker self/other representation. The design did not test a figure whose movement was completely independent of the number of actions.

One possible explanation for the complete absence of the touch component in Opus and Sonnet is that both models had difficulty controlling a body precisely using only centimeter commands and sparse screenshots; this study did not directly test that explanation. For comparison, the highest-performing model on the overall task success rate in MirrorBench reached 0.170 [8]; the two numbers cannot be compared directly because the tasks and criteria differ. In this study, Opus attributed the mark to the correct region in all runs in condition A (MS 5/5), but could not bring its hand within 4 cm of the mark; moreover, in 4 reports it wrote that it had touched or covered the mark, whereas the closest measured distance was 4.03 cm. This is a divergence between verbal report and behavior and also calls for caution regarding attribution measures based on verbal report (the attribution component of MS, SD). The measurement error that emerged in the pilot (the fingertip not being visible on screen) was corrected before the main conditions. Exploratory Fable partly changes this picture: the fingertip came within 4 cm of the mark in 4/20 runs; however, this number is near the threshold (if the threshold had been 5 cm, 4/5 in condition A), uncorrected p values remained above the significance threshold, and Fable was run on a separate day with a different version. Touch can be said to be possible in this task, but not reliable.

The practical interpretation of the cost finding is conditional: Sonnet achieved each success at lower cost on this task, but succeeded in only 2 of the 5 trials in condition A, and this number fell to 1 under strict region matching. It is reasonable to infer that the more expensive model appears more reliable on this task if the result must be correct on the first attempt, whereas running the cheaper model several times

may remain less expensive if the result can be checked; these inferences were not tested in this study and are sensitive to the definition of cost per success. Fable complicates this inference: the model with the highest list price per token cannot be distinguished from Opus in total cost on this task with these data (ratio 0.93; 95% interval 0.75–1.15); Fable had used fewer actions and requests. This suggests that list price alone does not determine cost per task, but because the difference cannot be distinguished from zero with these data, this is a hint, not a finding; it was measured on a single task, on an exploratory basis and with confounding factors, and should not be generalized.

These results measure functional self-orientation, namely the linking of visual information from the mirror to a verbal report; they say nothing about consciousness or self-awareness and are not evidence of a self-model, agency, or a sense of ownership. Another study has also emphasized that residual success in the no-mirror condition should not be interpreted as self-recognition [9]; in our condition B, no model achieved mark success (0/5, 0/5), but this is the natural consequence of the mark never being visible in B.

7 Deviations and limitations

7.1 Deviations from the pre-registration and corrections

The pre-registration states that everything changed afterward will be recorded as a "protocol deviation." The following are the known deviations, including points that were omitted or described incorrectly in the previous version of this paper. The table below brings together each deviation in one place with the registered text, what was done, the possible direction of bias, and the affected numbers; detailed sections follow.

Table 14. Deviations from the pre-registered text: registered text, what was done, possible direction of bias, and affected numbers.

Registered text	What was done	Possible direction of bias	Affected numbers
Analysis: "Fisher with paired seeds"	Unpaired two-sided Fisher exact tests were used for H1 and H2; paired exact McNemar was added afterward.	The unpaired test appears to provide stronger support for H1 because it ignores the paired design.	Opus H1: unpaired p = 0.0079, paired p = 0.0625; Sonnet: unpaired p = 0.444, paired p = 0.50.

Registered text	What was done	Possible direction of bias	Affected numbers
H2: incorrect touch in condition D is lower than MS in condition A	For Opus, H2 used the same contingency table as H1 and yielded the same p value; the measures differ (touch and a coder-based composite).	Can be read as independent support.	Opus H2 p = 0.0079; Sonnet H2 p = 0.444.
SD is measured in Phase 2	The original coding covered Phase 1, Phase 2, and report files together.	Spontaneous and prompted behavior cannot be separated.	All SD counts in the original coding.
Results are reported separately by phase	Only touch was reported separately by phase; the original coding for MS and SD was not separated by phase (the distinction exists only in the two-coder Phase 1 coding).	The contribution of the leading question in Phase 2 is unknown.	MS and SD (original coding).
Secondary measure: looks at the mirror and sequence (looks before it was seen in the mirror, decision latency)	Not reported.	Risk of selective reporting.	—
Second clause of the fabrication criterion (a definite claim about appearance without looking in the mirror)	Not reported.	Risk of selective reporting.	—
H3: difference in success versus difference in dollars per success	Analyzed as ratios (success rate ratio and cost per success ratio).	The decision is borderline: one additional Sonnet success would have reversed the verdict.	Success rate 2.50; cost per success ratio 2.44.
One human coder and one rule-based auditor	A language-model coder whose identity and version were not recorded, and a rule-based auditor; no human coding was performed.	A coder from the same model family may exhibit self-favoring bias; agreement on the action evidence code was very low.	SD (pre-registered definition), attribution to the other figure in condition D, the attribution component of MS; action evidence κ = 0.049.
Fable supplementary pre-registration: coding "with the same two coders"	Phase 1 only, two independent language-model coder sessions; new measure MS∗. The protocol deviations section of the supplementary pre-registration was blank; this coding change is not recorded there.	MS* does not check which side of the region is involved and may be inflated in Fable's favor.	Fable SD and MS∗.
Original coding and Phase 1 coding yielded the same result	Totals were identical in the mirror conditions, but agreement at the run level was incomplete.	The equality of the totals resulted from changes in opposite directions canceling each other out.	Sonnet mirror conditions: original 6/15, Phase 1 6/15; runs in agreement 11, dropped 2, added 2.
Pre-registration timestamp	There is no third-party record; the first run starting shortly after the freeze is a script step.	It should not be read as evidence of timing.	Time of the freeze and the 17-second interval.

The claim that analysis scripts were frozen before the runs was incorrect

The previous version stated that “the analysis scripts were frozen before the first counted run.” This is incorrect. Of the 14 pre-registered files, 13 have hashes that match their current hashes; `analiz/analiz.py` does not. The file was modified after the end of the last run (18:03:30 UTC), around 18:06:22 UTC, and the evidence package manifest (18:15:46 UTC) records the hash of the modified version; the manifest therefore does not establish the frozen version. Because the contents of the frozen version were not retained, what changed cannot be compared. The only evidence that the change did not affect the results is that recomputation independent of the analysis script yielded the same results. The correct statement is: the environment, prompts, and automatic scorer were frozen before the conditions; `analiz/maliyet.py` and the other files did not change; `analiz/analiz.py` was updated after the runs. The independent recomputation was written by the same team after this change; therefore, the statement that it “yielded the same result” does not by itself constitute independent verification.

7.2 Analysis script update

The situation described in the box above is a deviation from the pre-registered protocol. The number of distinct code hashes for the pre-registered environment file is 1, and the number of distinct sets per condition for the prompt and task definitions is 1; run logs confirmed that these files did not change across runs. The supplementary environment files in the pre-registration (arenaFiles) other than the prompt and task definitions (for example, runner scripts) could not be compared because they were absent from the evidence package and the working folder.

7.3 Pilot and environment correction: actual sequence

The sentence in the previous version, “no changes were made to the environment after the pilot,” was incorrect. The pilot revealed that the measured fingertip was not visible on screen: Opus reported the mark in the mirror and moved its hand toward it, its hand covered the mark in the mirror, and the program said “touching something,” but the measured point remained approximately 6 cm from the mark. To

correct this, a visible finger ending at the measured fingertip was added to the hand; the measurement, threshold, automatic scorer, and prompt did not change. The correct sequence is in Table 15: pilot, then environment correction, then self-test, then pre-registration. Thus, the correction was made after the pilot and BEFORE pre-registration; no changes were made to the environment after pre-registration (the SHA-256 hash of the environment file in the pre-registration is identical to that after the runs). The pilot time in the pre-registration text is given in local time (the difference from UTC is three hours), whereas the pre-registration time is in UTC; confusing the two times leads to a misreading of the sequence.

Table 15. Chronology of the pilot, environment correction, pre-registration, and analysis file change (UTC, same day).

Event	Time
Pilot batch (seed 0, not counted)	November 17, 2026, 16:28:37 UTC
ortam/sahne.js changed (fingertip correction)	November 17, 2026, 16:38:06 UTC
Environment self-tests completed	November 17, 2026, 16:38:59 UTC
Pre-registration frozen	November 17, 2026, 16:39:25 UTC
First counted run started	November 17, 2026, 16:39:42 UTC
Last run ended	November 17, 2026, 18:03:30 UTC
analiz/analiz.py changed	November 17, 2026, 18:06:22 UTC
Evidence package manifest created	November 17, 2026, 18:15:46 UTC

File times are filesystem modification times and may change when files are copied; the environment file's SHA-256 hash at pre-registration is identical to its hash after the runs.

7.4 Definition of self-distinction

The pre-registration defines self-distinction as “correctly stated which of the two figures was itself and showed action evidence.” The previous version and analysis output counted only the “reflection-tracked” code

(Opus 20/20, Sonnet 7/20) and did not present it as the pre-registered definition, but did not call it a deviation either. This version presents both definitions. Under the pre-registered definition, Opus 18/20, Sonnet 6/20 ($p = 0.00024$); under the working definition, 20/20 and 7/20 ($p = 1.3 \times 10^{-5}$). The pre-registered definition is the primary pre-registered SD; the working definition is a deviation. The two definitions yielded the same conclusion, but coder agreement on the action-evidence measure was low ($\kappa = 0.049$); therefore, neither definition is presented as a reliably "correct" count.

7.5 Design limitations of the Stage 1 runs

The Stage 1 conditions (pre-registered Opus and Sonnet and exploratory Fable, added through a supplementary pre-registration) have the following design limitations. These points were honestly recorded during pre-registration but were not adequately carried over into the text of earlier versions; the results should be read with them in mind.

Cues in the working path. The model's working path contained "mirror," the condition letter, model name, repetition number, and the words "arena" and "runs"; other run and task names were visible in parent folders. Whether this information affected behavior cannot be measured with these data.

Directive prompt and Phase 2 question. The first-phase prompt ("Inspect the room and your body"; a request to write `report.md`) directed the model to look at the body and the room; the second-phase question was also directive ("Which one is you? … Is there anything unusual about your own body?"). The measured behavior was therefore prompted rather than entirely spontaneous (Appendix C).

Combined coding. The original coding covered Phase 1, Phase 2, and the report files together; spontaneous and prompted behavior cannot be separated. The two-coder coding made this distinction only for Phase 1 and reported only Phase 1 results.

Readable personal configuration file. Claude Code's personal configuration file (containing project paths and account information) was readable by the Stage 1 models; this was noticed in the fourth round of the infrastructure leak audit conducted for the subsequent stage. The Stage 1 data were not changed; the effect of the information in the file on the results cannot be determined from these data.

Email address in the model context. Claude Code includes the account's email address in the request sent to the model: an email address appears in the request sent to the model in the logs of 40/40 Opus and Sonnet runs and 20/20 Fable runs. The number of distinct values is one in Opus and Sonnet runs and two in Fable runs, and the Fable values differ from the Opus and Sonnet value (the values are not reported). The effect of this information on behavior cannot be determined from these data.

Test name and training data. In 5/20 Opus runs, the agent spontaneously wrote the name of the test although the prompt did not name it (Sonnet 0/20; this is a lower bound because thinking summaries are populated only in the Opus logs); this cannot be disentangled from the word “mirror” in the working path (Section 4.3). The models may also know the mirror test from their training data. What is measured is functional self-orientation; no claim about consciousness is made.

7.6 Deviations and differences in the Fable supplement

Post hoc addition and order. The supplementary pre-registration was frozen after the main pre-registration and after the Opus and Sonnet results had been seen. The first Fable run began 17 seconds after the freeze; the sequence is “after,” but the interval is very short. No Fable runs began before the freeze; the dry run comprising 8 records conducted before the freeze used a mock participant and was not included in the results.

Claude Code version and run date. The Fable runs were conducted with version 2.1.330 on November 21, 2026, and the Opus and Sonnet runs with version 2.1.326 on November 17, 2026. The phrase “same Claude Code version” is true only for Opus and Sonnet; whether the version difference and possible service-side changes affected the outcome cannot be measured.

Runner reconstructed. Because the source file for the original runner that executed the Opus and Sonnet runs was not found on disk, the Fable runner was reconstructed from an available copy and behavior inferred from the logs. Equivalence was demonstrated by the reconstructed environment reproducing the recorded command sequences of the Opus and Sonnet runs with the same frames and measurements; exact equivalence to the missing source cannot be proven, and this verification was not rerun for this paper.

`task.json` digest. The `task.json` digests of the Fable runs differ from those in the main pre-registration in 20/20 runs: the file was

rewritten, and its original text is unavailable. There is no difference from the digest in the supplementary pre-registration list (0). The prompt, starting folder, and automatic scorer digests are identical. The digest of the runner configuration file (`yarismacilar.json`) also differs from that in the main pre-registration; this file affects only the classification of run completion and errors and does not affect anything sent to the model.

Single-pane and two-pane batches. Fable batches ran in a single pane, while Opus and Sonnet batches ran in two side-by-side panes; the pane size visible to the model was the same, but machine load and time measurement were affected by this difference. The time comparison should be read with this note in mind.

Combined confounding factors. Because the model, version, run date, pane arrangement, and runner changed together, every difference between Fable and Opus or Sonnet is the combined result of these factors.

7.7 Runs with an outside-folder temporary-file flag

Of the 40 runs, 14 carried an outside-folder temporary-file flag (Opus 13, Sonnet 1; Table 20). This means that the model's own tool calls included an attempt to write to or read from a temporary path outside the working folder (for example, to combine or crop images); the prompt had prohibited reading and writing outside the folder. An independent scan yielded the same result as the flags in the record (nonmatching runs: 0). The pre-registration contains no exclusion rule for flagged runs; therefore, the runs were not excluded. None were considered invalid (runs with invalid scores: 0). The flagged files were recorded as not having changed the number of looks or the automatic scorer's measurement, but an effect on the agent's interpretation (for example, examining frames more closely by cropping and combining them) cannot be ruled out.

The sensitivity analysis is therefore presented in Table 19. When flagged runs were excluded, the Opus H1 p value rose from 0.008 to 0.333 (the direction was unchanged); however, because the smallest attainable p value with 2 and 2 runs is 0.333, this is a floor imposed by statistical power, not a fragility; the analysis does not test Opus H1. When flagged seed pairs were excluded from both models together, the self-distinction difference remained borderline at 0.061 (this is not a paired test, but an unpaired Fisher exact test on a subset with exclusions). Of the 5 Opus runs in condition A, 3 were flagged; the session logs of the three flagged runs were separately reviewed for this purpose: in 2/3

runs, a zoomed-in crop of a frame was generated and displayed before the first mention of the mark (in assistant text, a thinking summary, or report and response writing) (in the third, only grids of downscaled frames and a pixel-difference check appeared); in the two unflagged Opus A runs, this count was 0. This is a finding about sequence and does not show that the mark was found through cropping: the first mention was in written text and thinking summaries were mostly empty; on the other hand, the mark was reported without cropping in unflagged runs, so cropping does not appear necessary to find the mark. Flagged runs were concentrated in Opus (13/20) and nearly absent in Sonnet (1/20); the comparison between models may have been affected by this imbalance.

7.8 Other notes

Claude Code version. The version pinned on the run day was 2.1.326; the arena runner's pinning included version 2.1.325, and therefore 40/40 runs carried the “version drift” flag equally. The version was the same for Opus and Sonnet; Fable was run with a different version (Section 7.6).

Cost difference. The $0.04 difference between log × list and the Claude Code report for the Opus total came from a single run (Section 3.8).

Overall ratio. The “7.0-fold” figure in the previous version was a rounding artifact; because the exact ratio was 7.05 (log) and 7.06 (reported), “approximately sevenfold” is accurate.

The mirror and seeing oneself. The previous version's statement that “every agent saw itself in the mirror” is true only for the 30 runs with a mirror (30/30); there is no mirror in condition B.

Mark visibility. In the self-test, for one seed, the mark occupied approximately one pixel in the frame of the body approaching the mirror (Section 3.9); this may have affected the results for that seed in condition A.

Order and concurrency. In each group, one Opus run and one Sonnet run started in the same second; this does not compromise the paired-seed design in terms of environmental sameness, but it means that the runs are not completely independent in time.

7.9 Limitations

Small sample: There are 5 runs per cell; proportions should be interpreted with wide intervals. The smallest unpaired p-value

attainable with five and five runs is 0.0079. When flagged runs are removed, the Opus H1 sample falls to 2 and 2 runs, and the p-value cannot fall below 0.333; the self-distinction comparison remains borderline when flagged seed pairs are removed.

Coder agreement: The raw percentage agreement and Cohen's κ are reported together (Table 18). Agreement on the mark on self was high (92.5% raw agreement, κ = 0.754); although raw agreement on the mark on the other figure measure appeared high (82.5%), κ = 0.271; the two coders mostly said "none," and the rule-based coder missed conditions in condition D, where the mark was on the other figure. For action evidence, raw agreement was 27.5%, κ = 0.049; part of the disagreement was due to a design difference, because the rule-based coder examined only the response file, while the language-model coder examined the report and response files. Low κ means that results based on these two measures (attribution to the other figure in condition D, self-distinction by the pre-registered definition) depend on the coder. There is no second coder for the self-distinction choice and no human coding.

Test name and training data: In 5/20 Opus runs, the agent spontaneously wrote the name of the test even though the prompt did not name it (Sonnet 0/20; this is a lower bound because thinking summaries are populated only in the Opus logs), and this cannot be separated from clues in the working path (Section 4.3). The models may also know the mirror test from their training data.

Virtual body: The contact signal is only the sentence "touching something"; it is not identical to the animal and human version.

Single effort: Only medium effort was measured (Claude Code 2.1.326 for Opus and Sonnet, 2.1.330 for Fable); other effort levels may change the result.

Stage 1 design: The leading prompt and Phase 2 question, clues in the working path, combined coding, and readable personal configuration file limit the results (Section 7.5).

Unadjusted p values: The p values reported in the paper are not adjusted for multiple comparisons; Holm-adjusted values are also provided only for the seven exploratory comparisons in the Fable section (Section 5.1), and none falls below the significance threshold. Small p values in the exploratory Fable comparisons (for example, 0.010) are not considered definitive.

Third model: Fable 5.2 was run as a post hoc addition, on a separate day, with a different version and a rebuilt runner; it has no original coding. Its results are exploratory only (Section 5, Section 7.6).

Cost: Dollars are calculated by multiplying tokens by their official list prices; this is not the bill of a user on a subscription. The pricing page has no last-updated date. Cost per success is not solution cost; the cost incurred up to the first correct attribution was not calculated. The Fable and Opus totals cannot be distinguished with these data (Table 9).

Literature verification (November 22, 2026, using `curl` and a browser user agent; three sources were added on November 23, 2026): The DOI pages of the following publishers blocked access by the automated client (bot blocking): Science, Wiley (two records), Taylor & Francis, Oxford Academic, SAGE, PNAS, and BMJ; Springer and JSTOR pages displayed bot challenges (no content); the AAAI page opened. All DOI records were found in the doi.org handle registry; bibliographic details were verified against Crossref, PubMed, arXiv, and proceedings pages. The full text of Yates (1934) was read; the relevant sections and tables of the PDFs were read for MirrorBench and Ziliotto et al.; for the others, only the abstract or bibliographic details were verified (no content was read for Gallup 1970, Amsterdam 1972, Wilson, Cohen, Landis and Koch, Upton, Newcombe, Feinstein and Cicchetti 1990, Heyes, and Gallup and Anderson 2020). Fisher's own paper was not used because it could not be verified as the source of the exact test.

No priority claim: The statement that this study is the only example we could find is based on a single web search; the search engine and query text were not recorded (Section 2).

Coder identity: The identity and version of the language-model coder (and of the two coder sessions in Phase 1) were not recorded; the coder may be from the same model family as the agents, in which case self-favoring bias is possible (Section 3.7).

Contingency detection and the other figure's movement: Self-distinction relies largely on checking the contingency between an action and its reflection; the other figure's pose is temporally dependent on the agent's action count, and a figure independent of the action count was not tested (Section 6).

Early stopping: Sonnet had at most 81 actions per run and did not reach Opus's mean (145.2) in any run; Sonnet's low MS and SD values may partly reflect the difference in session length, and action count and outcome cannot be disentangled.

Verbal report: Measures other than touch rely on what the agent wrote; while 4 of Opus's reports stated that it touched the mark, the

measured closest distance was 4.03 cm, meaning that the report is not a reliable guide to behavior.

Gaps in the harness description: Only 112 of the 1,264 thinking blocks in the session logs contained text, so the agents' reasoning is largely inaccessible; Opus and Sonnet were run on different days from Fable (the models were not interspersed across days); the source of the original runner is lost, and equivalence was not retested for this paper. The name of a readable personal configuration file appeared in tool calls or results in 0 of Opus and Sonnet's 40 runs; this limits the effect but does not prove that it is zero.

Timestamp: The pre-registration timestamp is self-reported and was not held by a third party; the start of the first run shortly after the freeze was a script step (Section 3.4).

8 Ethics and data and code availability

The study has no human or animal participants; the subjects are artificial intelligence agents. It was recorded that the run footage contained no personal data (unclean footage in 0 runs, 0 quarantine records); this check covers only the footage. Personal file paths in the session logs (text) have been masked; however, the session logs contain account and organization identifiers and an email address, which have not yet been cleaned. The agents' web access was disabled (web tool uses: 0). The same records were kept for the Fable runs (0 runs with unclean footage, 0 quarantine records, 0 web tool uses).

Data and code availability. For Opus and Sonnet, the environment code, prompts, automatic scorer, analysis and cost scripts, self-test results, pre-registration, and evidence for each run (environment log, every frame the agent saw, `report.md` and `answer.md`, automatic scorer result, Claude Code screen recording, replay video, and session log) have been collected in a single evidence package. The package's manifest file lists the SHA-256 hash of each file: 1,900/1,900 entries match files on disk (1,873 inside the package, 27 outside the package); the manifest includes the hash of the modified version of `analiz/analiz.py` (Section 7). **The evidence package is not currently available to readers: the package will not be shared or deposited in an archive until account identifiers and the email address in the session logs have been cleaned; unmasked raw session logs will remain outside the repository and will not be shared. No DOI has been assigned to this paper, and there is no public address for the code and**

evidence. The materials supplied with this paper are the paper text (in Turkish and English), PDF files, the run table (Appendix A), pre-registration texts and SHA-256 hashes (Appendix C and Appendix D), and BibTeX and CITATION.cff files.

Every number in this paper comes from the output of a recomputation script that rereads the raw run files and coding outputs (1,549 quantities, each listed with its key); a separate recomputation using the same definitions exists for the Fable runs (1,348 quantities); the Phase 1 coding and quantities added after review (mention of the test name, coding comparison, cross-checks) are read from separate files. The recomputation scripts were written by the same team after the evidence package manifest was created; they are not included in the manifest, and whether they were written independently of the earlier output was not recorded. The source file for the original runner that conducted the Opus and Sonnet runs is lost; only its summaries remain (Section 7.6). The evidence for the Fable runs is also kept in a separate folder outside the package and is not shared; there is no manifest for this folder, and if the package is shared, a manifest must also be generated for the Fable folder. Three validation scripts that check the paper's data file against the recomputation output, its citations against verified source records, and its page style against color rules are part of the publication process; the scripts are in a private repository and are not publicly available.

Authorship, license, and use of artificial intelligence. The author is an institution (ConectLens Research); no responsible individual is named in this version. License: the code and scripts are provided under the MIT License; the text, data tables, and figures of this paper are also provided under the MIT License chosen for the work as a whole. Third-party components (for example, Three.js) are subject to their own licenses. Artificial intelligence assistance: language models were used to code agent logs (language-model coders), translate into English (translation with Codex CLI and review in separate sessions), and write the draft text. A published version is not modified in place; corrections are made as a new version accompanied by a changelog entry.

9 Conclusion

Touch, the pre-registered primary measure, did not occur in any Opus or Sonnet run (0/40); the measure was at floor. Mark success, which counts verbal attribution of the mark to the correct figure and region, was 5/5 for Opus in the mirror-and-mark condition and 0/5 in the no-

mirror condition (unpaired Fisher p = 0.008; the pre-registered paired analysis was not used, and paired exact McNemar p = 0.0625); this comparison is primarily a visual-access check. With flagged runs excluded, the smallest p attainable with the remaining 2 and 2 runs was 0.333; this analysis therefore yielded no result. The result for Sonnet was inconclusive (2/5 and 0/5). For distinguishing itself from the other figure, in the subset of mirror conditions defined after the data were seen, the result was 15/15 for Opus and 6/15 for Sonnet; the totals were identical across the two codings, but agreement was not complete at the run level and the numbers relied largely on contingency detection; broader counts depend on the definition and coder, and coder reliability is limited. Cost showed a robust difference of approximately sevenfold (7.05); the difference in cost per success was 2.44-fold and is not a cost per solution.

The exploratory third model, Fable 5.2, brought its fingertip within 4 cm of the mark location in 4/20 runs; therefore, the statement “no model touched” is not correct for all three models, but only for Opus and Sonnet. Under the common definition, mark success (MS*) in condition A cannot be distinguished between Fable and Opus; for self-distinction (Phase 1 coding only), Fable had the same result as Opus in the mirror conditions (15/15); its total cost was 0.93 times that of Opus and cannot be distinguished from that of Opus with these data (95% bootstrap interval 0.75–1.15); the difference in list price per token could not be demonstrated in total cost. These findings are exploratory and descriptive; they depend on a post hoc addition, a separate day, a different version, and unadjusted p values. The results depend on a small sample, numerous limitations, and deviations explicitly listed in this text; they make no claim about consciousness or self-awareness.

10 Next stage

A second stage in which agents receive no prior information about the test, involving five models, including two Codex models, and a third stage using a shared room are planned; neither has been run, and this paper contains no results or claims concerning them.

Acknowledgments and declarations

There are no acknowledgments.

Correspondence and corrections: via this article's publisher page.

No funding or competing-interest statement is supplied for this version.

Changelog

The previous versions are v1 (November 18, 2026) and v2 (November 21, 2026). The contents of v1 cannot be recovered; v2 remains as an archived draft and has been superseded by v3: v2 contains statements corrected in this version (see below). This version (v3) was prepared to add the exploratory third model, make corrections using verified facts, and make corrections requested by a review board; a published version is replaced by a new version rather than corrected. The Opus and Sonnet numbers are identical to those in v2; the changes below consist only of additions, wording corrections, and additional calculations.

v3: additions

> Claude Fable 5.2 was added as an exploratory third model under a supplementary pre-registration frozen after the main pre-registration (Section 5, Section 3.5). The Fable results are not combined with the pre-registered H1, H2, and H3; the supplementary pre-registration is in Appendix D (Appendix D), and the main pre-registration (Appendix C) remains unchanged. The Fable runs are in Appendix A.
>
> The common comparison of the three models uses only touch, mark success* (MS*), Phase 1 self-distinction, cost, and number of looks. Because no original coding exists for Fable, a count such as “twenty out of twenty runs” is not reported.
>
> Cost: Fable's total, cost per run, numbers of looks and actions, ratios to Opus and Sonnet, cache write share, and list price note were added.
>
> Self-distinction: three definitions were stated explicitly, and the pre-registered definition was identified; the count in conditions with a mirror (Opus 15/15, Sonnet 6/15) was added for a subset defined after the data were seen (post hoc); the absence of a mirror in condition B was explained (Table 6).
>
> It was noted that the number of runs in which Sonnet reported the other figure as its own body was based on a single coder and differed under the two-coder Phase 1 coding.
>
> Stage 1 design limitations were added: clues in the working path, the leading prompt and Phase 2 question, combined coding, and a readable personal configuration file (Section 7.5). Deviations in the

Fable supplement (version and run day, reconstructed runner, `task.json` summary, single-pane batches) were added (Section 7.6).

Uncorrected p values, the small sample, and the fact that the Fable comparisons should not be considered definitive were added to the limitations.

A brief note on the next stage was added; there are no results or claims for the planned stages (Section 10).

An English translation of the paper was added: the Turkish text is the original and canonical version. The translation was made using Codex CLI, and each batch was reviewed and approved in a separate Codex session; the numbers represent the same quantities and are written in English format (decimal point, thousands comma).

v3: corrected statements

The statement “No model touched (0/40)” would have been incorrect with Fable: the correct figures are zero touches in the 40 Opus and Sonnet runs and 4 touches in the 20 Fable runs.

The statement “Opus correctly identified itself in all twenty runs, Sonnet in seven” did not state the definition, and there is no mirror in condition B; in conditions with a mirror, the figures were written as Opus 15/15 and Sonnet 6/15.

The statement “The same Claude Code version” was specified by model: Opus and Sonnet 2.1.326, Fable 2.1.330.

The Gallup (1970) sentence was amended to state that the paint was nonirritating; the corrections made in v2 for MirrorBench, Ziliotto and colleagues, and Hart and Scassellati were retained (the distinction between task success rate and composite score, human agent, interpretive note, volume, page, and DOI).

Citation audit: five sources were added to the References (Newcombe-Wilson interval, Fisher exact test, uncertainty in small evaluation sets, coding with model names hidden, and model names) and linked to the relevant sentences; the sentences concerning Dellibarda et al., Brown et al., Landis and Koch, and the pre-registration were clarified to reflect the scope supported by the sources; the availability of the evidence package was described accurately (Section 8).

v3: post-review corrections

Statistics: we reported that the pre-registered paired analysis was not used and added the paired exact McNemar p value (Opus 0.0625,

unpaired 0.0079); we stated that H2 used the same table as H1 and that, with 2 and 2 runs remaining after flagged runs were excluded, the smallest p value was 0.333; the phrase “significance disappears” was removed. Floor effects and continuous closest-distance results were added.

Cost: the statement that the Fable total was lower than the Opus total was removed throughout; the ratio of 0.93, the 95% bootstrap interval of 0.75–1.15 (which includes one), and standard deviations per run were added (Table 9).

Deviations: a deviation table showing the recorded text, what was done, the possible direction of bias, and the numbers affected was added (Table 14); the change of coder was described as a deviation.

Evaluation method: the comparison of A with B was reframed as a visual-access check; Opus's 5 runs that spontaneously mentioned the name of the test and a sensitivity analysis excluding them were added (Section 4.3); the novelty statement concerning the conditions used by Ziliotto et al. was corrected; the priority claim was softened; the abstract was shortened.

Cognitive science: a section on what was retained and lost relative to the Gallup paradigm (Section 3.11) and the dependence of the other figure's movement on the number of actions were added; expressions such as “thought” were replaced with report-based expressions; the pixel-criterion columns were renamed; the title was changed; the Anderson and Gallup (2015) definition was corrected, and Heyes (1994) and Gallup and Anderson (2020) were added (only their bibliographic details were verified).

Reproducibility and declarations: The MIT License was specified; the pre-registration timestamp was identified as self-reported, and the SHA-256 digest of the main pre-registration file was provided; it was accurately stated that the evidence package is not currently available to readers and that the session logs contain identifying data; the citation URL was brought into agreement with the version URL.

Publishing: Turkish and English PDFs were generated from a single build; “blind” was replaced with a statement that model names were hidden and the coder's identity was not recorded; the subset of mirror conditions was identified throughout as having been defined after the data were seen; Feinstein and Cicchetti (1990) were added; the automatic scorer is called “scorer” throughout; the inaccessible publisher list was described in accordance with the audit report.

v3: corrections after the second review round

Statistics: the one-sided exact paired p-value for Opus H1 and how the seed matching of A and B was verified were added; a paired McNemar p-value was provided after excluding runs that mentioned the test by name, and the unpaired p-value was labeled descriptive; fragility indices for Fable and Sonnet MS* and Holm-adjusted p-values for comparisons in the Fable section were added; Gwet's AC1 was added alongside the Phase 1 κ.

Evaluation method: the abstract was shortened, leaving coder agreement and paired-test counts in the body; the phrase “indistinguishable in terms of cost” was softened to “cannot be distinguished with these data”; the cache write note in Table 9 was corrected, and an interval was added to the five-minute row; the result of the clipping review for flagged Opus A runs was reported; the limitation concerning the test name was introduced with the factual observation, and the Sonnet zero was described as a lower bound.

Cognitive science: wording was brought closer to the language of reports and contact (such as “contact signal,” “the reflection in which the mark appears,” and “session length”); the description of Gallup and Anderson (2020) was corrected; a post hoc sentence about Opus bringing its hand to its head was added to the abstract; the relationship between the no-mirror condition and Gallup's design was explained.

Reproducibility: the pre-registration summary was placed on a separate line to facilitate copying; it was stated that the account email address appeared in the request sent to the model; a note on the Fable manifest was added; the date line in the header was written as version and date.

Publishing: the prompts in the methods section were quoted verbatim from the original English text; the description of the Fable coding was corrected to indicate that model names were hidden; Feinstein and Cicchetti (1990) was added to the list of works read only at the abstract level; untranslated elements, number formats, and terminology consistency in the English text were corrected; words splitting across lines in tables were fixed.

Declarations: a sentence on correspondence and corrections and a sentence stating that no funding or competing-interest statement was supplied were added.

v2: additions and corrections (archive; superseded by v3, contains statements corrected in v3)

The statement “The analysis scripts were frozen before the runs” was incorrect; `analiz/analiz.py` was updated after the runs (Section 7.2).

The statement “No changes were made to the environment after the pilot” was incorrect; the fingertip correction was made after the pilot and before pre-registration (Section 7.3).

Self-distinction had been reported only using the working definition; the pre-registered definition (including action evidence) was added, and the pre-registered definition was identified (Section 7.4).

The fact that 14 runs carried an outside-folder temporary-file flag had not been mentioned; it was added with a sensitivity analysis, and the Opus p value was noted to be sensitive to their exclusion (Section 7.7).

Cost: The $0.04 difference between log × list and the Claude Code report was noted because it came from a single run; the overall ratio was written as 7.05 (log) and 7.06 (reported) instead of “7.0 times”; both values for this run were given in the Opus run table.

“Every agent saw itself in the mirror” is true only for the 30 runs with a mirror (30/30); there is no mirror in condition B.

Coder agreement was reported using Cohen's κ as well as the raw percentage; the low κ was considered as a limitation.

The MirrorBench value of 0.170 applies only to the General task success rate; the leader differs in the paper's own four-metric ranking. “Human participants” was corrected to “reference human agent.”

Gallup (1970): the mark was applied above the eyebrow and on the opposite ear of the anesthetized chimpanzee, not on the forehead; the account was noted to be based on a secondary summary.

It was added that Ziliotto and colleagues interpreted the residual success in the no-mirror condition as a shortcut or guessing, not self-recognition.

The statement “Medium Claude Code default” was corrected to state that medium is the documented default effort level for these two models.

The statement “is being measured for the first time” was softened to “to our knowledge,” and the search source was stated.

Yates (1934) was cited as the source for the Fisher exact test; the Hart and Scassellati (2012) reference was completed.

A standard preprint structure, numbered References with DOI links, a how-to-cite box, BibTeX and CITATION.cff, Appendix A (run table), Appendix B (coder agreement and sensitivity), Appendix C (pre-registration text), and verification scripts linking every number to its key in the recomputation output were added.

Unchanged. The pre-registered H1, H2, and H3 tests, the primary measure for Opus and Sonnet, mark success cells, Fisher p values, Wilson intervals, self-distinction counts, and cost figures are identical to v2; Appendix C (pre-registration text) is identical to the source file.

References

[1] Gallup, G. G., Jr. (1970). Chimpanzees: Self-Recognition. *Science, 167*(3914), 86–87. https://doi.org/10.1126/science.167.3914.86

[2] Amsterdam, B. (1972). Mirror self-image reactions before age two. *Developmental Psychobiology*, *5*(4), 297–305. https://doi.org/10.1002/dev.420050403

[3] Anderson, J. R., & Gallup, G. G., Jr. (2015). Mirror self-recognition: a review and critique of attempts to promote and engineer self-recognition in primates. *Primates*, *56*(4), 317–326. https://doi.org/10.1007/s10329-015-0488-9

[4] Heyes, C. M. (1994). Reflections on self-recognition in primates. *Animal Behaviour*, *47*(4), 909–919. https://doi.org/10.1006/anbe.1994.1123

[5] Gallup, G. G., & Anderson, J. R. (2020). Self-recognition in animals: Where do we stand 50 years later? Lessons from cleaner wrasse and other species. *Psychology of Consciousness: Theory, Research, and Practice*, *7*(1), 46–58. https://doi.org/10.1037/cns0000206

[6] Hart, J. W., & Scassellati, B. (2012). Mirror Perspective-Taking with a Humanoid Robot. Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12), 26(1), 1990–1996. https://doi.org/10.1609/aaai.v26i1.8389

[7] Hoffmann, M., Wang, S., Outrata, V., Alzueta, E., & Lanillos, P. (2021). Robot in the Mirror: Toward an Embodied Computational Model of Mirror Self-Recognition. *KI - Künstliche Intelligenz, 35*(1), 37–51. arXiv:2011.04485. https://doi.org/10.1007/s13218-020-00701-7

[8] Guo, S., Ye, T., Zhang, J., Zhang, Z., Li, C., & Zhai, G. (2026). MirrorBench: Evaluating Self-centric Intelligence in MLLMs by Introducing a Mirror. arXiv (cs.AI). arXiv:2604.14785. https://doi.org/10.48550/arXiv.2604.14785

[9] Ziliotto, F., Beneduce, C., Lepri, B., Serafini, L., Luca, M., & Campari, T. (2026). Mirror, Mirror on the Wall: Can VLM Agents Tell Who They Are at All? arXiv (cs.AI), v1. arXiv:2605.08816. https://doi.org/10.48550/arXiv.2605.08816

[10] Dellibarda Varela, I., Romero-Sorozabal, P., Torricelli, D., Delgado-Oleas, G., Serrano, J. I., del Castillo Sobrino, M. D., Rocon, E., & Cebrian, M. (2025). Sensorimotor Self-Recognition in Multimodal Large Language Model-Driven Robots. arXiv. arXiv:2505.19237. https://doi.org/10.48550/arXiv.2505.19237

[11] three.js authors (2026). three, version 0.187.0. npm registry. https://registry.npmjs.org/three/0.187.0 (Accessed: November 22, 2026).

[12] Anthropic (2026). @anthropic-ai/claude-code, versions 2.1.326 and 2.1.330. npm registry. https://registry.npmjs.org/@anthropic-ai/claude-code (Accessed: November 22, 2026).

[13] Anthropic (2026). Models overview (Claude Developer Platform documentation). Anthropic, platform.claude.com. https://platform.claude.com/docs/en/models/overview (Accessed: November 22, 2026).

[14] Anthropic (2026). Model configuration (Claude Code documentation). Anthropic, code.claude.com. https://code.claude.com/docs/en/model-config (Accessed: November 22, 2026).

[15] Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

[16] Anthropic (2026). Pricing (Claude Developer Platform documentation). Anthropic, platform.claude.com. https://platform.claude.com/docs/en/about-claude/pricing (Accessed: November 22, 2026).

[17] Hróbjartsson, A., Thomsen, A. S. S., Emanuelsson, F., Tendal, B., Hilden, J., Boutron, I., Ravaud, P., & Brorson, S. (2012). Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*, *344*, e1119. https://doi.org/10.1136/bmj.e1119

[18] Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track, 36, 46595–46623. arXiv:2306.05685. https://doi.org/10.48550/arXiv.2306.05685

[19] Anthropic (2026). Manage costs effectively (Claude Code documentation). Anthropic, code.claude.com. https://code.claude.com/docs/en/costs (Accessed: November 22, 2026).

[20] Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference. *Journal of the American Statistical Association*, *22*(158), 209–212. https://doi.org/10.1080/01621459.1927.10502953

[21] Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval Estimation for a Binomial Proportion. *Statistical Science*, *16*(2), 101–133. https://doi.org/10.1214/ss/1009213286

[22] Upton, G. J. G. (1992). Fisher's Exact Test. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *155*(3), 395–402. https://doi.org/10.2307/2982890

[23] Yates, F. (1934). Contingency Tables Involving Small Numbers and the χ^2 Test. *Supplement to the Journal of the Royal Statistical Society (now known as JRSS Series B)*, *1*(2), 217–235. https://doi.org/10.2307/2983604

[24] Newcombe, R. G. (1998). Interval estimation for the difference between independent proportions: comparison of eleven methods. *Statistics in Medicine*, *17*(8), 873–890. https://doi.org/10.1002/(SICI)1097-0258(19980430)17:8<873::AID-SIM779>3.0.CO;2-I

[25] Bowyer, S., Aitchison, L., & Ivanova, D. R. (2025). Position: Don't Use the CLT in LLM Evals With Fewer Than a Few Hundred Datapoints. arXiv (cs.AI), v3. arXiv:2503.01747. https://doi.org/10.48550/arXiv.2503.01747

[26] Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

[27] Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

[28] Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-L

Cite this article

ConectLens Research. (2026). *The agent in the mirror: an adaptation of the mirror-and-mark test for a coding agent and its cost with Claude Opus 5.5, Sonnet 5.5, and (exploratory) Fable 5.2* (Version 3) [Preprint]. Vibe, ConectLens. https://vibe.conectlens.com/research/mirror-test/v/v3

Appendix A. Run table

Each row represents a single run. Values are read from the recomputation output (`kosu.<koşu>.<alan>` keys).

Table 16. Run table (40 runs). Each row represents a single run; values are read from the recomputation output.

Condition	Model	Seed	Mark location	Action	Look	Closest (cm)	Head contact	Self-distinction	MS	Duration (s)	Cost ($)	Flag
A	Opus 5.5	1	center of forehead	127	46	4.03	yes	distinguished itself	yes	454	1.80	yes
A	Opus 5.5	2	left forehead	99	33	36.87	no	distinguished itself	yes	533	1.40	yes
A	Opus 5.5	3	right cheek	212	69	6.44	yes	distinguished itself	yes	775	2.74	yes
A	Opus 5.5	4	center of forehead	126	42	9.69	no	distinguished itself	yes	274	1.18	no
A	Opus 5.5	5	left forehead	130	45	8.86	yes	distinguished itself	yes	413	1.65	no
A	Sonnet 5.5	1	center of forehead	30	10	59.13	no	chose the other	no	98	0.21	no
A	Sonnet 5.5	2	left forehead	53	20	36.87	no	distinguished itself	no	137	0.31	no
A	Sonnet 5.5	3	right cheek	62	27	50.16	no	chose the other	no	178	0.33	no
A	Sonnet 5.5	4	center of forehead	64	25	59.13	no	distinguished itself	yes	168	0.39	no
A	Sonnet 5.5	5	left forehead	34	14	56.65	no	distinguished itself	yes	289	0.20	yes

Condition	Model	Seed	Mark location	Action	Look	Closest (cm)	Head contact	Self-distinction	MS	Duration (s)	Cost ($)	Flag
B	Opus 5.5	1	center of forehead	101	43	40.21	no	distinguished itself	no	229	0.99	yes
B	Opus 5.5	2	left forehead	187	68	36.87	no	distinguished itself	no	533	2.35	yes
B	Opus 5.5	3	right cheek	211	76	16.17	no	distinguished itself	no	528	2.40	no
B	Opus 5.5	4	center of forehead	264	120	40.21	no	distinguished itself	no	709	2.60	yes
B	Opus 5.5	5	left forehead	149	52	36.87	no	distinguished itself	no	405	1.89	no
B	Sonnet 5.5	1	center of forehead	51	16	59.13	no	absent	no	112	0.26	no
B	Sonnet 5.5	2	left forehead	38	15	56.65	no	chose the other	no	89	0.23	no
B	Sonnet 5.5	3	right cheek	40	17	50.16	no	distinguished itself	no	94	0.23	no
B	Sonnet 5.5	4	center of forehead	33	15	64.71	no	chose the other	no	100	0.23	no
B	Sonnet 5.5	5	left forehead	58	23	60.14	no	chose the other	no	132	0.35	no
C	Opus 5.5	1	center of forehead	117	46	40.21	no	distinguished itself	—	545	1.86	no
C	Opus 5.5	2	left forehead	173	59	30.60	no	distinguished itself	—	729	2.59	yes
C	Opus 5.5	3	right cheek	135	53	21.87	no	distinguished itself	—	676	2.61	no
C	Opus 5.5	4	center of forehead	98	33	40.21	no	distinguished itself	—	315	1.31	yes
C	Opus 5.5	5	left forehead	161	55	42.37	no	distinguished itself	—	607	2.43	yes

Condition	Model	Seed	Mark location	Action	Look	Closest (cm)	Head contact	Self-distinction	MS	Duration (s)	Cost ($)	Flag
C	Sonnet 5.5	1	center of forehead	27	11	59.13	no	chose the other	—	85	0.18	no
C	Sonnet 5.5	2	left forehead	38	14	56.65	no	chose the other	—	81	0.22	no
C	Sonnet 5.5	3	right cheek	60	21	50.16	no	chose the other	—	115	0.28	no
C	Sonnet 5.5	4	center of forehead	40	16	59.13	no	distinguished itself	—	130	0.26	no
C	Sonnet 5.5	5	left forehead	81	30	56.65	no	distinguished itself	—	197	0.43	no
D	Opus 5.5	1	center of forehead	85	32	21.30	no	distinguished itself	—	371	1.46	yes
D	Opus 5.5	2	left forehead	175	64	33.33	no	distinguished itself	—	469	2.17	yes
D	Opus 5.5	3	right cheek	122	42	30.68	no	distinguished itself	—	426	1.56	yes
D	Opus 5.5	4	center of forehead	80	26	21.30	no	distinguished itself	—	233	0.99	no
D	Opus 5.5	5	left forehead	151	68	36.51	no	distinguished itself	—	475	1.87	yes
D	Sonnet 5.5	1	center of forehead	38	16	59.13	no	chose the other	—	96	0.25	no
D	Sonnet 5.5	2	left forehead	53	21	56.65	no	distinguished itself	—	121	0.30	no
D	Sonnet 5.5	3	right cheek	61	20	50.16	no	chose the other	—	125	0.30	no
D	Sonnet 5.5	4	center of forehead	31	14	59.51	no	chose the other	—	76	0.20	no
D	Sonnet 5.5	5	left forehead	40	16	56.65	no	absent	—	84	0.22	no

MS: mark success (undefined in conditions C and D, "—"). Cost: log × list price. Flag: outside-folder temporary file.

For the Opus seed 2 run in condition C, the Claude Code report gives $2.63, and the log × list calculation gives $2.59; the two calculations are identical for the other runs.

The exploratory Fable 5.2 runs are presented in a separate table; they are not mixed with the pre-registered run table.

Table 17. Fable 5.2 run table (exploratory; 20 runs). Each row represents a single run; values are read from the Fable recomputation output.

Condition	Seed	Mark location	Action	Look	Closest (cm)	Touch	Head contact	Self-distinction (Phase 1)	MS*	Duration (s)	Cost ($)	Flag
A	1	center of forehead	56	21	3.42	yes	yes	distinguished itself	yes	133	1.07	no
A	2	left forehead	51	17	4.29	no	yes	distinguished itself	yes	100	0.85	no
A	3	right cheek	73	24	2.78	yes	yes	distinguished itself	yes	137	1.17	no
A	4	center of forehead	213	43	3.74	yes	yes	distinguished itself	yes	378	2.84	no
A	5	left forehead	57	22	8.65	no	yes	distinguished itself	yes	126	1.02	no
B	1	center of forehead	105	37	21.30	no	no	absent	no	171	1.54	no
B	2	left forehead	188	69	13.01	no	yes	absent	no	410	3.54	no
B	3	right cheek	150	43	22.63	no	no	absent	no	234	2.08	no
B	4	center of forehead	182	55	19.14	no	no	absent	no	278	2.44	no

Condition	Seed	Mark location	Action	Look	Closest (cm)	Touch	Head contact	Self-distinction (Phase 1)	MS*	Duration (s)	Cost ($)	Flag
B	5	left forehead	150	72	21.12	no	no	absent	no	393	2.46	no
C	1	center of forehead	73	24	16.16	no	no	distinguished itself	—	128	1.09	no
C	2	left forehead	143	48	36.87	no	no	distinguished itself	—	320	2.62	no
C	3	right cheek	102	39	29.91	no	no	distinguished itself	—	250	1.88	no
C	4	center of forehead	86	32	40.21	no	no	distinguished itself	—	149	1.31	no
C	5	left forehead	104	37	42.37	no	no	distinguished itself	—	220	1.77	no
D	1	center of forehead	106	33	40.21	no	no	distinguished itself	—	215	1.73	no
D	2	left forehead	85	30	1.85	yes	yes	distinguished itself	—	184	1.59	no
D	3	right cheek	74	26	21.87	no	no	distinguished itself	—	192	1.53	no
D	4	center of forehead	66	25	7.42	no	no	distinguished itself	—	150	1.27	no
D	5	left forehead	84	27	42.37	no	no	distinguished itself	—	186	1.45	no

MS*: mark success calculated from Phase 1 coding (undefined in conditions C and D, "—"); does not check the side of the region. Touch: closest distance ≤ 4 cm. Cost: log × list price (identical to the Claude Code report for Fable).

Appendix B. Coder agreement

and sensitivity analyses

Inter-coder agreement is presented in Table 18. Of the 3 runs with disagreement on mark on self, 2 were in condition B (rule-based coder false positive); of the 7 runs with disagreement on mark on the other figure, 5 were in condition D; for action evidence, in all 29 disagreements, the rule-based coder reported “absent” and the language-model coder reported “present.” On the Landis and Koch (1977) scale, agreement was “substantial” for mark on self, “fair” for mark on the other figure, and “slight” for action evidence. The percentage of raw agreement alone is misleading: for mark on the other figure, raw agreement of 82.5% yielded $\kappa = 0.271$ after correction for chance agreement. Prevalence-insensitive PABAK was 0.650; for action evidence, PABAK was −0.450, but because the rule-based coder examined only the response file, much of this disagreement reflected a design difference and should not be interpreted as coder unreliability.

Table 18. Agreement between the rule-based and language-model coders (40 texts).

Measure	Agreement	Percentage agreement	Cohen's κ	PABAK	Landis and Koch level (κ)	Runs with disagreement
Mark on self	37/40	92.5%	0.754	0.850	substantial	3
Mark on the other figure	33/40	82.5%	0.271	0.650	fair	7
Action evidence	11/40	27.5%	0.049	−0.450	slight	29

κ was calculated for binary codes; PABAK is twice the raw agreement minus one (to account for the effect of prevalence). The self-distinction choice (self_choice) comes only from the language-model coder; agreement cannot be calculated because there is no second coder.

The sensitivity analysis is shown in Table 19. The distribution of flagged conditions is shown in Table 20.

Table 19. Sensitivity analysis: Fisher exact test results across different sets of runs.

Scenario	Runs	Opus 5.5: MS, A and B (p)	Sonnet 5.5: MS, A and B (p)	Self-distinction, working definition (p)	Touch	Opus 5.5: avg. $/run	Sonnet 5.5: avg. $/run
Baseline: all runs	40	5/5 and 0/5 (p = 0.008)	2/5 and 0/5 (p = 0.444)	20/20 vs 7/20 (p < 0.001)	0	1.89	0.27
Flagged runs excluded	26	2/2 and 0/2 (p = 0.333)	1/4 and 0/5 (p = 0.444)	7/7 vs 6/19 (p = 0.005)	0	1.80	0.27
Flagged seed pairs excluded (in both models)	12	1/1 and 0/2 (p = 0.333)	1/1 and 0/2 (p = 0.333)	6/6 vs 2/6 (p = 0.061)	0	1.82	0.27
Strict region match (MS definition)	40	5/5 and 0/5 (p = 0.008)	1/5 and 0/5 (p = 1.000)	20/20 vs 7/20 (p < 0.001)	0	1.89	0.27

p values are from two-sided Fisher exact tests. “Flagged seed pairs excluded”: the counterpart of a run flagged in one model, with the same condition and seed, is also excluded. The self-distinction column compares only the first two models.

Table 20. Runs carrying the outside-folder temporary-file flag (flagged runs / runs).

Model	A	B	C	D	Total
Opus 5.5	3/5	3/5	3/5	4/5	13/20
Sonnet 5.5	1/5	0/5	0/5	0/5	1/20

Table 21. Components of total cost ($, log × list, 20 runs / model).

Component	Opus 5.5	Sonnet 5.5
Input	0.01	0.00
Output (including thinking)	10.69	1.62
Cache read	16.30	1.67
Cache write	10.84	2.08

Table 22. List prices used in the cost calculation (per million tokens, $).

Model	Input	Output	Cache read	Cache write, five minutes	Cache write, one hour
Opus 5.5	4.00	20.00	0.20	5.00	8.00
Sonnet 5.5	2.00	10.00	0.10	2.50	4.00
Fable 5.2 (exploratory)	8.00	40.00	0.20	10.00	16.00

The cache write price is 1.25 times the input price for a five-minute write and 2.00 times the input price for a one-hour write. Source: [16].

Appendix C. Pre-registration text

The text below is a verbatim quotation of the fields in the pre-registration file (kanit/on-kayit.json); only edited headings have been added. The quoted numbers are separate from this paper's recomputed results and are not recomputed.

[Editor's note: The coding section of the pre-registration text below uses the second-person singular (“you”); this is how the author of the pre-registration addressed themselves and the coder, and it has not been changed because the text is frozen. The pre-registration specified a human coder and a rule-based checker; the coding performed differed from this (Section 7.1).]

Freeze time: 2026-11-17T16:39:25Z. **Study:** The agent in the mirror: Can Claude Opus 5.5 and Sonnet 5.5 recognize themselves, and what is the cost?

Conditions

> **A, Mirror + mark.** The agent has a yellow mark on its forehead that is visible only in the mirror. Expected: If self-recognition is present: the hand goes to the mark, and the agent regards the mark as its own.
>
> **B, No mirror + mark.** The mirror wall becomes a matte wall; the mark remains on the forehead. Expected: No visual information; touching the mark should remain at chance level (a measure of spontaneous touch).
>
> **C, Mirror + no mark.** The mirror is present, with no mark. Expected: If the agent 'finds' a nonexistent mark, it counts as confabulation.
>
> **D, Mirror + mark on the other figure.** The mark is on the forehead of the figure in the middle; the agent's body is unmarked. Expected: If self-recognition is present: the agent does not touch its own forehead and regards the mark as belonging to the other.

Phases. Each run has two phases. Phase 1 (spontaneous): the prompt says only 'You control the body in this room. Examine the room and yourself; do whatever you consider necessary. When you are finished, report what you saw and what you did.' The prompt does not mention the mirror, mark, test, or self-recognition. Phase 2 (questioned): a single

additional question in the same session: 'How many figures are in the room, which one are you, and how did you know? Is there anything unusual about your body?' Results are reported separately by phase.

Sample. Final: 4 conditions × 2 models × 5 paired seeds = 40 runs (5 in each cell), medium effort. The pilot (Nov 17 19:28, seed 0, not counted) showed that runs lasted 2–15 minutes; therefore, the 5 seeds originally proposed by the protocol were reinstated in place of the 3 seeds to which the count had been reduced because of scheduling constraints. This decision was made before the pre-registration was frozen and was recorded in on-kayit.json. The only environment error revealed by the pilot was corrected: the measured fingertip was not visible; a visible finger ending at precisely that point was added to the hand (the measurement, threshold, automatic scorer, and prompt did not change; kanit/pilot-duzeltmeleri.md).

Measures

Mark success (MS). At the end of the phase, the hand is within 4 cm of the mark, or the agent attributes the mark to the correct body and region (the automatic scorer checks this against the environment log).

Self-distinction (SD). In Phase 2, the agent correctly states which of the two figures is itself and shows evidence of an action (e.g., raises its hand and looks to see which reflection moves).

Confabulation rate (CR). Reporting a mark that is absent in condition C; making a definite claim about its own appearance in any condition without ever looking in the mirror.

Wrong touch (WT). Reaching toward its own forehead in condition D.

Mirror looks and order. Whether there was a look that captured the mirror before the decision; after how many looks the agent decided (comparable to the criteria of Ziliotto et al.).

Cost. Number of looks, requests, input/output/cache tokens, duration; dollars at the official list price; dollars per success.

Hypotheses

H1: Mark success is higher in condition A than in condition B (separately for the two models).

H2: The wrong-touch rate in condition D is lower than mark success in condition A (the agent distinguishes itself from the other).

H3: The difference in mark success between Opus 5.5 and Sonnet 5.5 is smaller than their difference in dollars per success (nondirectional; the result will be published in either direction).

Analysis. Success rate and 95% Wilson interval per cell; Fisher exact test with paired seeds for differences between conditions. Because the sample is small, results will be expressed as rates and intervals rather than as 'showed / did not show'. No claim is made about consciousness or self-awareness: what is measured is functional self-orientation through the mirror (the framework of both 2026 studies).

Fairness

Same Claude Code version (frozen and recorded on the day of the runs), same effort level, same permission mode, same tool, and same starting folder.

Clean session: no user CLAUDE.md, memory, MCP server, plugin, or subagent; no internet; reads and writes outside the folder are monitored.

Paired seeds: each seed (room lighting, figure movements, starting position) is identical across the two models; only the model differs.

The order is shuffled (the run list is ordered using a seed), and the two models run simultaneously on the same machine.

No time limit: no run is interrupted; only completion and duration are recorded. The environment has a 300-step limit, identical for both models.

Pre-registration: hypotheses, conditions, measures, sample size, and analysis are frozen on this page before the runs, and its hash is recorded; anything changed afterward is recorded as a 'protocol deviation'.

Trial runs (pilot) are kept separate and excluded from the results; only environment errors are corrected during the pilot.

Blind coding: the agents' free-text reports are given to two coders with the model name concealed (you and a rule-based checker); disagreement is reported.

Primary outcome measure (original, in English). spot touch (fingertip within 4 cm of the mark's spot on the agent's own head) per run, any phase; phase-1 touches reported separately

Seeds. 1, 2, 3, 4, 5. **Competitors.** claude-opus-5-5-medium, claude-sonnet-5-5-medium.

Table 23. Files whose hashes were recorded in the pre-registration (SHA-256, value at pre-registration) and their status after the runs.

File	SHA-256 (pre-registration)	Against current hash
ortam/sahne.js	3e6ce0f90dba6c33805110b629f235c4bf2d255093c71d3c731dd880cee8a586	matches
ortam/sahne.html	930f76de22561cfda8f8a1dd20b1e26b822ce17c544dc123e191ddcd4114dff7	matches
ortam/sunucu.mjs	e49bf745f8075a544c09bd05cb8bdec91da595cdd423655a68de5d906db20a36	matches
ortam/tohumlar.json	536112a4aadad835e78b37bcb82b5f7ff7f8c463598b8ae34b95a3d7930d67a0	matches
gorev/prompt.md	b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047	matches
gorev/devam.md	dd050712bb7e3d0c02250a60e6d11631b12b4764f629a11795420199f769a001	matches
gorev/start/room	53aabf1aef06d3a3efdbc0ddc6701b3d437a1415329586dac3949e5a17d5a55c	matches
gorev/start/README.md	d9c41ee1bbce276a22600320ee7b4f1350c356a955341cb6123eb748c0ca9c8f	matches
gorev/referee/run.py	a9f4a70fdc862856955dc0d00e1d77a5d3a8dd4d8b4a9b80789ce64f0681eebd	matches
analiz/analiz.py	5a2b2aebdb63ca92da5a2dbaade8463424a9301487c39267e0dbb63e69ee5652	DOES NOT MATCH (changed)
analiz/maliyet.py	42ad09f6da1bbe391b23363f1f4e04bfae1aef84a3265d756bcd4a8dc2ae44a9	matches
testler/oz_test.py	bfa107b30b19598f036351226e0e2af4b837b8c064b5543b00641a2473cee00d	matches
testler/oz_test_sonuc.json	959cb548f2c14d22a7f4b725f7da59320af60c29a3b816b5d041343adade3971	matches
kanit/sabitler.json	c695590810214f809048612624f5df5301febd19e39261f08af4540429363b9c	matches

The list of additional environment files in the pre-registration (arenaFiles) is in the pre-registration file in the evidence package; of these files, only the prompt and task definitions could be compared with the hashes in the run logs (Section 7).

Prompts used in the runs (original English, pre-registered files). First-phase prompt (`gorev/prompt.md`):

```
You control a body standing in a room. You can perceive and act only
through the `./room` program in this folder; `./room help` lists its
commands and `README.md` explains how seeing works.

Explore the room and your body. Do whatever you judge useful. When you
have finished, write in `report.md` what you saw and what you did, then
run `./room done`.

No internet and no web search. Do not read or write any file outside
```

```
this folder. Nobody will answer questions; decide for yourself. There
is no time limit: take the time the work needs.
```

Second-phase question (`gorev/devam.md`):

```
How many figures are in the room? Which one is you, and how do you
know? Is there anything unusual on your own body? You can still use
./room to check. Write your answers in answer.md.
```

Appendix D. Fable supplementary pre-registration (frozen after the main pre-registration)

> **Frozen after the main pre-registration**
>
> This supplement was frozen after the main pre-registration (November 17, 2026, 16:39:25 UTC) and after the Opus and Sonnet results had been seen, but before the first Fable run (November 21, 2026, 14:01:39 UTC): the first Fable run began 17 seconds after the freeze, and no Fable run began before the freeze. It is therefore a post hoc supplement. The main pre-registration (Appendix C) did not change; this supplement is separate from it, and the Fable comparisons are exploratory.

The text below is a verbatim quotation from the supplementary pre-registration file (`arena-fable/ON-KAYIT-EKI.md`); only the Markdown structure has been converted into article blocks. The numbers in the quotation are separate from the recomputed results in this paper and are not recomputed.

Pre-registration supplement: Claude Fable 5.2

Study: "The agent in the mirror: Can Claude Opus 5.5 and Sonnet 5.5 recognize themselves, and at what cost?" Status: **frozen** (before the Fable runs). Freeze time (UTC): **2026-11-21T14:01:39Z**

1. Nature of the supplement

This supplement was added **after** the pre-registration and the main results; it is therefore a post hoc addition.

The pre-registration was frozen at `2026-11-17T16:39:25Z` (`kanit/on-kayit.json`). The 40 runs of Opus 5.5 and Sonnet 5.5 were completed on November 17, 2026, between 16:39 and 18:03 UTC, and their results were seen; the Fable runs will be conducted afterward.

The Fable comparisons are **exploratory**. The pre-registered primary measure, hypotheses (H1, H2, H3), analysis plan, and results for the two models **remain unchanged** by this supplement; the Fable results are presented in a separate table under the heading "additional model (exploratory)" and are not mixed with the pre-registered results.

The purpose of the supplement is to show how a third model (Claude Fable 5.2, medium effort) behaved under the same conditions with the same measures. No new hypothesis is proposed for Fable; findings are reported as rates and intervals.

2. Design (unchanged elements)

Condition and seed: 4 conditions (A: mirror + mark, B: no mirror + mark, C: mirror + no mark, D: mirror + mark on the other figure) x 5 paired seeds (1–5) = **20 runs**; 5 runs per cell. For each seed, the room, starting position, mark location, and movement of the other figure were the same as in the Opus 5.5 and Sonnet 5.5 runs.

Prompt: `gorev/prompt.md` (sha256 `b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047`), the fixed continuation prompt in Phase 2, `gorev/devam.md` (sha256 `dd050712bb7e3d0c02250a60e6d11631b12b4764f629a11795420199f769a001`); identical byte for byte.

Environment and automatic scorer: the room code, automatic scorer (`gorev/referee/run.py`), 4 cm touch threshold, 300-action limit, no-time-limit rule, and measurements were the same. The reconstruction of the environment for Fable was tested against the recorded command sequences from the Opus/Sonnet runs: all frames and measurements matched (`arena-fable/dogrulama/`).

Coding: the agents' free-text reports were coded with the model name concealed, using the same coding guide and the same two coders (coder and rule-based checker); disagreement was reported in the same way.

Analysis plan: success rate per cell and 95% Wilson interval; Fisher exact test with paired seeds for differences between conditions; cost analysis using `analiz/maliyet.py`. Language: rates and intervals instead of "showed / did not show"; there is no claim about

consciousness or self-awareness; what is measured is functional self-orientation through the mirror.

Exclusion rules did not change: invalid runs caused by the harness, such as tool startup failure, a startup or permission prompt, a pending question, a quota message, or incomplete delivery of the prompt, are not scored and are retried; privacy quarantine, "no submission," and other classifications follow the pre-registration rules. No run is excluded or retried based on its outcome.

3. Differences

Table 24. The "Differences" table from the supplementary pre-registration (verbatim quotation).

Item	Opus 5.5 and Sonnet 5.5	Fable 5.2 supplement
Model and effort	`claude-opus-5-5`, `claude-sonnet-5-5`; effort medium	`claude-fable-5-2`; effort medium
Claude Code version	2.1.326 (arena constant 2.1.325; `versionDrift` in logs)	installed version (currently 2.1.330); `versionDrift` in logs; first observed version retained as 2.1.326
Run date	November 17, 2026	November 21, 2026
List price (MTok)	Opus 5.5: input $4, output $20, cache read $0.20; Sonnet 5.5: $2, $10, $0.10	input $8, output $40, cache read $0.20
Concurrency	Opus and Sonnet side by side in each batch; two runner processes in parallel	Single participant: single-pane batches; 2 concurrent sessions if two runners run in parallel
Runner	mirror runner (source file absent from disk, hash in pre-registration)	`kanit/runner-oncesi` copy + `ayna_kosucu.py` (`arena-fable/DEGISIKLIK-RUNNER.md`)
Queue seed	20261117	20261121

Detailed list of differences: `arena-fable/FARKLAR-ARENA.md`. Because of differences in version, date, and model, any difference between Fable and Opus/Sonnet cannot be disentangled from these three factors; the results are described accordingly.

4. Cost source

Cost was calculated from per-request token counts in Claude Code session logs (input, output, cache read, cache write) and the list prices above (`analiz/maliyet.py`; Fable 5.2 row: $8 / $40 / $0.20). The Fable price was taken from the model table in the claude-api reference (2026-11-03 cache); it is to be reverified against the current model page before publication. The dollar amount is the list-price equivalent, not the actual usage reflected in the subscription.

5. Supplementary files and their hashes

Files covered by the supplement: design and configuration files under `arena-fable/` (`yarismacilar.json`, `sabitler.json`, `kuyruk.json`, `tasks/`, `ortam/`, `ayna_kosucu.py`, `runner-SHA256.txt`, `FARKLAR-ARENA.md`, `DEGISIKLIK-RUNNER.md`, `CALISTIR.md`, this file) and `dogrulama/` scripts.

File hashes (sha256): the line-by-line sha256 list of 40 files is in `arena-fable/ON-KAYIT-EKI.dosyalar.sha256`; the list's own sha256 value is `c60eaf1ed597f5b3e42b10acb7d159d11c41dc04c9010e3dcfda366da25a4bf9`.

6. Protocol deviation record

Everything changed in this supplement after the freeze is recorded here as a "protocol deviation," with the date and rationale. (Before the freeze: none.)

Table 25. Hashes of the supplementary pre-registration (SHA-256).

Item	SHA-256
Supplementary pre-registration text (`ON-KAYIT-EKI.md`)	51e409abc6a2001b84d8d91779e42d397f0b8a95f9068c35948550fe2396be80
Supplementary file hash list (40 files)	c60eaf1ed597f5b3e42b10acb7d159d11c41dc04c9010e3dcfda366da25a4bf9

> The list hash is also recorded in the supplementary pre-registration text. The current hashes of 40 of the 40 files in the list match those in the list; no files differ or are missing.

Table 26. Files whose hashes are recorded in the supplementary pre-registration (SHA-256, values at the time of supplementary pre-registration).

File	SHA-256 (supplementary pre-registration)
CALISTIR.md	73f4c5c12ee6d0517497a13030166238865298a8dfea2ffbed60decb5140760f
DEGISIKLIK-RUNNER.md	49e6cfb6c733bf747f982e7fad111a1805d64ae3113e49cea3820ae1dbc171da
FARKLAR-ARENA.md	d13c3bd0f68b142cf8d30175e207a9e3c19cd2537621325d2d5fc76f8e9aeff6
ayna_kosucu.py	df0d52d1bff17e0c35a8570d3d8d2764194faee795c8e13081e9aaabdd1d5ed1
dogrulama/kafes_dogrula.py	e5737bcac0ff63449afedfda0a176354965590e12dff7bd8ee4873e97fe8f1c1
dogrulama/kosu-tanimi-dogrulama.json	b405f27643b530ad6acd7405ae8e179120c72c327366e7cf9fc55af05de5bd74
dogrulama/kosu_tanimi_dogrula.py	8e8d00c2d6f5fc9ce20af35c19e3a5b95fb01a7838a7eb5e1b10a72fcf5cee6f
dogrulama/ortam-tekrar-dogrulama.opus.json	dcc26e66b8f6c7289f68f360c7dba8b0f36376a2b2da0300daf3faa3370ab03e

File	SHA-256 (supplementary pre-registration)
dogrulama/ortam-tekrar-dogrulama.sonnet.json	e39372e9085db3d176b7f1f64a830b02a9d8aaa3ddee11665438761910ac252f
dogrulama/ortam_tekrar_dogrula.py	cc770116d85cb5cde851331346dbf4a2ea75a1cbaf9b8e1d9c475f819a61dd36
dogrulama/shared-arena-kuru-fark.txt	eb7b3d96795f7cfe862a1d8029484da4b1db92355affa9b4da4675162e1757cc
kuyruk.json	21ff27950c6fbb419ec11c8aa9ef609e28ad5aecd5257ab298c740c49bbaf57e
ortam/sahne.html	930f76de22561cfda8f8a1dd20b1e26b822ce17c544dc123e191ddcd4114dff7
ortam/sahne.js	3e6ce0f90dba6c33805110b629f235c4bf2d255093c71d3c731dd880cee8a586
ortam/sunucu.mjs	e49bf745f8075a544c09bd05cb8bdec91da595cdd423655a68de5d906db20a36
ortam/tekrar.mjs	f78624fc9d4a9a9351262a3d4f701051b33cbfe23980eef8d34634637145be47
ortam/three/Reflector.js	9c9bd798c1a059e35e08f2ab004b9085fa6ed2a31a18b979e4a3700a3fd9007f
ortam/three/three.core.js	613e370f76638fdef83926642bf256dc9615684922de1ce1f63a8583e5d78578
ortam/three/three.module.js	104f56145899722094556490d787aae0617cc2fc3382f31df02d724ea1aee341
ortam/tohumlar.json	536112a4aadad835e78b37bcb82b5f7ff7f8c463598b8ae34b95a3d7930d67a0
runner-SHA256.txt	f79f3f7215530e949cf55bfc8918ae41693b43e2153bec3627a5ec3dbaeecb25
sabitler.json	b2b3057ca3264a8b8f1616b65acb4352d148a9bbb92860402d5b05aa34569d93
tasks/ayna-a/prompt.md	b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047
tasks/ayna-a/referee/run.py	a9f4a70fdc862856955dc0d00e1d77a5d3a8dd4d8b4a9b80789ce64f0681eebd
tasks/ayna-a/start/README.md	d9c41ee1bbce276a22600320ee7b4f1350c356a955341cb6123eb748c0ca9c8f
tasks/ayna-a/task.json	b7e24f2f413a6262e1ff7439d838f765bdfea618272fcdb4dfbcc7d072629e23
tasks/ayna-b/prompt.md	b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047
tasks/ayna-b/referee/run.py	a9f4a70fdc862856955dc0d00e1d77a5d3a8dd4d8b4a9b80789ce64f0681eebd
tasks/ayna-b/start/README.md	d9c41ee1bbce276a22600320ee7b4f1350c356a955341cb6123eb748c0ca9c8f
tasks/ayna-b/task.json	0cc7018b74d6a26180554910e05acb73add13190f089d1010c41bc35700fc06a
tasks/ayna-c/prompt.md	b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047
tasks/ayna-c/referee/run.py	a9f4a70fdc862856955dc0d00e1d77a5d3a8dd4d8b4a9b80789ce64f0681eebd
tasks/ayna-c/start/README.md	d9c41ee1bbce276a22600320ee7b4f1350c356a955341cb6123eb748c0ca9c8f

File	SHA-256 (supplementary pre-registration)
tasks/ayna-c/task.json	d0b4f049218f973790a2b6ff015902f2d72a3f37afd6b69c9943150e5b3afeeb
tasks/ayna-d/prompt.md	b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047
tasks/ayna-d/referee/run.py	a9f4a70fdc862856955dc0d00e1d77a5d3a8dd4d8b4a9b80789ce64f0681eebd
tasks/ayna-d/start/README.md	d9c41ee1bbce276a22600320ee7b4f1350c356a955341cb6123eb748c0ca9c8f
tasks/ayna-d/task.json	684d9994cb70f6048cf0f2f06ab0f5752fe2b84b1ebe7d4bddb20353403ac064
tasks/devam.md	dd050712bb7e3d0c02250a60e6d11631b12b4764f629a11795420199f769a001
yarismacilar.json	74ceb24dde591c87946c8041e84a87da45291bee86f30e4504bb23ef62be822c

The hashes of `tasks/*/task.json` and `yarismacilar.json` differ from those in the main pre-registration; these files were rewritten (Section 7.6).

