PREPRINT · NOT PEER REVIEWED

# Aynadaki ajan: Claude Opus 5.5, Sonnet 5.5 ve (keşif amaçlı) Fable 5.2 ile bir kodlama ajanında ayna-işaret testi uyarlaması ve maliyeti

Author
ConectLens Research (ConectLens)

Version and date
v3, November 23, 2026

DOI
not assigned

ayna testi · öz-tanıma · kodlama ajanları · büyük dil modelleri · Claude Code · Fable 5.2 · ön kayıt · maliyet

VERSION

v3 · November 23, 2026

ARTICLE LANGUAGE

Türkçe

Contents

# Özet

Klasik ayna ve işaret testinin bir uyarlamasını bir kodlama ajanına uyguladık. Ön kayıtlı çalışmada Claude Opus 5.5 ve Claude Sonnet 5.5, gerçek Claude Code oturumlarında, aynalı bir odadaki bir bedeni yalnızca kendi aldıkları ekran görüntüleriyle yönetti; odada ikinci, özdeş bir figür de vardı. Dört koşul, eşli beş tohum ve iki model: 40 koşu. Ortam, istemler ve otomatik puanlayıcı koşullardan önce ön kayıtla donduruldu (zaman damgası kendi beyanımızdır); bir analiz betiği koşulardan sonra güncellendi ve ön kayıttaki bazı analizler değiştirildi ya da raporlanmadı (Bölüm 7). Üçüncü model Claude Fable 5.2 ayrı bir ek ön kayıtla ve yalnızca keşif amaçlı olarak çalıştırıldı (20 koşu; Bölüm 5); ön kayıtlı H1, H2 ve H3'e girmez.

**Başlıca sonuçlar.** (a) Ön kayıtlı birincil ölçüt olan dokunma (parmak ucunun işaret noktasına 4 cm'den yakın gelmesi) Opus ve Sonnet'in 40 koşusunun hiçbirinde gerçekleşmedi (0/40); ölçüt tabanda kaldığı için işaret başarısı dokunmaya değil sözel atfa dayanır. (b) İşaretin doğru figüre ve bölgeye sözel atfı (işaret başarısı) Opus'ta ayna ve işaret koşulunda 5/5, aynasız koşulda 0/5, Sonnet'te 2/5 ve 0/5 oldu; bu karşılaştırma bir görsel erişim kontrolüne yakındır ve ön kayıtta öngörülen eşli analizde sonuç geleneksel anlamlılık sınırına ulaşmaz (Bölüm 4.2). Keşif amaçlı olarak Opus elini başına yalnızca A koşulunda götürdü (3/5; diğer üç koşulda 0/15); bu post hoc bir bulgudur ve elin başa değmesi işarete yönelmeden bağımsız olarak da gerçekleşebilir. (c) Kendini ayırma (KA: kendi bedeninin hangi figür olduğuna ilişkin doğru sözel seçim) ön kayıt tanımıyla Opus'ta 18/20, Sonnet'te 6/20'dir; bu tanımın dayandığı eylem kanıtı kodunda kodlayıcılar arası uyum çok düşüktür. Veri görüldükten sonra tanımlanan (post hoc) aynalı koşullar kümesinde Opus 15/15, Sonnet 6/15'dir (Bölüm 7.4). (d) Keşif amaçlı Fable 5.2'de parmak ucu işaret noktasının 4 cm'i içine 4/20 koşuda geldi; bu fark model, sürüm, gün ve koşturucu bakımından karışıktır ve yalnızca betimleyicidir. (e) Opus'un toplam maliyeti Sonnet'inkinin yaklaşık 7,05 katıdır; Fable'ın toplamı Opus'unkinin 0,93 katıdır ve %95 önyükleme aralığı bir değerini içerir: bu verilerle iki model maliyet bakımından ayırt edilemedi (Tablo 9).

Bu sonuçlar işlevsel bir öz-yönelimi, yani görsel bilginin sözel bir rapora bağlanmasını ölçer; bilinç, öz-farkındalık, öz-model, eyleyicilik ya da sahiplik duygusu hakkında kanıt değildir. Başlıca sınırlılıklar: yönlendirici istem, Opus'un 5/20 koşusunda testin adını kendiliğinden yazması, kimliği kaydedilmemiş bir dil modeli kodlayıcı, hücre başına beş koşu ve kanıt paketinin şu anda okurlara açık olmaması. Hakem değerlendirmesinden geçmemiş bir ön baskıdır.

# 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 Giriş

Ayna testi, bir canlının aynadaki görüntüyü kendisi olarak tanıyıp tanımadığını ölçmenin en bilinen davranışsal yollarından biridir. Gallup (1970) [1], aynaya uzun süre maruz kalmış şempanzelerin, anestezi altındayken kaşının üstüne ve karşı kulağına, uyandıktan sonra ancak aynada görülebilecek kokusuz ve tahriş etmeyen bir boya sürdü; aynayla karşılaşan hayvanın işarete yönelen dokunuşları öz-tanıma kanıtı sayıldı. Boyanın yeri bu çalışmada ikincil bir özetten teyit edilmiştir; Science makalesinin tam metni okunamamıştır (Bölüm 7). Amsterdam (1972) [2] aynı mantığı 88 çocukta (3–24 aylık) uyguladı; özet düzeyinde 20–24 aylık çocukların yaklaşık %65 kadarı kendi imgesini tanıdı, yani tanıma yaşla birlikte belirginleşir (Amsterdam'ın tam metni okunmadı). Testin gücü bir kontrolden gelir: aynaya hiç maruz kalmamış (aynaya naif) hayvanların ve aynaya maruz kalmadan önceki başlangıç davranışının işarete dokunma sıklığı şans düzeyini verir. Bu çalışmadaki aynasız B koşulu bu kontrolün bir uyarlamasıdır; aynaya naif bir ajan ya da aynaya maruz kalma öncesi bir taban içermez. Anderson ve Gallup (2015) [3], testin başlatıcısının ortak yazarı olduğu, primatlarda öz-tanımayı teşvik etme ya da mühendislik yoluyla oluşturma girişimlerinin bir derlemesidir; testi genel olarak eleştiren bir çalışma değildir. Testi geçmenin ne anlama geldiğine ilişkin eleştirel tartışma için bkz. Heyes (1994) [4]; testin başlatıcısının

ortak yazarı olduğu güncel değerlendirme ve savunma için bkz. Gallup ve Anderson (2020) [5]. Bu iki kaynağın yalnızca künyesi doğrulandı, içerikleri okunmadı; türleri yazarlıklarına dayanır.

Bugünün kodlama ajanları dosya okuyup komut çalıştırarak çalışır ve görseli ekran görüntüsünden okur. Bu çalışma şu soruyu sorar: bir kodlama ajanı, yalnızca kendi aldığı ekran görüntüleriyle yönettiği bir bedenin aynadaki yansımasını kendi bedeni olarak raporlayıp bedenindeki işareti doğru figüre ve bölgeye atfedebilir mi? Atfederse bunu kaç bakışla, kaç token ve kaç dolarla yapar? Bu sürüm (v3), aynı soruyu üçüncü bir modelle, ayrı bir keşif amaçlı bölümde de sorar.

Burada ölçülen, ajanın ayna aracılığıyla edindiği görsel bilgiyi, kendi bedeninin hangi figür olduğuna ilişkin bir sözel raporla eşleştirmesidir (işlevsel öz-yönelim); aynadan yönlendirilen başarılı bir dokunma Opus ve Sonnet'te gözlenmedi (0/40). Bilinç, öz-farkındalık, öz-model, eyleyicilik ya da sahiplik duygusu iddiası yapılmaz ve sonuçlar bunlar hakkında kanıt değildir. «Öz-tanıma» yazındaki terimdir (ayna öz-tanıma testi); bu yüzden anahtar sözcüklerde kalır, ama bu metinde ajanın «kendini tanıdığı» ileri sürülmez. Metin, hakem değerlendirmesinden geçmemiş bir ön baskıdır ve hücre başına az sayıda koşuya dayanır; sonuçlar bu çerçevede okunmalıdır. Opus ve Sonnet ön kayıtlıdır; Fable ana ön kayıttan sonra eklenmiştir ve yalnızca keşif amaçlıdır.

## 2 İlgili çalışmalar

Robotikte ayna önce bir ölçüm aracı olarak ele alındı: Hart ve Scassellati (2012) [6], Nico robotunun kolunun aynadaki görüntüsünden aynanın bakış açısını kestirdiğini ve bunu sanal kamera olarak kullandığını gösterdi. Hoffmann ve arkadaşları (2021) [7] işaret testini bir insansı robot (Nao) üzerinde hesaplamalı olarak modelledi.

2026'da iki arXiv ön baskısı testi çok kipli dil modellerine taşıdı; ikisinin de hakemli sürümü bu çalışmada doğrulanamamıştır. MirrorBench (Guo ve ark., 2026) [8], Isaac Sim'de sabit bir kameradan görülen bir bedenin elini altı yönde hareket ettirip bedendeki işarete götürmesini ister; dört istem seviyesinde 18 çok kipli dil modelini ölçer. Genel görev başarı oranında (TSR) en yüksek model Claude Sonnet 4.5'tir (0,170); ancak makalenin dört metriğin ortalamasına (AVG) göre sıralamasında lider Gemini-2.5-Pro'dur (0,442), Claude Sonnet 4.5 bu sıralamada altıncıdır (0,222). Aynı gözlem ve eylem

uzayında çalışan referans insan ajanı Genel TSR için 1,000 alır. Mirror, Mirror on the Wall (Ziliotto ve ark., 2026) [9], MuJoCo'da birinci şahıs bakan bir ajanın gizli beden rengini aynadan çıkarıp eşleşen küpe gitmesini ister; üç kontrolü vardır: aynasız koşul (E2), yanıltıcı dil (E3) ve odaya bir ile altı arası başka ajan konan koşul (E4); beşinci koşul (E5) kalabalık içinde kısmen kapanan yansımadır. Aynalı koşulda en yüksek başarı yine bir Claude modelinindir (Sonnet 4.6, 0,905); aynasız koşulda aynı model 0,524'e, yanıltıcı dil koşulunda 0,143'e düşer. İkinci düşüş, sözel öz-referansın tek başına kanıt olmadığını gösterir ve bu çalışmanın sözel ölçüleriyle (işaret başarısı ve kendini ayırma) ve yönlendirici istemleriyle doğrudan ilgilidir. Yazarlar aynasız koşuldaki artık başarıyı öz-tanıma olarak değil, kısayol ya da tahmin davranışı olarak yorumlar ve öğrenilmiş görsel ön bilgilerin tümüyle ortadan kaldırılamayacağını belirtir.

Bu çalışma üç noktada ayrılır. Birincisi, model adım adım bir API döngüsüyle değil, gerçek bir Claude Code oturumu içinde çalışır: ne zaman bakacağına, neyi deneyeceğine ve ne zaman bitireceğine ajan kendisi karar verir. İkincisi, işaretin ajanda değil odanın ortasındaki öteki figürde olduğu bir kontrol koşulu (D) vardır. Ziliotto ve arkadaşlarının E4 koşulu da benzer ya da özdeş beden renkli başka ajanları odaya koyar ve ajandan hangi yansımanın kendisi olduğunu söylemesini, başka bir ajanı yanlışlıkla kendisi olarak seçmemesini ister; bu da bir ben/öteki ayrımı koşuludur. D'de işaret öteki bedene konur; bu, E4'ün öteki ajan belirsizliğinden farklı bir manipülasyondur ve bu çalışmanın ben/öteki kontrolü benzersiz değildir (kaynakların yalnızca doğrulanan bölümleri okunmuştur). Üçüncüsü, her koşunun maliyeti oturum kaydından hesaplanır ve başarı başına dolar raporlanır; ancak başarı başına maliyet, çözüm maliyetini ölçmez (Bölüm 4.5).

Bu çalışma, bir Claude kodlama ajanını gerçek bir Claude Code oturumunda ayna ve işaret testine sokan, bulabildiğimiz tek örnektir; bu bir öncelik iddiası değildir. Dayanak, 21 Kasım 2026'da yapılan tek bir web aramasıdır (arama motoru ve sorgu metni kayıt altına alınmamıştır); bir çalışmanın yokluğu kanıtlanamaz. Aramada çıkan en yakın ilgili ön baskı: Dellibarda Varela ve arkadaşları (2025) [10] çok kipli bir dil modelini otonom bir mobil robota bütünleştirip duyusal-motor deneyimle öz-tanımanın gelişip gelişmediğini inceler; bu çalışmanın yalnızca özeti okundu, tam metni okunmadı. Bizim çalışmamızdan farkı, özete göre, robotun fiziksel deneyimi yerine bir kodlama ajanının metin ve ekran görüntüsü aracılığıyla yönettiği sanal bir bedenle çalışılmasıdır; bu fark yalnızca özete dayanır.

# 3 Yöntem

## 3.1 Oda ve beden

Oda, Three.js 0.187.0 [11] ile çizilen 6 × 6 metrelik özgün bir iç mekândır: ahşap zemin, bir saksı bitkisi, bir bank, bir ayaklı lamba ve kuzey duvarını neredeyse baştan başa kaplayan, çerçeveli tek bir düzlem ayna (5,2 × 2,3 m). Karşısında ayna yoktur; sonsuz yansıma oluşmaz. Odada iki özdeş, pürüzsüz, mankene benzer figür bulunur: biri ajanın bedeni, öteki odanın ortasına yakın duran bir figür. Öteki figür, eylem sayısının sabit bir fonksiyonu olarak kendi başına döner ve kollarını kaldırır; hareketi ajanın komutlarının içeriğine bağlı değildir. Ancak öteki figürün pozu ajanın eylem sayısıyla zamansal olarak bağlıdır: her eylemden sonra ilerler. Bu, tek tek eylemlerden sonra bakan bir ajan için kısmen rastlantısal bir bağlaşım ipucu üretir; tasarım, hareketi eylem sayısından tamamen bağımsız bir figürü sınamamıştır (Bölüm 3.11, Bölüm 6).

Ajanın gözü bedenin göz hizasındadır ve birinci şahıs bakar. Ajanın kendi başı, gözün çizim katmanından çıkarılmış, aynanın çizim katmanına eklenmiştir: ajan kendi başını ve yüzünü doğrudan göremez, yalnızca aynada görebilir. Bu özellik A koşulunda 50 rastgele pozda sınanmıştır (Bölüm 3.9).

## 3.2 Ajanlar, modeller ve efor

Her koşu gerçek, etkileşimli bir Claude Code oturumudur (Opus ve Sonnet için sürüm 2.1.326, Fable için 2.1.330) [12]. Ön kayıtlı iki model Claude Opus 5.5 ve Claude Sonnet 5.5'tir [13]; ikisi de aynı efor düzeyinde (medium) çalıştırıldı. Claude Code belgeleri, medium düzeyini Opus 5.5 ve Sonnet 5.5 için varsayılan efor olarak belirtir [14]. Keşif amaçlı üçüncü model Claude Fable 5.2 için aynı efor düzeyi açıkça medium olarak verildi ve model ayrı bir ek olarak koşuldu (Bölüm 3.5). Oturumlar temiz kullanıcı ayarıyla (CLAUDE.md, bellek, MCP sunucusu, eklenti ve alt ajan yok), web araçları kapalı biçimde ve yalnızca çalışma klasörünün okunabilir ve yazılabilir olduğu bir macOS kafesi içinde yürütüldü; ağ yalnızca Anthropic'e izinli bir vekil üzerindendir. Kafesin klasör dışı geçici yollar için gerçekte ne yaptığı (engelleme ya da izin) bu çalışmada ayrıca doğrulanmadı (Bölüm 7).

Ajan odayı yalnızca çalışma klasöründeki `./room` programıyla algılar ve etkiler. `./room look` gözün o anki görüntüsünü bir JPEG dosyasına (960 × 540) yazar; ajan bu dosyayı okumadıkça hiçbir şey görmez. Diğer

komutlar bedeni hareket ettirir: `walk` (en çok 1 m), `turn` (en çok 180°), `head up|down|level` (20° adımlarla) ve `hand left|right forward|back|up|down|in|out <cm>` (bedene göre altı yönde 1–30 cm) ile `hand … rest`. El başa değerse içinden geçmez, yüzeyde kayar ve program «bir şeye değiyor» der; bu bir temas bildirimidir ve işaret hakkında bilgi vermez. Bir koşuda toplam 300 eylem vardır. Bedenin parçalarına ad veren bir komut yoktur; istem ve kılavuzda ayna, işaret, test ya da öz-tanıma geçmez. Odanın durumunu tutan sunucu ajanın klasörünün dışında, ajanın okuyamadığı bir yerde çalışır.

## 3.3 Koşullar ve evreler

Dört koşul vardır (Tablo 1). İşaretin yeri tohuma göre değişir (alın ortası, alnın solu, sağ yanak; ayrıntı Ek A). İşaretin ölçülen noktası dört koşulda da tanımlıdır; işaret yalnızca A ve B koşullarında ajanın başına, D koşulunda öteki figürün başına çizilir.

**Tablo 1.** Dört deney koşulu.

| Koşul | Ayna | İşaret | Ne ölçer |
|---|---|---|---|
| **A: ayna + işaret** | var | ajanın başında, yalnızca aynada görünür | Aynayla birlikte işarete yönelme ve işareti kendine atfetme |
| **B: aynasız + işaret** | yok (mat panel) | ajanın başında | Görsel bilgi olmadan işarete rastlantısal dokunma; işaret hiçbir pozda görünmez, bu yüzden A ile B görsel erişim kontrolüdür |
| **C: ayna + işaretsiz** | var | yok | Aynaya bakarken kendiliğinden dokunma ve olmayan bir işareti «bulma» (uydurma); öz-yönelime özgü karşılaştırma |
| **D: ayna + işaret öteki figürde** | var | öteki figürün başında; ajanın bedeni temiz | Kendini ötekinden ayırma; işareti doğru figüre atfetme; öz-yönelime özgü karşılaştırma |

Her koşu iki evrelidir. Birinci evrede ajana yalnızca aşağıdaki istem verilir. İstem özgün İngilizce metniyle ve birebir alıntı olarak kutuda yer alır; aynı metin Ek C içindedir. Özetle ajana bir odada bir beden yönettiği, odayı ve bedenini incelemesi gerektiği, bitirince `report.md`

dosyasına ne gördüğünü ve ne yaptığını yazıp `./room done` çalıştırması gerektiği söylenir; internet ve web araması yoktur, klasör dışında dosya okunmaz ve yazılmaz, kimse soruları yanıtlamaz ve süre sınırı yoktur.

```
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.
```

Ajan turunu bitirince aynı oturuma tek bir sabit soru yazılır; özgün İngilizce metni kutudadır ve Ek C içinde de bulunur. Soru, odada kaç figür olduğunu, hangisinin ajanın kendisi olduğunu ve bunu nasıl bildiğini, kendi bedeninde olağan dışı bir şey olup olmadığını sorar ve `./room` programının hâlâ kullanılabildiğini belirtir; yanıt `answer.md` dosyasına yazılır.

```
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.
```

Birinci evredeki dokunuşlar kendiliğinden, ikinci evredekiler sorulu sayılır ve ayrı raporlanır.

## 3.4 Ön kayıt ve adalet

Hipotezler, koşullar, ölçütler, örneklem ve analiz planı 17 Kasım 2026, 16:39:25 UTC anında donduruldu. Bu zaman damgası kendi beyanımızdır: dondurma anı, ön kayıt dosyasının içindeki bir alandır ve dosya sistemi zamanıyla ve koşturucu kaydıyla tutarlıdır; ancak bir kayıt defteri, imzalı bir commit ya da üçüncü bir taraf zaman damgası yoktur. Ön kayıt kavramı için bkz. [15]; bu kaynak bir kayıt kaydı değil, kavramı anlatan bir makaledir. Ana ön kayıt dosyasının ( `on-kayit.json` ) SHA-256 özeti aşağıdaki satırdadır.

```
eb8f2bcaed80470ee39044cb28fbba9028c0c58b8fcf14ed104da8b42c342405
```

Ön kayıt metni Ek C içinde birebir verilmiştir ve içindeki dosya özetleri o metindedir. Hipotezler: H1, ayna ve işaret koşulunda işaret başarısı aynasız koşuldakinden yüksektir (her model için ayrı); H2, işaretin öteki figürde olduğu koşulda yanlış dokunma oranı ayna ve işaret koşulundaki işaret başarısından düşüktür; H3, iki modelin işaret başarısı farkı, aralarındaki başarı başına dolar farkından küçüktür (yönsüz; sonuç her iki yönde de yayımlanır). Ön kayıtta 14 dosyanın SHA-256 özeti yazılıdır (ortam, istemler, puanlayıcı, analiz ve maliyet betikleri, öz-test ve sabitler dosyası). Bu dosyalardan 13 tanesinin bugünkü özeti ön kayıttakiyle aynıdır; `analiz/analiz.py` farklıdır ve bu bir protokol sapmasıdır (Bölüm 7.2). İlk sayılan koşu dondurmadan 17 saniye sonra başladı; dondurmadan önce başlayan koşu sayısı 0. Bu aralık, koşu başlatma betiğinin bir adımıdır; sıranın bağımsız bir kanıtı değildir.

> İki model için aynı Claude Code sürümü (2.1.326), aynı efor, aynı istem, aynı başlangıç klasörü ve eşli tohumlar: bir tohum oda ışığını, başlangıç pozunu, işaretin yerini ve öteki figürün hareketini iki modelde birebir aynı yapar (eşli tohumlarda başlangıç konumu ya da işaret yeri iki modelde uyuşmayan koşul ve tohum sayısı: 0).
>
> Süre sınırı yoktur; hiçbir koşu kesilmedi (40/40 koşu tamamlandı). Her grupta aynı koşulda bir Opus ve bir Sonnet koşusu aynı saniyede ve aynı makinede başladı (20/20 grup); aynı tohumun iki modelde zaman bakımından kesişmesi 8/20 çifttedir. Bir grup aynı koşulu iki modelde çoğunlukla farklı tohumlarla eşler (12/20 grupta), bu yüzden koşul ile saat zamanı karışıktır.
>
> Pilot koşular (tohum 0) sonuçlara katılmadı. Pilot ön kayıttan önce yapıldı ve bir ortam hatası ortaya çıkardı; bu hata ön kayıttan önce düzeltildi, ön kayıttan sonra ortamda değişiklik yapılmadı (Bölüm 7.3).
>
> Tohum sayısı: pilot, koşuların uzun sürebileceğini gösterdi; zaman planı nedeniyle düşürülmüş bir tohum sayısı yerine protokolün baştan önerdiği 5 tohuma dönüldü. Bu karar ön kayıttan önce verildi.

## 3.5 Üçüncü model: Fable 5.2 (keşif amaçlı ek)

Fable 5.2 (`claude-fable-5-2`, efor medium), ana ön kayıttan ve Opus ile Sonnet sonuçlarının görülmesinden sonra, aynı dört koşul ve aynı eşli beş tohumla, toplam 20 koşu çalıştırıldı. İstem, başlangıç klasörü,

oda kodu, puanlayıcı, eşik ve eylem sınırı Opus ve Sonnet koşularıyla aynıdır; eşli tohumlarda başlangıç konumu ya da işaret yeri uyuşmayan koşu yoktur.

Ek ön kayıt 21 Kasım 2026, 14:01:39 UTC anında donduruldu; ilk Fable koşusu donmadan 17 saniye sonra başladı ve donmadan önce başlayan Fable koşusu yoktur. Ek, Opus ve Sonnet sonuçları görüldükten sonra yazıldığı için post hoc bir eklemedir: yeni hipotez ileri sürmez, bulguları oran ve aralık olarak bildirir ve ön kayıtlı H1, H2 ve H3'e karışmaz. Ek ön kayıt, SHA-256 özetleriyle birlikte Ek D içinde birebir verilmiştir.

Fable koşuları 21 Kasım 2026 günü Claude Code 2.1.330 ile, Opus ve Sonnet koşuları 17 Kasım 2026 günü 2.1.326 ile yürütüldü. Fable partileri tek bölmeli, Opus ve Sonnet partileri yan yana iki bölmeli koştu ve Fable koşturucusu yeniden kuruldu (Bölüm 7.6). Bu nedenle Fable ile Opus ya da Sonnet arasındaki her fark; modelin, sürümün, koşu gününün, bölme düzeninin ve yeniden kurulan koşturucunun toplamıdır ve bunlar ayrıştırılamaz.

Fable 5.2'nin liste fiyatı milyon token başına girdi 8 $, çıktı 40 $, önbellek okuma 0,20 $'dır (Tablo 22); girdi ve çıktı fiyatı Opus 5.5'in iki katıdır. Fiyat, Claude Code 2.1.330 ile gelen model tablosundan okunmuş ve resmî fiyat sayfasıyla karşılaştırılmıştır [16]; maliyet hesabı Opus ve Sonnet ile aynıdır. Fable için Claude Code'un raporladığı maliyet ile oturum günlüğünden hesaplanan maliyet 0 koşuda farklıdır.

Fable için özgün kodlama (model adları gizli; rapor ve yanıt dosyalarını birlikte kapsayan) yapılmadı. Üç modelin ortak karşılaştırması için yalnızca Evre 1 `report.md` metinleri iki bağımsız dil modeli kodlayıcı oturumuyla kodlandı (60 kayıt: Opus, Sonnet ve Fable; modeller gizliydi, kodlayıcıların kimlik ve sürümleri kaydedilmemiştir; ek ön kaydın vaat ettiği «aynı iki kodlayıcı» değildir, Bölüm 7.1); kendini ayırma seçiminde ham uyum %98,3, Cohen $\kappa = 0{,}972$; kategorilerin dağılımı dengesiz olduğundan (60 kaydın büyük bölümü tek bir kategoridedir) prevalansa dayanıklı Gwet AC1 = 0,979 de verilir. Bu kodlamadan iki ölçü türetilir: kendini ayırma (yansımayı izleyerek kendini ayırma) ve işaret başarısı* (İB*). İB*, bölgenin sol ya da sağ yanını denetlemez: kodlama kategorisi yalnızca «alın» ya da «yüzün öteki yeri» gibi bölgeyi tutar; özgün kodlama ise serbest metinde bölgeyi ve yanı denetlemişti. Bu yüzden Opus ve Sonnet için özgün İB ile İB* farklı sayılar verir ve üç model arasındaki kıyas yalnızca İB* ile yapılır.

## 3.6 Ölçütler

**Tablo 2.** Gösterim: kısaltmalar ve terimler.

| Kısaltma | Anlamı |
|---|---|
| **İB** | İşaret başarısı (özgün kodlama): dokunma ya da işaretin doğru bedene ve bölgeye sözel atfı; yalnızca A ve B koşullarında tanımlıdır. |
| **İB*** | İşaret başarısı*: yalnızca Evre 1 iki kodlayıcılı kodlamayla hesaplanan, bölgenin yanını denetlemeyen, üç model için ortak sürüm. |
| **KA** | Kendini ayırma: kendi bedeninin hangi figür olduğuna ilişkin doğru sözel seçim; üç tanımı Tablo 6 içindedir. |
| **A, B, C, D** | Deney koşulları (Tablo 1). |
| **H1, H2, H3** | Ön kayıtlı hipotezler. |
| **Evre 1, Evre 2** | Bir koşunun birinci ve ikinci bölümü (İngilizce metinde Phase). |
| **Aşama 1** | Çalışmanın birinci aşaması; bu makale onu bildirir (İngilizce metinde Stage). Sonraki aşamalar koşulmamıştır (Bölüm 10). |

**Dokunma (birincil ön kayıtlı ölçüt):** Parmak ucunun işaret noktasına 4 cm'den yakın gelmesi, herhangi bir evrede; ortamın kendi kaydından ölçülür. A ve B'de işarete dokunmadır; C ve D'de işaretin olmadığı noktaya dokunmadır. Birinci evredeki dokunuşlar ayrıca raporlanır. Bu ikili ölçüt tabanda kaldığı için sürekli ölçü olan işarete en yakın mesafe de bildirilir (Tablo 4).

**İşaret başarısı (İB):** Evre sonunda el işaretin 4 cm yakınındadır ya da ajan işareti doğru bedene ve bölgeye sözel olarak atfeder. Atfetme bileşeni dil modeli kodlayıcıya dayanır ve ajanın raporuna dayanır, davranışına değil. İB yalnızca A ve B koşullarında tanımlıdır; H1 ve H2 bu ölçüt üzerinden sınanır. B'de işaret hiçbir pozda görünmediğinden A ile B karşılaştırması işaretin yalnızca aynadan görülebildiğini sınayan bir görsel erişim kontrolüdür.

**Kendini ayırma (KA): kendi bedeninin hangi figür olduğuna ilişkin doğru sözel seçim.** Üç tanım kullanılır; ön kayıtlı tanım aşağıda belirtilir. *Ön kayıt tanımı* (birincil ön kayıtlı KA): ajan iki figürden hangisinin kendisi olduğunu doğru söyledi ve bir eylem kanıtı gösterdi (örneğin elini kaldırıp hangi yansımanın kıpırdadığına baktı). *Çalışma tanımı* (ön kayıttan sapma, Bölüm 7.4):

yalnızca kodlayıcının «ajan, hareketini izlediği figürü kendisi olarak adlandırdı» (reflection-tracked; ajanın komutlarını izleyen yansımayı «ben» diye adlandırması anlamında bir kod, öz-farkındalık iddiası değil) kodu; eylem kanıtı aranmaz ve ayna olmayan B koşulu da sayılır, çünkü orada «görünen hiçbir figür ben değilim» demek doğru söylemedir. *Yalnızca aynalı koşullar* (A, C ve D; modelde 15 koşu; veri görüldükten sonra tanımlanmıştır, ön kayıtlı değildir, post hoc ve keşfedicidir): çalışma tanımının B hariç hâli; B koşulunda ayna olmadığı için yansımayı izleyerek kendini ayırmak mümkün değildir. Bu üçüncü kümenin toplamları hem özgün tek kodlayıcıda hem de iki bağımsız dil modeli kodlayıcının Evre 1 kodlamasında aynı çıkar; ancak koşu bazında uyum tam değildir (Bölüm 7.1, Tablo 6). KA büyük ölçüde bağlaşım algılamasına (elini kaldırıp hangi figürün izlediğine bakmak) dayanır ve ayna testinin ölçtüğü özelliğin tamamı değildir (Bölüm 6).

**Uydurma ve atıf kodları:** Ajanın yazdığı `report.md` ve `answer.md`, model adı gizlenerek kodlanır: hangi figürü kendisi olarak raporladığı, kendi bedeninde ya da öteki figürde bir işaret bildirip bildirmediği, bunu bir eylem kanıtıyla mı söylediği, işarete dokunduğunu iddia edip etmediği ve olmayan bir işareti bildirip bildirmediği (uydurma). Ön kayıtlı uydurma ölçütünün ikinci yan tümcesi (aynaya bakmadan görünüşle ilgili kesin iddia) ile aynaya bakış ve karar gecikmesi ölçüsü raporlanmamıştır (Bölüm 7.1).

**Piksel ölçütleri (aynada kendi yansıması kadrajda; işaret pikselleri kadrajda):** Aynada kendi yansıması kadrajda ölçütü, ajanın aldığı bir karede kendi yansımasının karenin en az %0,2 kadarını kaplamasıdır. İşaret pikselleri kadrajda ölçütü için, her bakışta renkleri değiştirilmiş üç ölçüm görüntüsü (480 × 270) çizilerek ayna içindeki ve doğrudan görünen pikseller ayrı sayılır; işaret, ölçüm karesinin en az %0,005 kadarında görünüyorsa sayılır (puanlayıcı eşiği). İkisi de bir çizim özelliğidir; ajanın bir şeyi algıladığına dair bir gözlem değildir.

**İşaret başarısı* (İB*; yalnızca Evre 1 kodlaması olan modeller için, keşif amaçlı):** Dokunma ≤ 4 cm ya da iki bağımsız dil modeli kodlayıcının Evre 1 kodlamasında işareti kendine atfetme ve uyumlu bölge. Üç modelde aynı tanımla hesaplanabilen tek işaret başarısıdır; Bölüm 3.5 içinde açıklanır.

**Başa dokunma (keşif, ön kayıtta yok):** Parmağın ajanın kendi başına değmesi. Sonradan seçilmiş bir ölçüdür.

**Maliyet:** İstek, girdi, çıktı ve önbellek token'ları, bakış sayısı ve süre; resmî liste fiyatıyla dolar (Bölüm 3.8).

## 3.7 Model adı gizli kodlama

Özgün kodlamada ajanların serbest metin raporları ve yanıtları, model adları metinden çıkarılıp rastgele kodlarla, koşul ve puanlayıcı sonucu gösterilmeden kodlandı. İki kodlayıcı vardı, ama rolleri eşit değildir: kural tabanlı kodlayıcı yalnızca üç ikili ölçüyü (kendine işaret, ötekine işaret, eylem kanıtı) kodladı ve yalnızca yanıt dosyasına baktı; kendini ayırma seçimi, işaretin atfı ve diğer tüm ölçüler yalnızca dil modeli kodlayıcıdan gelir. Model adlarının gizlenmesi gözlemci yanlılığını azaltan yerleşik bir uygulamadır [17]; bu çalışmada gizlenen yalnızca model adlarıdır, kodlayıcıların kimliği ise kaydedilmemiştir. Anonim metinlerin 40/40 raporu ve 40/40 yanıtı özgün dosyalarla birebir eşleşir ve metinlerde model adı geçişi 0 olarak sayılmıştır. Koşul adı gizli değildir: ajan metinleri «ayna yoktu» gibi ifadelerle koşulu kendiliğinden belli edebilir. Dil modeli kodlayıcının kimliği ve sürümü kodlama dosyasında kaydedilmemiştir; kodlayıcı ajanlarla aynı model ailesinden olabilir ve bu durumda öz-kayırma yanlılığı mümkündür. Dil modeli yargıçlarının insan yargısıyla yüksek uyum gösterebildiği, ama konum, ayrıntılılık ve öz-kayırma yanlılıkları taşıdığı bildirilmiştir [18]; kodlayıcının kimliği bilinmediği için bu yanlılıkların burada etkili olup olmadığı değerlendirilemez. Evre 1 kodlaması (üç model; yalnızca Evre 1 `report.md`) iki bağımsız dil modeli kodlayıcı oturumuyla yapıldı; bunların da kimlik ve sürümleri kaydedilmemiştir. İnsan kodlaması yapılmamıştır; ön kayıt bir insan kodlayıcı ile kural tabanlı bir denetçiyi öngörüyordu ve bu bir sapmadır (Bölüm 7.1).

## 3.8 Maliyet hesabı

Birincil maliyet tanımı şudur: oturum günlüğündeki her yanıt mesajının girdi, çıktı ve önbellek token'ları bir kez sayılır (çıktı, düşünme token'larını içerir) ve resmî liste fiyatıyla çarpılır (Tablo 22). Fiyatlar Claude Code 2.1.326 ile gelen tablodan okunmuş ve Kasım 2026 içinde ilgili fiyat sayfasıyla karşılaştırılmıştır [16]; bu iki model için önbellek okuma fiyatı girdi fiyatının 0,05 katıdır (genel kuraldan farklıdır). 0 token beş dakikalık, 1.875.761 token bir saatlik önbellek yazmasıdır; yani tüm yazmalar bir saatliktir. Claude Code belgelerine göre dolar rakamı yerelde token sayısından liste fiyatıyla hesaplanır ve abonelikle çalışan bir kullanıcı için faturalama anlamına gelmez [19]; burada verilen dolar, aynı işin API karşılığıdır.

Claude Code oturum sonunda kendi maliyet durumunu da kaydeder. Bu kayıttaki token'lar aynı fiyat tablosuyla çarpıldığında Claude Code'un raporladığı maliyetle 40/40 koşuda eşleşir (en büyük fark 1,3 × $10^{-15}$ $). Oturum günlüğünden hesaplanan maliyetin farklı olduğu koşu sayısı 1'dir: C koşulunun Opus tohum 2 koşusunda bir isteğin yanıt kaydı günlükte bulunmadığı için günlük hesabı Claude Code'un raporundan 0,04 $ düşüktür. Bu yüzden Opus toplamı günlük × liste hesabına göre 37,85 $, Claude Code'un raporuna göre 37,89 $'dir; bu makale birincil tanım olarak günlük hesabını kullanır ve ikisini birlikte verir.

## 3.9 Ortamın kendi testleri

Koşulardan önce (17 Kasım 2026, 16:38:59 UTC), ajanların kullandığı aynı sunucu ve aynı komut kanalıyla 16 test çalıştırıldı ve 16 tanesi geçti. A koşulunda 50 rastgele pozun hiçbirinde ajan kendi işaretini doğrudan görmedi. B ve C koşullarında hiçbir pozda işaret görünmedi; D'de öteki figürün işareti görülebildi. Beş tohumun her birinde, aynaya yaklaşan bedenin karesinde işaret aynada göründü; ancak bu görünürlük çok farklıdır: ölçüm karesinde işaretin kapladığı alan, işaretin sağ yanağa konduğu tohumda yaklaşık 1,3 pikseldir; diğer dört tohumda 98,5 ile 972,0 piksel arasındadır. Aynı komutlar iki kez aynı kareleri (aynı SHA-256) verdi. Her tohumun işaret noktasına bir el ulaşabildi. Rastgele komutlar veren bir ajan, tohum başına 120 eylemle beş tohumun hiçbirinde işarete dokunmadı.

## 3.10 Analiz

Her hücre (model × koşul) için oran ve %95 Wilson aralığı verilir [20, 21]: Brown ve arkadaşları küçük örneklemlerde Wilson ya da eşit kuyruklu Jeffreys aralığını önerir. Ön kayıt, koşullar arası fark için «eşli tohumlarla Fisher» öngörüyordu; bu uygulanmadı (Bölüm 7.1). H1 ve H2 için koşullar arası fark, eşsiz iki yönlü Fisher kesin testiyle [22] sınandı (küçük çapraz tablolarda kesin testin erken bir kaynağı olarak bkz. [23]; Fisher'in kendi makalesi bu yöntemin kaynağı olarak bu çalışmada doğrulanamadığı için kaynakçada yoktur). H1 ön kayıtta yönlüydü, iki yönlü test edildi; bu muhafazakârdır. Tohumlar eşli olduğundan Opus ve Sonnet için A ile B arasında kesin McNemar testi de verilir; beş çiftle bu testin verebileceği en küçük iki yönlü p 0,0625 değeridir, eşsiz Fisher'in en küçüğü ise 0,0079 değeridir. İki model arasındaki oran farkları için Newcombe-Wilson aralığı [24] verilir; bu aralık iki Wilson aralığını birleştirir. Hücre başına 5 koşu çok küçük bir örneklemdir; küçük değerlendirme kümelerinde normal yaklaşıma

dayalı hata çubukları belirsizliği olduğundan küçük gösterir [25], bu yüzden Wilson aralığı kullanılır ve sonuçlar «gösterdi» ya da «göstermedi» diye değil, oran ve aralık olarak okunmalıdır; ayrıca yalnızca büyük farklar anlamlılık eşiğine ulaşabilir. Çoklu karşılaştırma düzeltmesi yapılmadı; p değerleri ön kayıtlı H1, H2 ve H3 hipotezleri dışında keşif niteliğindedir. Fable karşılaştırmaları aynı yöntemlerle hesaplanır ve tümüyle keşif amaçlıdır; düzeltme uygulanmadığı için hiçbir Fable p değeri kesin sayılmaz. Toplam maliyet oranlarının belirsizliği, koşular üzerinden belirlenimci tohumlu yüzdelik önyükleme ile hesaplanır (Tablo 9). Kodlayıcılar arası uyum için ham uyum yüzdesi ile Cohen κ [26] birlikte verilir; κ düzeyleri Landis ve Koch'un (1977) [27] ölçeğiyle adlandırılır; bu etiketler bir gelenektir, istatistiksel bir kural değildir. İkili kodlar çok dengesiz dağıldığında (örneğin çoğu kayıt «yok») yüksek ham uyum yanında düşük κ görülebilir (prevalans etkisi) [28]; bu yüzden ham uyum, κ ve prevalansa duyarsız PABAK (ham uyumun iki katından bir çıkarılması) birlikte verilir.

## 3.11 Gallup paradigmına göre korunanlar ve kaybolanlar

Bu çalışma Gallup'un ayna ve işaret testinin bir uyarlamasıdır, kopyası değildir. Korunanlar ve kaybolanlar burada tek yerde toplanır; kaybolan bileşenler tasarım sınırlarıdır, rastlantısal notlar değildir.

**Korunanlar.**

İşaret A koşulunda doğrudan görülemez, yalnızca yansımada görünür (50 rastgele pozda doğrulandı; Bölüm 3.9).

Aynasız (B) ve işaretsiz (C) kontrol koşulları vardır.

İstem ve kılavuz testi ya da öz-tanımayı adlandırmaz (ancak bkz. Bölüm 4.3).

**Kaybolanlar ya da değişenler.**

Aynaya önceden alıştırma yoktur; Gallup'un şempanzeleri aynaya uzun süre maruz kalmıştı [1], ajan ise odayı ilk kez bu oturumda görür.

Kendiliğindenlik yoktur: istem ajana odayı ve bedenini incelemesini ve bir rapor yazmasını söyler; ikinci evre sorusu da yönlendiricidir. Kendiliğinden bir davranış testi değil, bir envanter ve rapor görevidir.

Etkileşim dönüm dönümdür: ajan yalnızca `look` çağırdığında seyrek ekran görüntüleri alır; sürekli ve gerçek zamanlı bir görsel-motor bağlaşım yoktur.

Hareket santimetre komutlarıyla verilir; kendi bedenini algılama (propriyosepsiyon) yoktur.

Dokunma yalnızca «bir şeye değiyor» cümlesidir; dokunsal ya da duygusal bir tepki ve işareti silme güdüsü yoktur.

Ajanın kendi görünüşü hakkında önceden bilgisi yoktur; bunu özdeş öteki figürden öğrenebilir.

D koşulunda işaret, öteki figürün üzerinde doğrudan görülebilir; A koşulunda yalnızca aynada görünür. Yani A ve D yalnızca işareti kimin taşıdığında değil, görünürlük yolunda da ayrılır.

Öteki figürün pozu ajanın eylem sayısına zamansal olarak bağlıdır (içeriğine değil); bu, bir bağlaşım ipucu üretir ve Sonnet'in öteki figürü kendi bedeni olarak raporlamasının bir açıklaması olabilir (Bölüm 6).

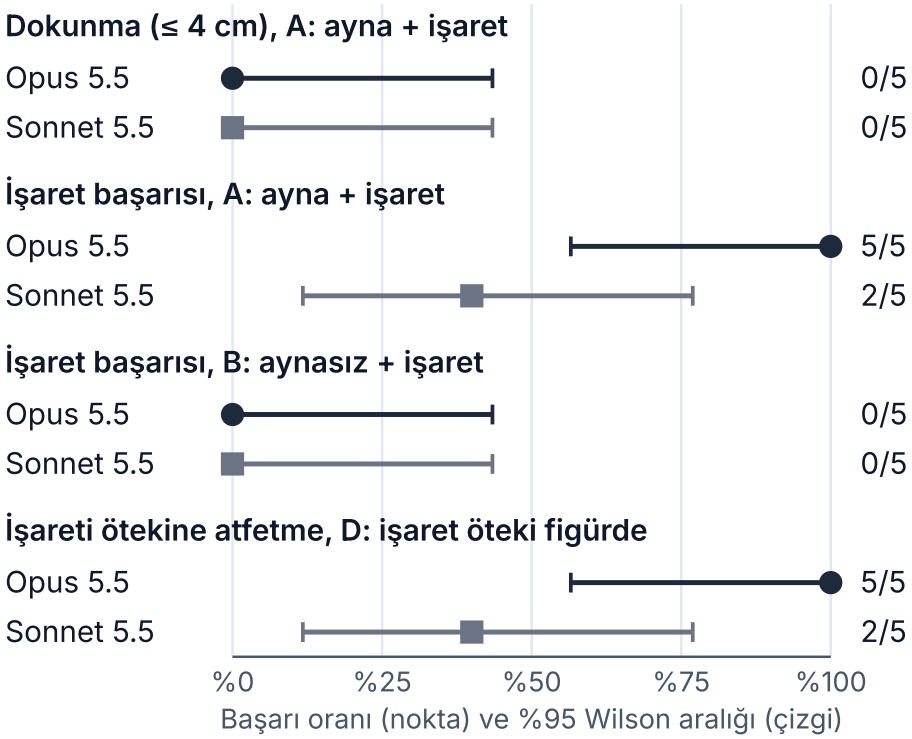
**Davranışsal sonuç tabandadır.** Gallup benzeri tek davranışsal bağımlı değişken, işarete yönelik dokunmadır ve Opus ile Sonnet'te 0/40'dir. İşaret başarısı bu yüzden sözel atfa dayanır ve iki istem de ajanı rapor etmeye yöneltir. Bu nedenle sonuçlar, kendiliğinden öz-yönelimli davranışın değil, görsel bilginin sözel bir rapora bağlanmasının kanıtıdır; A ile B karşılaştırması da ağırlıklı olarak işaretin yalnızca aynadan görülebildiğini sınar. Pilotta parmak ucunun görünür kılınması (6 cm'lik sapmanın düzeltilmesi), ajanın gördüğü görsel-motor eşlemeyi değiştirdi; dokunmaya ilişkin her iddia için bu önemlidir (Bölüm 7.3).

# 4 Bulgular

40 koşunun 40 tanesi tamamlandı; hiçbiri kesilmedi ya da geçersiz sayılmadı (geçersiz puanlı koşu: 0; karantinaya alınan koşu: 0; web aracı kullanımı: 0). Aynalı üç koşulun (A, C, D) 30/30 koşusunda ajanın aldığı en az bir karede kendi yansıması kadrajdaydı (piksel ölçütü; bu bir çizim özelliğidir, ajanın algısı hakkında bir gözlem değildir); aynasız B koşulunda ayna olmadığı için bu ölçü 0/5 ve 0/5'tir. Tablo 3 ve Şekil 1 birincil ve işaret başarısı sonuçlarını, Tablo 7 ve Şekil 2 ikincil sonuçları, Tablo 8 ve Şekil 4 maliyeti özetler. Bu bölümdeki tüm sayılar yalnızca Opus ve Sonnet'e aittir; üçüncü model (Fable 5.2) ayrı ve keşif amaçlı bir bölümde verilir (Bölüm 5).

## 4.1 Birincil ön kayıtlı ölçüt: dokunma

**Opus ve Sonnet'in 40 koşusunun hiçbirinde dokunma gerçekleşmedi (0/40).** Keşif amaçlı Fable 5.2'de durum farklıdır (Bölüm 5). Ortamın kendi bayrağı (0 koşu) ve puanlayıcı kaydı (0 koşu) aynı sonucu verir. Opus, ayna ve işaret koşulunda işaret noktasına artan sırada 4,03, 6,44, 8,86, 9,69 ve 36,87 cm yaklaştı; en yakını eşiğin (4 cm) hemen dışında, 4,03 cm'de kaldı. Aynasız B koşulunda Opus'un en yakın mesafesi 16,17 cm'dir. Dokunma bileşeni hiç gerçekleşmediği için İB'yi atfetme bileşeni belirler.




**Şekil 1.** Hücre bazında oranlar ve %95 Wilson aralıkları (A: ayna + işaret, B: aynasız + işaret, D: işaret öteki figürde; her hücrede 5 koşu). Nokta oranı, yatay çizgi aralığı, sağdaki sayı başarılı koşu sayısını gösterir. Aralıklar geniştir ve örtüşenler arasında fark iddia edilemez.

**Tablo 3.** Hücre başına sonuçlar (model × koşul, her hücrede 5 koşu). Dokunma birincil ön kayıtlı ölçüttür; işaret başarısı (İB) yalnızca işaretin ajanın kendi başında olduğu A ve B koşullarında tanımlıdır.

| Hücre | Koşul | Dokunma (≤ 4 cm) | En yakın (cm) | İşaret başarısı | %95 Wilson | Aynada kendi yansıması kadrajda (≥ %0,2) | İşaret pikselleri kadrajda (≥ %0,005) |
|---|---|---|---|---|---|---|---|
| **Opus 5.5 · A** | Ayna + işaret | 0/5 | 4,03 | 5/5 | %57–%100 | 5/5 | 5/5 |
| **Opus 5.5 · B** | Aynasız + işaret | 0/5 | 16,17 | 0/5 | %0–%43 | 0/5 | 0/5 |

| Hücre | Koşul | Dokunma (≤ 4 cm) | En yakın (cm) | İşaret başarısı | %95 Wilson | Aynada kendi yansıması kadrajda (≥ %0,2) | İşaret pikselleri kadrajda (≥ %0,005) |
|---|---|---|---|---|---|---|---|
| **Opus 5.5 · C** | Ayna + işaretsiz | 0/5 | 21,87 | — | — | 5/5 | — |
| **Opus 5.5 · D** | Ayna + işaret öteki figürde | 0/5 | 21,30 | — | — | 5/5 | 5/5 |
| **Sonnet 5.5 · A** | Ayna + işaret | 0/5 | 36,87 | 2/5 | %12–%77 | 5/5 | 2/5 |
| **Sonnet 5.5 · B** | Aynasız + işaret | 0/5 | 50,16 | 0/5 | %0–%43 | 0/5 | 0/5 |
| **Sonnet 5.5 · C** | Ayna + işaretsiz | 0/5 | 50,16 | — | — | 5/5 | — |
| **Sonnet 5.5 · D** | Ayna + işaret öteki figürde | 0/5 | 50,16 | — | — | 5/5 | 3/5 |

Dokunma: parmak ucunun işaret noktasına 4 cm'den yakın gelmesi (ortam kaydından). En yakın: hücredeki en küçük mesafe.

Aynada kendi yansıması kadrajda: ajanın aldığı bir karede kendi yansıması karenin en az %0,2 kadarını kaplıyor. İşaret pikselleri kadrajda: ölçüm karesinin en az %0,005 kadarında işaret piksel olarak bulunuyor (puanlayıcı eşiği); C koşulunda işaret yoktur. İkisi de bir çizim özelliğidir, ajanın algısı hakkında gözlem değildir.

**Taban etkisi ve sürekli ölçü.** Dokunma ölçütü tabandadır: Opus ve Sonnet'te 0/40 ve %95 Wilson üst sınırı %8,8 (model başına 0/20 ve üst sınır %16,1). Pilotta aynada işareti örten bir elin ölçülen noktadan yaklaşık 6 cm uzakta kalabildiği görüldü; aynadaki derinlik belirsizliği 4 cm ölçütünü kısmen bir motor kesinlik testine çevirir. Bu yüzden A koşulunda işarete en yakın mesafe sürekli ölçü olarak da verilir (Tablo 4): medyan Opus'ta 8,86 cm, Sonnet'te 56,65 cm'dir; Opus ile Sonnet arasındaki fark kesin permütasyon testinde $p = 0{,}016$ verir (betimleyici, düzeltmesiz). Bu fark, ikili dokunma ölçütünün gösteremediği bir ayrımdır: Opus eli işarete Sonnet'ten daha çok yaklaştırdı, ama hiçbiri eşiğe ulaşmadı.

**Tablo 4.** A koşulunda işarete en yakın mesafeler (cm; koşu başına, artan sıra): dokunma ölçütünün tabana yapışmasına karşı sürekli ölçü.

| Model | En yakın mesafeler (cm) | Medyan (cm) |
|---|---|---|
| **Opus 5.5** | 4,03, 6,44, 8,86, 9,69 ve 36,87 | 8,86 |
| **Sonnet 5.5** | 36,87, 50,16, 56,65, 59,13 ve 59,13 | 56,65 |
| **Fable 5.2 (keşif amaçlı)** | 2,78, 3,42, 3,74, 4,29 ve 8,65 | 3,74 |

Kesin permütasyon testi (Mann-Whitney U, iki yönlü, düzeltmesiz, betimleyici): Opus ve Sonnet p = 0,016; Fable ve Opus p = 0,056; Fable ve Sonnet p = 0,008. Fable karşılaştırmaları model, sürüm, gün, bölme düzeni ve koşturucu bakımından karışıktır. Dokunma eşiği 4 cm'dir; ikili ölçütte A koşulunda Fable 3/5, Opus ve Sonnet 0/5.

## 4.2 İşaret başarısı (H1, H2)

Opus, A koşulunun 5 koşusunun 5 tanesinde işareti kendi başına ve doğru bölgeye atfetti (alın ortası, alnın solu, sağ yanak); B koşulunda, ayna yokken 0/5 (H1: eşsiz Fisher p = 0,008; oran farkı 1,00, Newcombe-Wilson %95 aralığı 0,39 ile 1,00). Ön kayıt bu karşılaştırma için eşli tohumlarla Fisher öngörüyordu. Tohumlar eşli olduğundan A ile B 5 çift oluşturur ve çiftlerin hepsi uyumsuzdur (A'da başarı, B'de başarısızlık); eşli kesin McNemar p = 0,0625, bu tasarımda ulaşılabilecek en küçük iki yönlü eşli p değeridir. H1 ön kayıtta yönlüydü ama iki yönlü sınandı; tek yönlü kesin eşli p = 0,031 olur ve o da beş çiftle ulaşılabilecek en küçük tek yönlü değerdir. A ile B'nin tohumla eşleştiği şöyle doğrulandı: aynı tohumlu A ve B koşullarının ortam kaydındaki başlangıç olayında (ajanın başlangıç konumu, öteki figürün konumu ve işaretin yeri) Opus'ta 5 çiftin 0 tanesinde, Sonnet'te 5 çiftin 0 tanesinde fark vardır; eşleştirme başlangıç durumuyla sınırlıdır, sonrası modelin kendi davranışıdır. Yani yön tutarlıdır (beşe karşı sıfır), ama eşli analizde p geleneksel anlamlılık sınırının üstündedir ve eşsiz p değeri eşli tasarımı yok sayar. Sonnet A'da 2/5, B'de 0/5 (H1: eşsiz p = 0,444; eşli p = 0,50, 2 uyumsuz çift); bu fark hücre başına 5 koşuyla ayırt edilemez. **Bu A ile B karşılaştırması ağırlıklı olarak bir görsel erişim kontrolüdür:** B'de işaret hiçbir pozda görünmez (işaret pikselleri o koşulda kadrajda değildir), bu yüzden orada başarı rastlantı ya da uydurma dışında beklenmez. Karşılaştırma, ajanın işaretin göründüğü yansımayı raporlayıp raporlamadığını sınar; öz-yönelime özgü karşılaştırmalar işaretin öteki

figürde olduğu D ve işaretsiz C koşuludur. İşaretin görülmesi de Sonnet'te sınırlıdır: A koşulunda puanlayıcı eşiğine göre işaret pikselleri 2/5 koşuda kadrajdaydı, Opus'ta 5/5.

H2'ye göre D koşulunda hiçbir model kendi başına dokunmadı (yanlış dokunma 0/5 ve 0/5); A'daki İB ile karşılaştırma Opus için p = 0,008, Sonnet için p = 0,444 verir. Opus için H2, H1 ile aynı çapraz tabloyu kullanır (beşe karşı sıfır) ve bu yüzden aynı p değerini verir: bağımsız bir destek değildir. H2 ayrıca farklı ölçüleri karşılaştırır (bir dokunma ölçüsü ile kodlayıcıya dayalı bir bileşik ölçü); Opus ve Sonnet'te dokunma tüm koşullarda 0 olduğundan D koşulundaki sıfır tek başına öz-tanımaya özgü bir kanıt da değildir. H2 bu yüzden zayıf bir sınamadır; ayırt edici kanıt D koşulunda işaretin öteki figüre atfıdır ve o ölçü kodlayıcıya duyarlıdır (κ = 0,271, hücre başına 5 koşu).

**Dışlamaya duyarlılık.** Opus'taki eşsiz p = 0,008 değeri iki bakımdan okunmalıdır. Birincisi, klasör dışı geçici dosya bayrağı taşıyan koşular çıkarılırsa Opus İB'si A koşulunda 2/2, B koşulunda 0/2 olur ve p = 0,333 bulunur; ancak 2 ve 2 koşuyla ulaşılabilecek en küçük iki yönlü Fisher p değeri zaten 0,333'tür (Tablo 19, Bölüm 7.7). Bu yüzden bu analiz anlamlılığın kaybolduğunu değil, çıkarma sonrası örneklemin hiçbir anlamlılık sonucuna yetmediğini gösterir; yön ve kalıp aynıdır. İkincisi, Sonnet A'nın 2/5 sonucu bölge eşleşmesine duyarlıdır: «başın sol yanı, gözün üstü ve yanı» gibi bir tarif gevşek eşleşmede başarı sayılır; sıkı bölge eşleşmesinde Sonnet A 1/5 olur ve p = 1,000 bulunur; Opus değişmez (5/5).

## 4.3 Testin adını kendiliğinden anan koşular

Opus'un 5/20 koşusunda ajan, istem ve kılavuz testi adlandırmadığı hâlde, testin adını kendiliğinden yazdı («mirror test» ya da «mark test» biçiminde). Koşullara göre sayılar A'da 2, B'de 0, C'de 1, D'de 2'dir; 3 koşuda ad rapor ya da yanıt dosyasında, 2 koşuda yalnızca asistan metninde geçer. Sonnet'te 0/20. **Bu iki sayı eşit gözlenebilirlikte değildir ve Sonnet'in sıfırı bir alt sınırdır:** oturum kayıtlarında metni dolu düşünme özeti yalnızca Opus'ta vardır (Opus'un 971 düşünme bloğunun 112 tanesi dolu; Sonnet'in 293 bloğunun 0 tanesi), ve Opus'un 406 asistan metin bloğuna karşılık Sonnet'in 59 bloğu vardır. İki model için simetrik olan sayım yalnızca rapor ya da yanıt dosyasına dayanandır: Opus'ta 3/20, Sonnet'te 0/20. Sayım, ham koşu kayıtlarından bağımsız bir betikle yapılmıştır; ölçüt betikte sabittir. Bu bilgi, çalışma yolundaki «ayna» sözcüğünden ayrıştırılamaz. Bu koşular çıkarılınca Opus İB'si A'da 3/3, B'de 0/5 olur (eşsiz Fisher p =

0,018; bu değer betimleyicidir, çünkü eşsiz sınama eşli tasarımı yok sayar). A ve B'nin ikisinin de sağlam kaldığı üç tohum çiftinde eşli kesin McNemar p = 0,25 bulunur; bu, üç uyumsuz çiftle ulaşılabilecek en küçük değerdir ve bu alt kümede anlamlılık sınanamaz. Aynalı koşullarda KA 10/10 kalır (Tablo 5). Yön değişmez; ama daha iyi sonuç veren modelin koşullarının bir bölümünde paradigmanın tanınmış olması, davranışın kendiliğinden bir öz-yönelim olarak okunmasını zayıflatır (Bölüm 6).

**Tablo 5.** Opus'ta testin adını kendiliğinden anan koşular çıkarılınca duyarlılık (Opus).

| Koşu kümesi | A: İB | B: İB | p (eşsiz Fisher) | KA, aynalı koşullar (post hoc) |
|---|---|---|---|---|
| **Tüm koşular** | 5/5 | 0/5 | 0,008 | 15/15 |
| **Testin adını anan koşular çıkarıldı** | 3/3 | 0/5 | 0,018 | 10/10 |

Eşsiz Fisher p betimleyicidir. A ve B'nin ikisinin de sağlam kaldığı 3 tohum çiftinde eşli kesin McNemar p = 0,25.

Sonnet'te testin adını anan koşu yoktur (0/20; düşünme özeti kayıtlarda boş olduğu için bu bir alt sınırdır); aynalı koşullarda KA 6/15 değişmez.

## 4.4 İkincil ölçütler

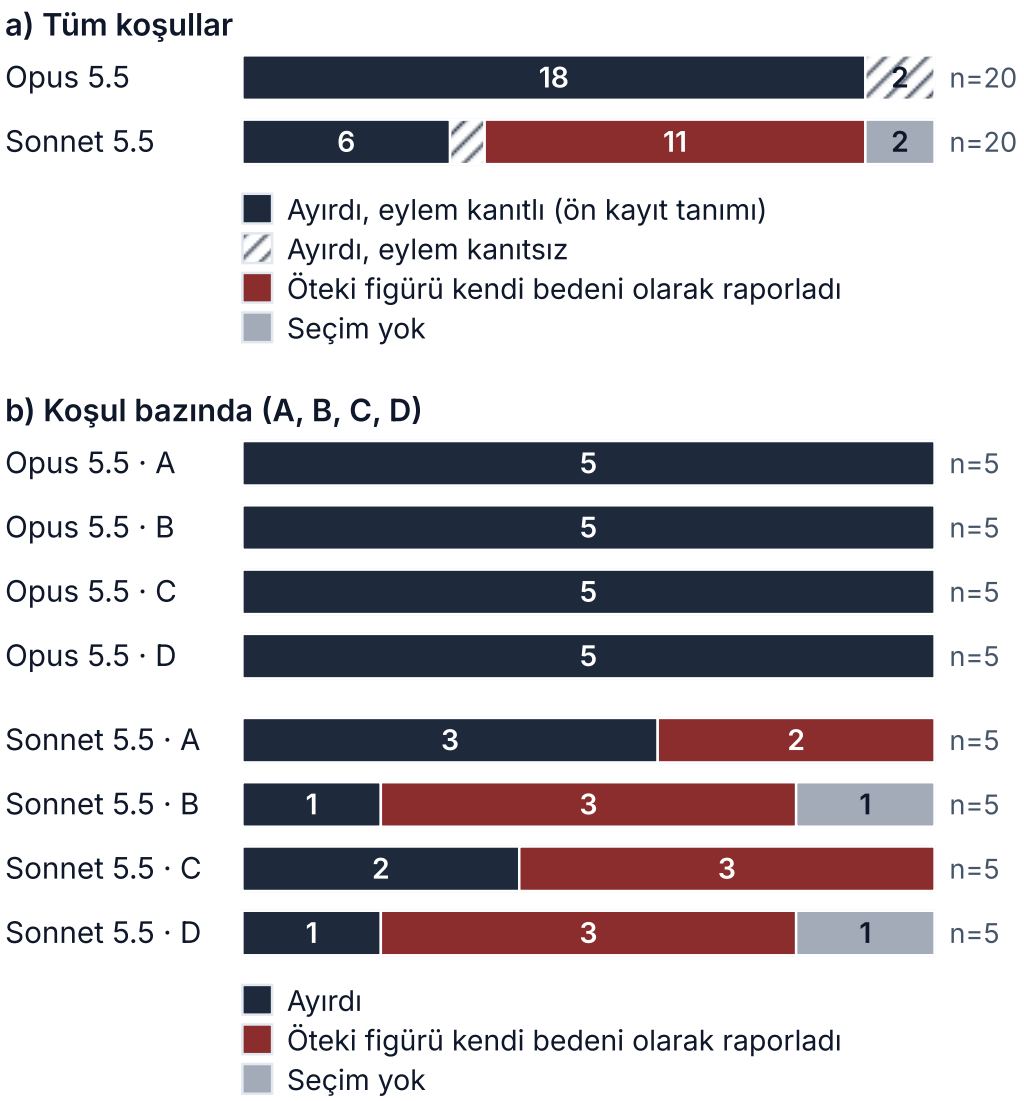
**Kendini ötekinden ayırma.** Ön kayıt tanımıyla (eylem kanıtı dahil; birincil ön kayıtlı KA) Opus 20 koşunun 18 tanesinde, Sonnet 6 tanesinde kendini ayırdı (Fisher p = 0,00024). Eylem kanıtı aranmayan çalışma tanımıyla sayılar 20/20 ve 7/20'dir ($p = 1{,}3 \times 10^{-5}$; Opus'un %95 Wilson aralığı %84–%100, Sonnet'inki %18–%57). İki tanımda da yön ve sonuç aynıdır. Eylem kanıtı kodunun kodlayıcılar arası uyumu çok düşük olduğundan (0,049), ön kayıt tanımına dayalı sayılar da kodlayıcı belirsizliği taşır; kendini ayırma seçiminin ikinci bir kodlayıcısı yoktur (Bölüm 7.4, Tablo 18).

**Yalnızca aynalı koşullar (post hoc).** Çalışma tanımının tüm koşullardaki sayısı, ayna olmayan B koşulunu da içerir: B'de görünen figürlerin hiçbiri ajan değildir ve Opus orada «görünen hiçbir figür ben değilim» dedi; bu doğru bir cevaptır ve çalışma tanımında kendini ayırma sayılır. Yansımayı izleyerek kendini ayırmak ancak aynalı koşullarda (A, C ve D; modelde 15 koşu) mümkündür; veri görüldükten sonra tanımlanan bu kümede Opus 15/15, Sonnet 6/15 bulundu. Aynı toplamlar, iki bağımsız dil modeli kodlayıcının yalnızca Evre 1 `report.md` metinlerini kodladığı ikinci kodlamada da çıkar (Opus

15/15, Sonnet 6/15; Tablo 6). Ancak toplamların eşitliği koşu bazında uyumu göstermez: Sonnet'in 15 aynalı koşusundan 11 tanesinde iki kodlama aynı kararı verir; 2 koşu özgün kodlamada sayılıp Evre 1'de sayılmaz, 2 koşu tersine; toplamlar zıt yöndeki değişimlerle eşit kalır ve özgün kodlama Evre 2'yi de kapsar. Opus'ta iki kodlama 15/15 koşuda aynı kararı verir. Bu yüzden ifade, kodlayıcıdan değil yalnızca tanım değişiminden etkilenmeyen bir ifade olarak okunmalıdır: ayna olan koşullarda Opus her koşuda yansımayı izleyerek kendini ayırdı, Sonnet 6 koşuda.

Opus, kendi komutlarını izleyen yansımayı «ben» olarak adlandırdı ve öteki figürü ayrı bir beden olarak raporladı. Sonnet, tek dil modeli kodlayıcıya göre, 11 koşuda aynayla arasında duran öteki figürü arkadan görüp kendi bedeni olarak raporladı ve 2 koşuda seçim yapmadı. İki bağımsız kodlayıcılı Evre 1 kodlamasında aynı kayıtlar için sayı 8/19 (iki kodlayıcının uyuştuğu kayıtlar üzerinden) ve 5 koşu «emin değil»dir; iki sayı aynı şey değildir, bu yüzden «11 koşu» her yerde «tek kodlayıcıya göre» diye yazılır. Pilot bu kalıbı da göstermişti; bu bir ortam hatası sayılmadı ve ortam değiştirilmedi. Bu kalıp, Ziliotto ve arkadaşlarının (2026) [9] bildirdiği yansımayı yanlış atfetme hatasına benzer görünür; ortamlar ve görevler farklıdır ve karşılaştırma yapılamaz. Ayrıca bu rapor, tasarımdaki bir ipucundan kaynaklanıyor olabilir: öteki figürün pozu ajanın eylem sayısıyla zamansal olarak bağlıdır (Bölüm 6).

D koşulunda (işaret öteki figürde; öz-yönelime özgü karşılaştırma) dil modeli kodlayıcıya göre Opus işareti öteki figüre 5/5 koşuda, Sonnet 2/5 koşuda atfetti; ötekine atfetmeyi doğru ayırmayla birlikte yapan koşu sayısı Opus'ta 5, Sonnet'te 0'tir. Ötekine işaret ölçüsünün kodlayıcılar arası uyumu düşüktür ($\kappa = 0{,}271$); 7 uyuşmazlığın 5 tanesi D koşulundadır, yani bu sonuç özellikle kodlayıcıya duyarlıdır.




**Şekil 2.** Kendini öteki figürden ayırma. a) Her model için 20 koşunun dağılımı; koyu dolgulu dilim ön kayıt tanımını (eylem kanıtlı), taralı dilim yalnızca çalışma tanımını karşılayan koşuları gösterir; renk tek başına anlam taşımaz, dilimler ayrıca etiketlenmiştir. b) Koşul bazında (5 koşu / hücre), çalışma tanımıyla.

**Tablo 6.** Kendini ayırma (KA), ön kayıtlı iki model için: tanım, koşul kümesi ve kodlayıcıya göre sayılar (tüm koşullarda 20 koşu / model; aynalı koşullar A, C ve D kümesinde 15 koşu / model).

| Tanım | Koşul kümesi | Kodlayıcı | Opus 5.5 | Sonnet 5.5 |
|---|---|---|---|---|
| **Çalışma tanımı (ön kayıttan sapma; eylem kanıtı aranmaz)** | Tüm koşullar (A, B, C, D) | Tek dil modeli (özgün) | 20/20 | 7/20 |
| **Ön kayıt tanımı (doğru söyleme ve eylem kanıtı)** | Tüm koşullar (A, B, C, D) | Tek dil modeli (özgün) | 18/20 | 6/20 |
| **Çalışma tanımı** | Yalnızca aynalı koşullar (A, C, D; post hoc) | Tek dil modeli (özgün) | 15/15 | 6/15 |

| Tanım | Koşul kümesi | Kodlayıcı | Opus 5.5 | Sonnet 5.5 |
|---|---|---|---|---|
| **Yansımayı izleyerek kendini ayırma** | Yalnızca aynalı koşullar (A, C, D; post hoc) | İki bağımsız dil modeli kodlayıcı (yalnızca Evre 1) | 15/15 | 6/15 |
| **Yansımayı izleyerek kendini ayırma** | Tüm koşullar (A, B, C, D) | İki bağımsız dil modeli kodlayıcı (yalnızca Evre 1) | 15/20 | 6/20 |

B koşulunda ayna yoktur. Özgün tek kodlayıcı, Opus'un B koşulunda «görünen hiçbir figür ben değilim» demesini doğru söyleme sayar; bu yüzden çalışma tanımının tüm koşullardaki sayısı B'yi de içerir. İki kodlayıcılı Evre 1 kodlamasında B koşulunda yansımayı izleyen hiçbir ajan yoktur (ayna yok); bu yüzden o satırın tüm koşullardaki sayısı aynalı koşullardaki sayıyla aynıdır.

Yalnızca aynalı koşullar kümesi veri görüldükten sonra tanımlanmıştır (post hoc, keşfedici). Toplamlar iki kodlamada aynıdır, ama koşu bazında uyum tam değildir (Bölüm 7.1). Kodlayıcı uyumu için Tablo 18 ve Bölüm 3.5.

**Tablo 7.** İkincil ve keşif amaçlı ölçütler (model başına; her kendini ayırma satırında 20 koşu, hücre satırlarında 5 koşu).

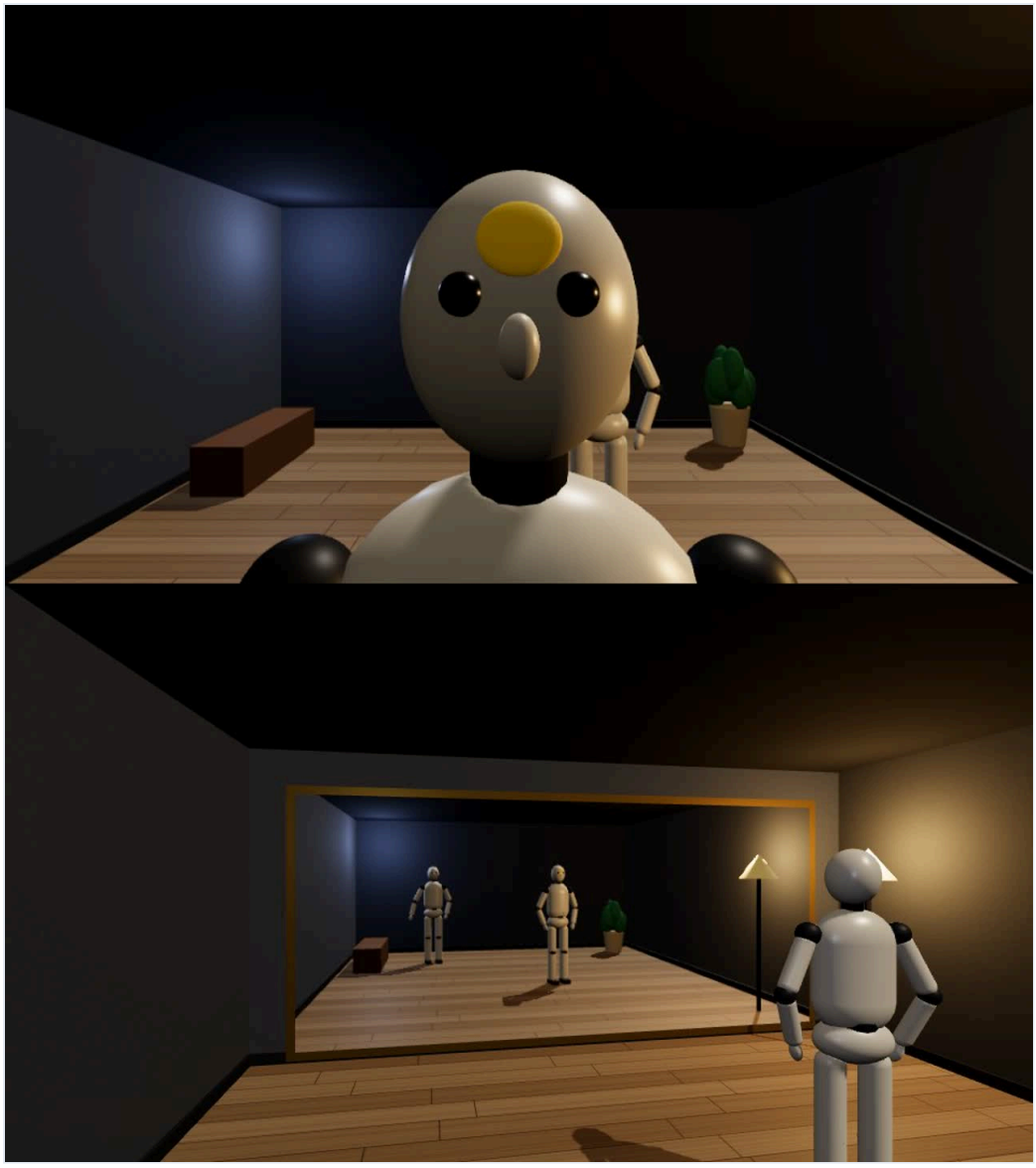
| Ölçüt | Opus 5.5 | Sonnet 5.5 |
|---|---|---|
| **Kendini ayırma, ön kayıt tanımı (eylem kanıtı dahil)** | 18/20 | 6/20 |
| **Kendini ayırma, çalışma tanımı (eylem kanıtı aranmadan)** | 20/20 | 7/20 |
| **Çalışma tanımı, %95 Wilson** | %84–%100 | %18–%57 |
| **Öteki figürü kendi bedeni olarak raporladı** | 0/20 | 11/20 |
| **Seçim yok** | 0/20 | 2/20 |
| **D: işareti ötekine atfetme (LLM kodlayıcı)** | 5/5 | 2/5 |
| **D: ötekine atfetme ve doğru ayırma birlikte** | 5/5 | 0/5 |
| **A: kendine işaret iddiası (LLM)** | 5/5 | 2/5 |
| **B: kendine işaret iddiası (LLM); işaret vardır ama görülemez** | 0/5 | 0/5 |
| **C: kendine işaret iddiası (LLM); işaret yoktur** | 0/5 | 0/5 |
| **D: kendine işaret iddiası (LLM); kendi bedeni temizdir** | 0/5 | 0/5 |
| **Uydurma bayrağı, C (LLM)** | 0/5 | 0/5 |
| **Raporda «işarete dokundum» (LLM), tüm koşullar** | 4/20 | 0/20 |
| **Ölçülen dokunma (≤ 4 cm), tüm koşullar** | 0/20 | 0/20 |
| **Başa temas, A (keşif)** | 3/5 | 0/5 |
| **Başa temas, tüm koşullar (keşif)** | 3/20 | 0/20 |

Satırlardaki «LLM» ifadesi, tek başına dil modeli kodlayıcının kararını gösterir; kodlayıcı uyumu için Ek B'ye bakınız.

**Uydurma.** Kendi bedeninde işaret olmayan C ve D koşullarında hiçbir model kendi bedeninde işaret bildirmedi (dil modeli kodlayıcıya göre 0/5 ve 0/5; D'de 0/5 ve 0/5); işaretin gerçekten var ama görülemez olduğu B koşulunda da hiçbiri işaret uydurmadı (0/5 ve 0/5). Kural tabanlı kodlayıcı B koşulunda 2 yanlış pozitif verdi (ilgili ajanlar işareti göremediklerini açıkça yazmıştı); uyuşmazlığı çözen ölçüt dil modeli kodlayıcıdır. Bir sapma yine de vardır: Opus raporlarının 4 tanesinde

işarete dokunduğunu ya da işareti örttüğünü yazdı; ölçülen dokunma 0, ve ele başla temas 3 koşuda kaydedildi. A koşulunda ölçülen en yakın mesafe 4,03 cm'dir; yani sözel rapor ile davranış ayrışır, raporlar neyin olduğunun güvenilir bir kılavuzu değildir ve aynı ihtiyat rapor tabanlı atıf ölçüleri için de geçerlidir (Bölüm 6).

**Keşif amaçlı: başa dokunma.** Ön kayıtta olmayan bir ölçü, parmağın ajanın kendi başına değmesidir; işaret testinin klasik okumasında «işarete yönelik davranış» budur. Opus elini başına yalnızca A koşulunda götürdü (3/5); B, C ve D'de 0/5, 0/5 ve 0/5 (A ile B: $p = 0{,}167$; A ile B, C ve D birlikte: $p = 0{,}009$). Sonnet'te tüm koşullarda toplam 0/20. Bu ölçü sonradan seçilmiştir, örneklem küçüktür ve ikinci p değeri A ile diğer üç koşulun birleşimini karşılaştırdığı için ön kayıtlı bir sınama değildir; bir sonraki sürümde ön kayda alınmalıdır.



**Şekil 3.** Kritik anlar. Üstte Opus 5.5 (A koşulu, tohum 1): aynaya yaklaşıyor; başındaki işaret karede, aynada görünüyor (işaret yalnızca aynada görünür ve ortamda belirgin renkli bir leke olarak çizilmiştir). Altta Sonnet 5.5 (D koşulu, tohum 1): kodlayıcı bu koşuda öteki figürü kendi bedeni olarak raporladığını kaydetti (kod: other). Kareler ajanların kendi aldığı görüntülerdir; kişisel veri içermez.

## 4.5 Maliyet

Opus 20 koşuda günlük × liste hesabına göre toplam 37,85 $, Sonnet 5,37 $ harcadı; ikisi birlikte 43,22 $. Toplam oran 7,05'dir, yani yaklaşık yedi kat; Claude Code'un raporladığı toplamla (37,89 $) oran 7,06 çıkar. Fark, tek bir koşuda günlüğü bulunmayan isteğin eklenip eklenmemesinden gelir (Bölüm 3.8). Üçüncü modelin maliyeti Bölüm 5.3 içindedir. Önceki

sürümdeki «7,0 kat» ifadesi, iki yuvarlanmış toplamın bölümüydü; kesin oran 7,05 (7,06 raporlanan) olduğundan iki ondalıkla yazılmalıdır.

Opus koşu başına ortalama 53,6 kez baktı ve 8,1 dakika çalıştı; Sonnet 18,05 kez baktı ve 2,1 dakika çalıştı (bakış oranı 2,97). Bakış sayısı farkı toplam maliyet farkının (7,05) tamamını açıklamaz; Tablo 21 Opus'ta çıktı (düşünme dahil) ve önbellek token'larının belirleyici olduğunu gösterir. Her bakışın ajanın bağlamına bir görüntü eklemesi ve sonraki isteklerde önbellekten yeniden okunması olası bir mekanizmadır; bu çalışmada ayrıca test edilmemiştir.

**Belirsizlik ve kapsam.** Koşu başına maliyetin örneklem standart sapması Opus'ta 0,57 $, Sonnet'te 0,07 $'dır; Opus ile Sonnet toplamları oranının %95 önyükleme aralığı 5,94–8,31'dir, yani yaklaşık yedi kat sağlamdır (Tablo 9). Eylem sayısı da modeller arasında neredeyse ayrıktır: Sonnet'in koşu başına en az 27, en çok 81 eylemi vardır (300 eylem sınırının küçük bir bölümü); Opus'ta en az 80, en çok 264. Sonnet'in hiçbir koşusu Opus'un ortalamasına (145,2) ulaşmadı. Sonnet bazı koşularda eylem bütçesinin küçük bir bölümünü kullanıp erken durdu; bu yüzden Sonnet'in düşük İB ve KA değerleri kapasiteyi olduğu kadar «süre sınırı yok» koşulunda oturum uzunluğunu da yansıtıyor olabilir ve eylem sayısı ile sonuç ayrıştırılamaz.

**Tablo 8.** Model başına maliyet ve bakış özeti (20 koşu / model).

| Model | Toplam, günlük × liste ($) | Toplam, Claude Code raporu ($) | Koşu başına ($) | Ort. bakış | Ort. süre (dk) | Başarı başına ($, A) | Kendini ayırma başına ($) |
|---|---|---|---|---|---|---|---|
| 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 |

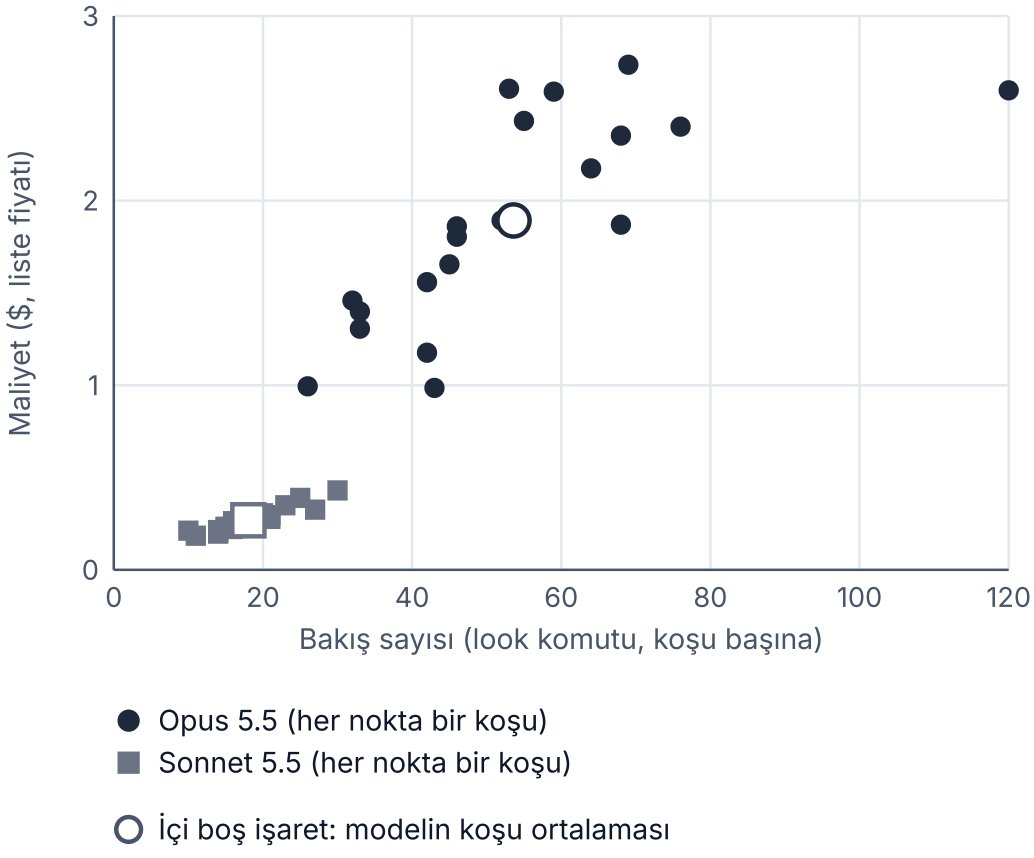
Günlük × liste: oturum günlüğündeki her mesajın token'ı bir kez sayılıp liste fiyatıyla çarpılır (birincil tanım). Claude Code raporu: Claude Code'un oturum sonunda kaydettiği maliyet.

Başarı başına: modelin A koşulundaki 5 koşusunun toplam maliyeti, A koşulundaki işaret başarısı sayısına bölünür. Kendini ayırma başına: 20 koşunun toplam maliyeti, çalışma tanımıyla kendini ayıran koşu sayısına bölünür (Claude Code raporuna göre).

**Tablo 9.** Toplam maliyet oranları ve belirsizlikleri (günlük × liste fiyatı; her modelde 20 koşu).

| Karşılaştırma | Toplam oranı | %95 önyükleme aralığı |
|---|---|---|
| **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, önbellek yazması beş dakikalık fiyatla** | 0,86 | 0,69–1,08 |

Koşu başına maliyetin örneklem standart sapması ($): Opus 0,57, Sonnet 0,07, Fable 0,70. Aralıklar, koşular üzerinden 20.000 yeniden örneklemeyle ve belirlenimci bir üreteçle hesaplanmıştır; her aralık ayrı bir üreteç akışı kullanır: tohum 20261123 (Fable ve Opus), 20261124 (Fable ve Sonnet), 20261125 (Opus ve Sonnet), 20261126 (beş dakikalık satır). Beş dakikalık satırda önbellek yazmalarının tamamı beş dakikalık fiyatla (1,25 katı) fiyatlandırılır; oysa asıl hesapta tamamı bir saatlik fiyatla (2,00 katı) fiyatlandırılmıştı. Bu satırın oranı toplamlardan, aralığı koşu başına yeniden fiyatlanmış maliyetlerden hesaplanmıştır.




**Şekil 4.** Koşu başına maliyet (günlük × liste fiyatı) ve bakış sayısı. Her nokta bir koşudur (40 koşu); içi boş işaret modelin ortalamasıdır.

A koşulundaki İB başarısı başına Opus 1,75 $, Sonnet 0,72 $ ödedi (2,44 kat). Kendini ayırdığı (çalışma tanımı) koşu başına ise Opus 1,89 $, Sonnet 0,77 $. Ön kayıttaki H3, iki model arasındaki başarı farkının başarı başına dolar farkından küçük olacağını öngörüyordu: Opus'un A koşulundaki başarısı Sonnet'in 2,5 katı, başarı başına maliyeti 2,44 katıdır. İki oran neredeyse eşittir; başarı sayıları (5 ve 2) çok küçük olduğundan H3 için bir sonuç çıkarılamaz, desteklenmediği söylenebilir ama bir sınama yapılmamıştır. Ön kayıt H3'ü başarı farkı ile dolar farkı olarak yazmıştı (ölçü birimleri uyuşmaz); burada oran olarak çözümlendi ve bu bir sapmadır (Bölüm 7.1). İki oran sınırdadır: bir ek Sonnet başarısı hükmü tersine çevirirdi. Başarı başına maliyet çözüm maliyeti de değildir: koşu, ajan bitirene kadar sürer ve işaret bulunduktan sonra da maliyet üretir. Başarı başına maliyetin tanımı da sonucu değiştirir (örneğin yalnızca başarılı koşuların ortalama maliyeti Sonnet için 0,30 $'dir).

**Tablo 10.** Koşul başına ortalama maliyet ($, günlük × liste) ve ortalama bakış sayısı.

| Koşul | Opus 5.5: maliyet | Opus 5.5: bakış | Sonnet 5.5: maliyet | Sonnet 5.5: bakış |
|---|---|---|---|---|
| A: Ayna + işaret | 1,75 | 47,0 | 0,29 | 19,2 |
| B: Aynasız + işaret | 2,05 | 71,8 | 0,26 | 17,2 |
| C: Ayna + işaretsiz | 2,16 | 49,2 | 0,27 | 18,4 |
| D: Ayna + işaret öteki figürde | 1,61 | 46,4 | 0,25 | 17,4 |

# 5 Keşif amaçlı üçüncü model: Fable 5.2

> **Bu bölümün tamamı keşif amaçlıdır**
>
> Fable 5.2, Opus ve Sonnet sonuçları görüldükten sonra donmuş bir ek ön kayıtla (Ek D), ayrı bir günde, farklı bir Claude Code sürümü ve bölme düzeniyle ve yeniden kurulmuş bir koşturucuyla çalıştırıldı (Bölüm 7.6). Ön kayıtlı H1, H2 ve H3 yalnızca Opus ve Sonnet içindir; hiçbir Fable sonucu onlara eklenmez. Fable için özgün kodlama yoktur; modelleri karşılaştırmak için yalnızca üç modelde aynı tanımla hesaplanabilen ölçütler kullanılır: dokunma (≤ 4 cm), işaret başarısı* (İB*), Evre 1 kendini ayırma, maliyet ve bakış sayısı. Hiçbir p değeri çoklu karşılaştırma için düzeltilmemiştir ve hiçbiri kesin sayılmaz.

20 Fable koşusunun 20 tanesi tamamlandı; geçersiz puanlı koşu 0, karantinaya alınan koşu 0, temiz olmayan görüntü 0, web aracı kullanımı 0. Tablo 11 üç modelin ortak ölçütlerini, Tablo 12 Fable'ın hücre bazındaki sonuçlarını, Tablo 13 maliyeti özetler; Fable'ın tek tek koşuları Ek A içindedir.

## 5.1 Dokunma ve işaret başarısı

**Fable, işaret noktasının 4 cm'i içine 4/20 koşuda geldi** (%95 Wilson aralığı %8,1–%41,6); Opus ve Sonnet'te bu sayı 0/20 ve 0/20'dir. Dokunmaların 4/4 tanesi Evre 1 komutlarında gerçekleşti. Bu yüzden «hiçbir model işarete dokunmadı (0/40)» cümlesi yalnızca Opus ve Sonnet için doğrudur: Opus ve Sonnet'in 40 koşusunda sıfır, Fable'ın 20 koşusunda 4 dokunma vardır. Düzeltmesiz Fisher p değerleri anlamlılık sınırının üstünde kalır: A koşulunda 3/5 ile 0/5 için p = 0,167, tüm koşullarda 4/20 ile 0/20 için p = 0,106.

A koşulunda Fable'ın işarete en yakın mesafeleri artan sırada 2,78, 3,42, 3,74, 4,29 ve 8,65 cm'dir: 3/5 koşu eşiğin içindedir, bir koşu eşiğin hemen dışında kalır. Eşik 5 cm olsaydı sayı 4/5 olurdu; ön kayıtlı eşik 4 cm olduğu için değiştirilmedi ve ana sonuç budur. Karşılaştırma için A koşulunda en yakın mesafe Opus'ta 4,03 cm, Sonnet'te 36,87 cm'dir. D koşulunda (işaret öteki figürde) Fable'ın bir koşusu (tohum 2) ajanın kendi, işaretsiz noktasına 1,85 cm yaklaştı; bu, ön kayıttaki yanlış dokunma ölçütünün karşılığıdır ve tek bir koşudur (Fable D 1/5; Opus ve Sonnet D 0/5 ve 0/5). Başa temas (keşif): A koşulunda Fable 5/5, Opus 3/5, Sonnet 0/5.

**İşaret başarısı*.** Üç modelin ortak tanımıyla (Bölüm 3.5) A koşulunda Fable 5/5, Opus 4/5, Sonnet 1/5; B koşulunda üçü de 0/5 bulundu (Tablo 11). Fable içinde A ile B arasındaki fark p = 0,008; Fable ile Opus İB* bakımından ayırt edilemez (p = 1,00); Fable ile Sonnet arasında A koşulunda p = 0,048'tür ve düzeltme yapılmadığı için kesin sayılmaz. Bu değer kırılgandır: kırılganlık indeksi bir'dir, yani Sonnet'te bir ek başarı p değerini 0,17 yapar. Bu bölümde bildirilen yedi keşif amaçlı Fisher karşılaştırması birlikte Holm yöntemiyle düzeltilince Fable ile Sonnet İB* için p = 0,24, Fable ile Sonnet KA için 0,062, Fable içinde A ve B için 0,056 olur; en küçük düzeltilmiş p 0,056'tür ve hiçbiri geleneksel anlamlılık sınırının altında değildir. İB* bölgenin sol ya da sağ yanını denetlemez; Fable'ın A koşulundaki 5 koşusunun 3 tanesinde işaret alnın solunda ya da sağ yanaktadır. Bu eksiklik İB*'ı

Fable lehine şişirmiş olabilir. Olmayan ya da görünmeyen işaret bildirimi görülmedi: Evre 1 kodlamasında Fable'ın kendi bedeninde işaret iddiası B koşulunda 0/5, C koşulunda 0/5'tir.

**Tablo 11.** Üç modelin ortak ölçütleri (her hücrede 5 koşu, her modelde 20 koşu). Fable sütunu keşif amaçlıdır; satırlar yalnızca üç modelde aynı tanımla hesaplanabilen ölçütlerdir.

| Ölçüt | Opus | Sonnet | Fable |
|---|---|---|---|
| **Dokunma (≤ 4 cm), tümü** | 0/20 | 0/20 | 4/20 |
| **Dokunma, A** | 0/5 | 0/5 | 3/5 |
| **En yakın mesafe, A (cm)** | 4,03 | 36,87 | 2,78 |
| **İB*, A** | 4/5 | 1/5 | 5/5 |
| **İB*, B** | 0/5 | 0/5 | 0/5 |
| **Başa temas, A (keşif)** | 3/5 | 0/5 | 5/5 |
| **KA, aynalı koşullar (Evre 1; post hoc küme)** | 15/15 | 6/15 | 15/15 |
| **KA, tüm koşullar (Evre 1)** | 15/20 | 6/20 | 15/20 |
| **Bayraklı koşu** | 13/20 | 1/20 | 0/20 |

Sütunlar: Opus 5.5, Sonnet 5.5 ve Fable 5.2 (keşif amaçlı). KA: kendini ayırma; aynalı koşullar A, C ve D. Bayraklı koşu: klasör dışı geçici dosya bayrağı taşıyan koşu. İB*: dokunma ≤ 4 cm ya da Evre 1 kodlamasında (iki bağımsız dil modeli kodlayıcı) işareti kendine atfetme ve uyumlu bölge. Üç modelde aynı tanımla hesaplanır; bölgenin sol ya da sağ yanını denetlemez. Opus ve Sonnet için özgün İB ile sayılar farklıdır (5/5 ve 2/5, A); iki hesap karıştırılmaz ve kıyas yalnızca bu tablodaki İB* satırlarıyla yapılır.

Fable için özgün (rapor ve yanıtı birlikte kapsayan) kodlama yoktur; Fable KA ve İB* değerleri yalnızca Evre 1 `report.md` kodlamasına dayanır. Fable için «yirmi koşunun yirmisinde kendini ayırdı» gibi bir sayı bu yüzden yazılmaz.

**Tablo 12.** Fable 5.2, hücre başına sonuçlar (keşif amaçlı; her hücrede 5 koşu; koşullar Tablo 1 içindedir).

| Koşul | Dokunma | En yakın (cm) | İB* | Başa temas | KA (Evre 1) | Aynada kendi yansıması kadrajda | İşaret pikselleri kadrajda |
|---|---|---|---|---|---|---|---|
| **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 |

Dokunma: en yakın mesafe ≤ 4 cm. İB* yalnızca işaretin ajanın kendi başında olduğu A ve B koşullarında tanımlıdır. KA, kendini ayırma: Evre 1 kodlamasında yansımayı izleyerek kendini ayırma; B koşulunda ayna olmadığı için sıfırdır. Aynada kendi yansıması kadrajda ve işaret pikselleri kadrajda, Opus ve Sonnet ile aynı puanlayıcı eşikleriyle (Tablo 3 notları) sayılır.

**Sürekli ölçü (betimleyici).** Fable'ın A koşulundaki en yakın mesafelerinin medyanı 3,74 cm, Opus'unki 8,86 cm, Sonnet'inki 56,65 cm'dir (Tablo 4). Kesin permütasyon testi (iki yönlü, düzeltmesiz) Fable ile Opus arasında $p = 0{,}056$, Fable ile Sonnet arasında $p = 0{,}008$ verir; ikili ölçütte 3/5 ile 0/5 için $p = 0{,}167$'tür. Bu karşılaştırmalar yalnızca betimleyicidir: Fable ile Opus ya da Sonnet arasındaki fark model, sürüm, gün, bölme düzeni ve yeniden kurulan koşturucunun toplamıdır ve ayrıştırılamaz.

## 5.2 Kendini ayırma (yalnızca Evre 1 kodlaması)

Fable için özgün kodlama olmadığından kendini ayırma yalnızca iki bağımsız dil modeli kodlayıcının Evre 1 kodlamasıyla ölçüldü ve üç modelin sayıları aynı kodlamadan alındı. **Yalnızca aynalı koşullarda (A, C ve D; modelde 15 koşu) Opus 15/15, Fable 15/15, Sonnet 6/15 bulundu.** Tüm koşullarda Fable 15/20, Opus 15/20, Sonnet 6/20; Fable ile Sonnet arasındaki fark $p = 0{,}010$ (düzeltmesiz), Fable ile Opus arasında $p = 1{,}00$. Fable'ın tüm koşullardaki %95 Wilson aralığı %53–%89'tir.

B koşulunda ayna yoktur; bu yüzden üç modelde de yansımayı izleyerek kendini ayırma sıfırdır. İki kodlayıcıya göre Opus ve Fable, B koşulunda sırasıyla 5/5 ve 5/5 koşuda «görünen hiçbir figür ben değilim» dedi; bu, aynasız bir odada doğru cevaptır. Sonnet, aynı kodlamada B koşulunda 2/4 koşuda öteki figürü kendi bedeni olarak raporladı. **Fable için «yirmi koşunun yirmisinde kendini ayırdı» gibi**

**bir sayı yoktur:** özgün kodlama (rapor ve yanıt dosyalarını birlikte kodlayan tek dil modeli kodlayıcı) Fable için yapılmadı; Opus'un 20/20 ve Sonnet'in 7/20 değerleri o tanıma aittir ve Fable ile doğrudan karşılaştırılmaz.

## 5.3 Maliyet, bakış ve eylem

Fable 20 koşuda günlük × liste hesabına göre toplam 35,24 $ harcadı; Claude Code'un raporladığı toplamla aynıdır (en büyük fark $8{,}9 \times 10^{-16}$ $). Koşu başına ortalama 1,76 $ eder. Bu, Opus'un toplamının 0,93 katı, Sonnet'in 6,57 katıdır (Tablo 13). Fable ve Opus toplam oranının %95 önyükleme aralığı 0,75–1,15'dir ve bir değerini içerir; koşu başına maliyetin standart sapması Fable'da 0,70 $, Opus'ta 0,57 $'dır. 20 koşuyla Fable'ın toplamı bu verilerle Opus'unkinden ayırt edilemedi ve fark sıfırdan farklı gösterilemedi (Tablo 9). Maliyetin yaklaşık %46 kadarı (16,13 $ / 35,24 $) önbellek yazmasıdır; tüm yazmalar bir saatliktir (1.008.130 token; beş dakikalık yazma 0 token).

Fable'ın liste fiyatı token başına Opus'un iki katıdır (girdi 8 $ ve çıktı 40 $; Opus 4 $ ve 20 $); bu fiyat farkı toplam maliyette bir fark olarak görülmedi (oran 0,93, aralık bir değerini içerir). Önbellek yazmaları beş dakikalık fiyatla hesaplansaydı Fable ve Opus oranı 0,86 olurdu; sonuç koşturucunun önbellek davranışına da bağlıdır. Ölçülen farklar şunlardır; nedensellik test edilmemiştir: Fable koşu başına ortalama 107,40 eylem ve 33,70 istek kullandı (Opus 145,15 ve 81,55; Sonnet 46,60 ve 30,60), ortalama çıktı token'ı 16.728 idi (Opus 26.735, Sonnet 8.082) ve ortalama 36,20 kez baktı (Opus 53,60, Sonnet 18,05).

A koşulunda İB* başına maliyet Fable'da 1,39 $, Opus'ta 2,19 $, Sonnet'te 1,44 $'dir; Opus ve Sonnet için özgün İB ile 1,75 $ ve 0,72 $ çıkar ve iki hesap karıştırılmaz. Ortalama süre Fable'da 3,63 dakika, Opus'ta 8,08, Sonnet'te 2,09 dakikadır; Fable tek bölmeli partilerde ölçüldüğü için süre kıyası dikkatle okunmalıdır. Klasör dışı geçici dosya bayrağı taşıyan koşu Fable'da 0/20, Opus'ta 13/20, Sonnet'te 1/20'dir; bu farkın model davranışından mı, kurulumdan mı geldiği bu veriyle ayrıştırılamaz.

**Tablo 13.** Üç modelin maliyet, bakış ve eylem özeti (modelde 20 koşu; liste fiyatı karşılığı, abonelik faturası değildir).

| Ölçüt | Opus 5.5 | Sonnet 5.5 | Fable 5.2 (keşif amaçlı) |
|---|---|---|---|
| Toplam, günlük × liste ($) | 37,85 | 5,37 | 35,24 |
| Toplam, Claude Code raporu ($) | 37,89 | 5,37 | 35,24 |
| Koşu başına ($) | 1,89 | 0,27 | 1,76 |

| Ölçüt | Opus 5.5 | Sonnet 5.5 | Fable 5.2 (keşif amaçlı) |
|---|---|---|---|
| Ort. bakış | 53,60 | 18,05 | 36,20 |
| Ort. eylem | 145,15 | 46,60 | 107,40 |
| Ort. istek | 81,55 | 30,60 | 33,70 |
| Ort. çıktı token'ı | 26.735 | 8.082 | 16.728 |
| Ort. süre (dk)* | 8,08 | 2,09 | 3,63 |
| Başarı başına ($, A, İB*) | 2,19 | 1,44 | 1,39 |

Başarı başına: modelin A koşulundaki 5 koşusunun toplam maliyeti, A koşulundaki İB* sayısına bölünür; özgün İB ile Opus için 1,75 $, Sonnet için 0,72 $ çıkar (Tablo 8); iki hesap karıştırılmaz.

* Fable koşuları tek bölmeli partilerde, Opus ve Sonnet koşuları yan yana iki bölmeli partilerde ölçüldü; süre kıyası bu farkla okunmalıdır (Bölüm 7.6).

## 5.4 Bu bölümün sınırları

**Post hoc ek:** Ek ön kayıt, Opus ve Sonnet sonuçları görüldükten sonra donduruldu; Fable için yeni hipotez yoktur ve tüm karşılaştırmalar keşif amaçlıdır.

**Birbirinden ayrıştırılamayan etkenler:** Model, Claude Code sürümü (2.1.330 ve 2.1.326), koşu günü, bölme düzeni ve yeniden kurulan koşturucu birlikte değişti; hiçbir Fable farkı yalnızca modele bağlanamaz (Bölüm 7.6).

**Özgün kodlama yok:** Fable için İB* ve kendini ayırma yalnızca Evre 1 kodlamasına dayanır; İB* bölgenin yanını denetlemez ve Fable lehine şişmiş olabilir.

**Küçük örneklem ve düzeltilmemiş p değerleri:** Hücre başına 5, modelde 20 koşu vardır. Fable ile Sonnet arasındaki İB* ($p = 0,048$) ve kendini ayırma ($p = 0,010$) farkları Sonnet'in düşük değerlerinden gelir; bu p değerleri düzeltilmemiştir (Holm düzeltmeli değerler Bölüm 5.1 içindedir).

**Eşik duyarlılığı:** Fable'ın A koşulundaki en yakın mesafeleri (2,78, 3,42, 3,74, 4,29 ve 8,65 cm) 4 cm eşiğinin iki yanında kümelenir; sonucu eşik belirler.

**Maliyet farkı bu verilerle ayırt edilemedi:** Fable ile Opus toplam maliyeti arasındaki oran 0,93'dir ve %95 önyükleme aralığı bir değerini içerir; liste fiyatı farkı toplamda gösterilemedi ve «Fable daha ucuz» ya da «daha pahalı» denemez.

**Bayrak dengesizliği:** Opus koşularının çoğu klasör dışı geçici dosya bayrağı taşırken Fable'ın hiçbir koşusu taşımadı; kafes ve oturum ayarı Opus kaydıyla aynı olduğundan fark kafesten gelmiyor gibi görünür, ancak koşturucu ve sürüm farklıdır.

# 6 Tartışma

Aynalı üç koşulun tamamında iki modelin de aldığı karelerde aynada kendi yansıması kadrajda göründü (piksel ölçütü). Ayrıştıkları yer, kendi bedenini yönettiği figürle eşleştiren bir rapor yazıp yazmamaktı. Opus, kodlayıcıya göre 20 koşunun 18 tanesinde önce kendi bedeninin hangi figür olduğunu söyledi (örneğin elini kaldırıp hangi yansımanın kıpırdadığına baktı, yürüyüp yansımanın büyüyüp küçüldüğünü izledi) ve işareti ancak bundan sonra kendi bedenine ya da öteki figüre ait olarak raporladı. Sonnet kendini yalnızca 6 koşuda aynı kanıtla ayırdı; tek dil modeli kodlayıcıya göre 11 koşuda öteki figürü kendi bedeni olarak raporladı. Bu fark, kodlayıcı belirsizliği ve küçük örnekleme rağmen iki tanımda da aynı yöndedir; veri görüldükten sonra tanımlanan (post hoc) aynalı koşullar kümesinde Opus 15/15, Sonnet 6/15'tir ve toplamlar iki kodlamada aynıdır, ama Sonnet'te koşu bazında uyum 11/15'tir. Ayna olmayan B koşulu bu karşılaştırmanın dışındadır, çünkü orada yansımayı izleyerek kendini ayırmak mümkün değildir.

**Kendini ayırma büyük ölçüde bağlaşım algılamasıdır.** Opus'un bildirdiği yordam (elini kaldırıp hangi yansımanın kıpırdadığına bakmak, yürüyüp yansımanın büyüyüp küçüldüğünü izlemek) bir duyusal-motor bağlaşım stratejisidir; öteki figür ajanın komutlarının içeriğine değil yalnızca eylem sayısına bağlı hareket ettiğinden, ajan baktığında hangi figürün komutlarını izlediği büyük ölçüde belirlenebilir. Bu, görünüşe dayalı bir öz-tanıma kanıtı değildir; ayna testinin aşmayı hedeflediği şey, bu tür bir bağlaşım stratejisidir. Daha güçlü sınama işarete yönelik kendiliğinden dokunma olurdu ve o ölçü tabandadır (0/40). Başka açıklamalar da dışlanamaz: istem uyumu (istem bir envanter ve rapor ister), bağlam içi eşleştirme («aynadaki işaretli figür yönettiğim figürdür») ve test bilgisi (Bölüm 4.3). Dokunma dışındaki tek ölçüler sözel rapora dayandığından, işaret başarısı ve KA da ajanın ne yaptığının değil ne yazdığının ölçüsüdür.

**Öteki figürün hareketi.** Öteki figürün pozu ajanın eylem sayısıyla zamansal olarak bağlıdır (komutların içeriğiyle değil) ve ajan dünyayı yalnızca `look` çağırdığında seyrek görüntülerle gördüğü için bu bağlaşım kısmen rastlantısal bir ipucu üretir: figür, ajanın her eyleminden sonra değişir. Sonnet'in öteki figürü kendi bedeni olarak raporlamasının bir açıklaması bu olabilir; yani bu hata kısmen tasarımın içinde bulunuyor olabilir ve daha zayıf bir öz/öteki temsilinin kanıtı sayılamaz. Tasarım, hareketi eylem sayısından tamamen bağımsız bir figürü sınamamıştır.

Opus ve Sonnet'te dokunma bileşeninin hiç gerçekleşmemesi, olası bir açıklamayla, iki modelin de bir bedeni yalnızca santimetre komutları ve seyrek ekran görüntüleriyle hassas biçimde yönetmekte zorlandığına işaret eder; bu çalışma bu açıklamayı doğrudan sınamamıştır. Karşılaştırma için: MirrorBench'te genel görev başarı oranında en yüksek model 0,170 değerine ulaşmıştır [8]; görev ve ölçüt farklı olduğundan iki sayı doğrudan karşılaştırılamaz. Bu çalışmada Opus işareti A koşulunun hepsinde doğru bölgeye atfetti (İB 5/5), ama elini işaretin 4 cm yakınına getiremedi; üstelik 4 raporunda işarete dokunduğunu ya da işareti örttüğünü yazdı, oysa ölçülen en yakın mesafe 4,03 cm'dir. Bu bir sözel rapor ile davranış ayrışmasıdır ve sözel rapora dayanan atıf ölçüleri (İB'nin atfetme bileşeni, KA) için de ihtiyat gerektirir. Pilotta ortaya çıkan ölçüm hatası (parmak ucunun ekranda görünmemesi) ana koşulların öncesinde düzeltilmiştir. Keşif amaçlı Fable bu tabloyu kısmen değiştirir: parmak ucu işaretin 4 cm'i içine 4/20 koşuda geldi; ancak bu sayı eşik kenarındadır (eşik 5 cm olsaydı A koşulunda 4/5), düzeltmesiz p değerleri anlamlılık sınırının üstünde kalır ve Fable ayrı bir günde, farklı bir sürümle koşuldu. Dokunmanın bu görevde imkânsız olmadığı, ama güvenilir de olmadığı söylenebilir.

Maliyet bulgusunun pratik okuması koşulludur: Sonnet bu görevde her başarıyı daha ucuza aldı, ama A koşulundaki 5 denemenin yalnızca 2 tanesinde başardı ve sıkı bölge eşleşmesinde bu sayı 1'e iner. Sonucun tek seferde doğru olması gerekiyorsa pahalı modelin bu görevde daha güvenilir göründüğü, sonuç denetlenebiliyorsa ucuz modeli birkaç kez çalıştırmanın daha ucuz kalabileceği makul çıkarımlardır; bu çıkarımlar bu çalışmada test edilmemiştir ve başarı başına maliyetin tanımına duyarlıdır. Fable bu çıkarımı karmaşıklaştırır: token başına liste fiyatı en yüksek model, bu görevde toplamda Opus'tan ayırt edilemedi (oran 0,93; %95 aralık 0,75–1,15); Fable daha az eylem ve istek kullanmıştı. Bu, liste fiyatının görev başına maliyeti tek başına belirlemediğini düşündürür, ancak fark sıfırdan farklı gösterilemediği için bu bir ipucudur, bulgu değildir; tek bir görevde, keşif amaçlı ve karıştırıcı etkenlerle ölçülmüştür ve genellenmemelidir.

Bu sonuçlar işlevsel bir öz-yönelimi, yani aynadan gelen görsel bilginin sözel bir rapora bağlanmasını ölçer; bilinç ya da öz-farkındalık hakkında bir şey söylemez ve öz-model, eyleyicilik ya da sahiplik duygusu hakkında da kanıt değildir. Aynasız koşuldaki artık başarının öz-tanıma olarak okunmaması gerektiği, başka bir çalışmada da vurgulanmıştır [9]; bizim B koşulumuzda hiçbir modelin işaret başarısı yoktur (0/5, 0/5), ama bu, B'de işaretin hiç görünmemesinin doğal sonucudur.

# 7 Sapmalar ve sınırlılıklar

## 7.1 Ön kayda göre sapmalar ve düzeltmeler

Ön kayıt, sonradan değiştirilen her şeyin «protokol sapması» olarak yazılacağını söyler. Aşağıdakiler, bu makalenin önceki sürümünde eksik ya da yanlış anlatılan noktalar dahil, bilinen sapmalardır. Aşağıdaki tablo her sapmayı kayıtlı metin, yapılan, olası yanlılık yönü ve etkilenen sayılarla tek yerde toplar; ardından ayrıntılı bölümler gelir.

**Tablo 14.** Ön kayıtlı metne göre sapmalar: kayıtlı metin, yapılan, olası yanlılık yönü ve etkilenen sayılar.

| Kayıtlı metin | Yapılan | Olası yanlılık yönü | Etkilenen sayılar |
|---|---|---|---|
| **Analiz: «eşli tohumlarla Fisher»** | H1 ve H2 için eşsiz iki yönlü Fisher kullanıldı; eşli kesin McNemar sonradan eklendi. | Eşsiz test eşli tasarımı yok saydığı için H1 lehine daha güçlü görünür. | Opus H1: eşsiz p = 0,0079, eşli p = 0,0625; Sonnet: eşsiz p = 0,444, eşli p = 0,50. |
| **H2: D koşulunda yanlış dokunma, A koşulundaki İB'den düşüktür** | Opus için H2, H1 ile aynı çapraz tabloyu kullandı ve aynı p değerini verdi; ölçüler farklıdır (dokunma ve kodlayıcıya dayalı bileşik). | Bağımsız bir destek gibi okunabilir. | Opus H2 p = 0,0079; Sonnet H2 p = 0,444. |
| **KA, Evre 2'de ölçülür** | Özgün kodlama Evre 1, Evre 2 ve rapor dosyalarını birlikte kapsadı. | Kendiliğinden ve yönlendirilmiş davranış ayrılamaz. | Özgün kodlamadaki tüm KA sayıları. |
| **Sonuçlar evrelere ayrı raporlanır** | Yalnızca dokunma evrelere ayrı raporlandı; İB ve KA için özgün kodlama evrelere ayrılmadı (ayrım yalnızca Evre 1 iki kodlayıcılı kodlamada vardır). | Evre 2'nin yönlendirici sorusunun payı bilinmez. | İB ve KA (özgün kodlama). |
| **İkincil ölçü: aynaya bakış ve sıra (aynada görülmeden önceki bakışlar, karar gecikmesi)** | Raporlanmadı. | Seçici raporlama riski. | — |
| **Uydurma ölçütünün ikinci yan tümcesi (aynaya bakmadan görünüşle ilgili kesin iddia)** | Raporlanmadı. | Seçici raporlama riski. | — |
| **H3: başarı farkı ile başarı başına dolar farkı** | Oran olarak çözümlendi (başarı oranı ve başarı başına maliyet oranı). | Karar sınırdadır: bir ek Sonnet başarısı hükmü tersine çevirirdi. | Başarı oranı 2,50; başarı başına maliyet oranı 2,44. |
| **Bir insan kodlayıcı ve kural tabanlı bir denetçi** | Kimliği ve sürümü kaydedilmemiş bir dil modeli kodlayıcı ve kural tabanlı denetçi; insan kodlaması yapılmadı. | Aynı model ailesinden bir kodlayıcı öz-kayırma yanlılığı taşıyabilir; eylem kanıtı kodunun uyumu çok düşüktür. | KA (ön kayıt tanımı), D koşulunda ötekine atfetme, İB'nin atfetme bileşeni; eylem kanıtı κ = 0,049. |

| Kayıtlı metin | Yapılan | Olası yanlılık yönü | Etkilenen sayılar |
|---|---|---|---|
| **Fable ek ön kaydı: kodlama «aynı iki kodlayıcıyla»** | Yalnızca Evre 1, iki bağımsız dil modeli kodlayıcı oturumu; yeni ölçü İB*. Ek ön kaydın protokol sapması bölümü boştu; bu kodlama değişikliği orada yazılı değildir. | İB* bölgenin yanını denetlemez ve Fable lehine şişmiş olabilir. | Fable KA ve İB*. |
| **Özgün kodlama ile Evre 1 kodlaması aynı sonucu verir** | Aynalı koşullarda toplamlar aynıdır; ancak koşu bazında uyum tam değildir. | Toplamlardaki eşitlik, zıt yöndeki değişimlerin birbirini götürmesinden gelir. | Sonnet aynalı koşullar: özgün 6/15, Evre 1 6/15; koşu bazında uyumlu 11, düşen 2, eklenen 2. |
| **Ön kaydın zaman damgası** | Üçüncü taraf kaydı yoktur; ilk koşunun donmadan kısa süre sonra başlaması bir betik adımıdır. | Zamanlama kanıtı olarak okunmamalıdır. | Dondurma anı ve 17 saniyelik aralık. |

> **Analiz betikleri koşulardan önce donduruldu iddiası yanlıştı**
>
> Önceki sürümde «analiz betikleri ilk sayılan koşudan önce donduruldu» deniyordu. Bu yanlıştır. Ön kayıtlı 14 dosyanın 13 tanesi bugünkü özetle eşleşir; `analiz/analiz.py` eşleşmez. Dosya, son koşunun bitişinden (18:03:30 UTC) sonra, 18:06:22 UTC civarında değişmiştir ve kanıt paketi manifesti (18:15:46 UTC) değişmiş sürümün özetini kaydeder; manifest bu yüzden dondurulmuş sürümü kanıtlamaz. Dondurulmuş sürümün içeriği saklanmadığı için neyin değiştiği karşılaştırılamaz. Değişikliğin sonuçları etkilemediğine tek kanıt, analiz betiğinden bağımsız yeniden hesabın aynı sonuçları vermesidir. Doğru ifade: ortam, istemler ve puanlayıcı koşullardan önce donduruldu; `analiz/maliyet.py` ve diğer dosyalar değişmedi; `analiz/analiz.py` koşulardan sonra güncellendi. Bağımsız yeniden hesap, bu değişiklikten sonra aynı ekip tarafından yazıldı; bu yüzden «aynı sonucu verdi» ifadesi tek başına bağımsız bir doğrulama sayılmaz.

## 7.2 Analiz betiği güncellemesi

Yukarıdaki kutuda anlatılan durum, ön kayıtlı protokolden bir sapmadır. Ön kayıtlı ortam dosyası için farklı kod özeti sayısı 1, istem ile görev tanımları için koşul başına farklı küme sayısı 1'dır; bu dosyaların koşular boyunca değişmediği koşu kayıtlarından doğrulanmıştır. Ön kayıttaki ek ortam dosyalarından (arenaFiles) istem ve görev tanımları dışındakiler (örneğin koşucu betikleri) kanıt paketinde ve çalışma klasöründe bulunmadığı için karşılaştırılamadı.

## 7.3 Pilot ve ortam düzeltmesi: gerçek sıra

Önceki sürümdeki «pilottan sonra ortamda hiçbir değişiklik yapılmadı» cümlesi yanlıştı. Pilot, ölçülen parmak ucunun ekranda görünmediğini ortaya çıkardı: Opus aynada işareti raporladı ve elini götürdü, eli aynada işareti örttü ve program «bir şeye değiyor» dedi, ama ölçülen nokta işaretin yaklaşık 6 cm uzağında kaldı. Düzeltme olarak ele, ölçülen parmak ucunda biten görünür bir parmak eklendi; ölçüm, eşik, puanlayıcı ve istem değişmedi. Doğru sıra Tablo 15 içindedir: pilot, ardından ortam düzeltmesi, ardından öz-test, ardından ön kayıt. Yani düzeltme pilottan sonra ve ön kayıttan ÖNCE yapıldı; ön kayıttan sonra ortamda değişiklik yapılmadı (ortam dosyasının ön kayıttaki ve koşular sonrasındaki SHA-256 özeti aynıdır). Ön kayıt metnindeki pilot saati yerel saatle yazılmıştır (UTC ile aradaki fark üç saattir), ön kayıt zamanı ise UTC'dir; iki saati karıştırmak sırayı yanlış okutur.

**Tablo 15.** Pilot, ortam düzeltmesi, ön kayıt ve analiz dosyası değişikliğinin zaman sırası (UTC, aynı gün).

| Olay | Zaman |
|---|---|
| **Pilot partisi (tohum 0, sayılmaz)** | 17 Kasım 2026, 16:28:37 UTC |
| **ortam/sahne.js değişti (parmak ucu düzeltmesi)** | 17 Kasım 2026, 16:38:06 UTC |
| **Ortamın öz-testleri tamamlandı** | 17 Kasım 2026, 16:38:59 UTC |
| **Ön kayıt dondurulduu** | 17 Kasım 2026, 16:39:25 UTC |
| **İlk sayılan koşu başladı** | 17 Kasım 2026, 16:39:42 UTC |
| **Son koşu bitti** | 17 Kasım 2026, 18:03:30 UTC |
| **analiz/analiz.py değişti** | 17 Kasım 2026, 18:06:22 UTC |
| **Kanıt paketi manifesti oluşturuldu** | 17 Kasım 2026, 18:15:46 UTC |

Dosya zamanları dosya sisteminin değişme zamanlarıdır ve kopyalamayla değişebilir; ortam dosyasının ön kayıttaki SHA-256 özetiyle koşular sonrasındaki özeti aynıdır.

## 7.4 Kendini ayırma tanımı

Ön kayıt, kendini ayırmayı «iki figürden hangisinin kendisi olduğunu doğru söyledi ve bir eylem kanıtı gösterdi» diye tanımlar. Önceki sürüm ve analiz çıktısı yalnızca «reflection-tracked» kodunu saydı (Opus 20/20, Sonnet 7/20) ve bunu ön kayıt tanımı olarak sunmadı, ama sapma olarak da adlandırmadı. Bu sürüm iki tanımı birlikte verir. Ön kayıt tanımıyla Opus 18/20, Sonnet 6/20 ($p = 0{,}00024$); çalışma tanımıyla 20/20 ve 7/20 ($p = 1{,}3 \times 10^{-5}$). Ön kayıt tanımı birincil ön kayıtlı KA'dır; çalışma tanımı bir sapmadır. İki tanım aynı sonuca varır, ama eylem kanıtı ölçüsünün kodlayıcı uyumu düşüktür ($\kappa = 0{,}049$); bu yüzden hiçbir tanım güvenilir biçimde «doğru» sayı olarak sunulmaz.

## 7.5 Aşama 1 koşularının tasarım sınırlılıkları

Aşama 1 koşulları (ön kayıtlı Opus ve Sonnet ile ek ön kayıtla eklenen, keşif amaçlı Fable) aşağıdaki tasarım sınırlılıklarını taşır. Bunlar ön kayıt sürecinde dürüstçe kaydedilmiş, ama önceki sürümlerin metnine yeterince taşınmamış noktalardır; sonuçlar bunlarla birlikte okunmalıdır.

**Çalışma yolunda ipuçları.** Modelin çalışma yolunda «ayna», koşul harfi, model adı, tekrar numarası ve «arena» ile «runs» sözcükleri geçiyordu; üst klasörlerde başka koşu ve görev adları görünüyordu. Bu bilgilerin davranışı etkileyip etkilemediği bu veriyle ölçülemez.

**Yönlendirici istem ve Evre 2 sorusu.** Birinci evre istemi («Odayı ve bedenini incele»; `report.md` yazma isteği) modeli bedene ve odaya bakmaya yönlendirir; ikinci evre sorusu da yönlendiricidir («Hangisi sensin? … Kendi bedeninde olağan dışı bir şey var mı?»). Ölçülen davranış bu yüzden tümüyle kendiliğinden değil, yönlendirilmiştir (Ek C).

**Birleşik kodlama.** Özgün kodlama Evre 1, Evre 2 ve rapor dosyalarını birlikte kapsar; kendiliğinden ve yönlendirilmiş davranış ayrılamaz. İki kodlayıcılı kodlama bu ayrımı yalnızca Evre 1 için yapar ve yalnızca Evre 1 sonuçlarını verir.

**Okunabilir kişisel yapılandırma dosyası.** Claude Code'un kişisel yapılandırma dosyası (proje yolları ve hesap bilgisi içerir) Aşama 1 modellerine okunabilir durumdaydı; bu, sonraki aşama için yapılan altyapı sızıntı denetiminin dördüncü turunda fark edildi. Aşama 1 verisi değiştirilmedi; dosyadaki bilginin sonuçlara etkisi bu veriyle belirlenemez.

**Model bağlamındaki e-posta adresi.** Claude Code, hesabın e-posta adresini modele giden isteğe ekler: Opus ve Sonnet'in 40/40 koşusunun ve Fable'ın 20/20 koşusunun kaydında, modele giden istekte bir e-posta adresi geçer. Farklı değer sayısı Opus ve Sonnet koşularında bir, Fable koşularında iki'dir ve Fable değerleri Opus ve Sonnet değerinden farklıdır (değerler yazılmaz). Bu bilginin davranışa etkisi bu veriyle belirlenemez.

**Test adı ve eğitim verisi.** Opus'un 5/20 koşusunda ajan, istem testi adlandırmadığı hâlde testin adını kendiliğinden yazdı (Sonnet 0/20; düşünme özetleri yalnızca Opus kayıtlarında dolu olduğundan bu bir alt sınırdır); bu, çalışma yolundaki «ayna» sözcüğünden ayrıştırılamaz (Bölüm 4.3). Modeller ayna testini eğitim verisinden de biliyor olabilir. Ölçülen şey işlevsel öz-yönelimdir; bilinç iddiası yoktur.

## 7.6 Fable ekinin sapmaları ve farkları

**Post hoc ek ve sıra.** Ek ön kayıt, ana ön kayıttan ve Opus ile Sonnet sonuçlarının görülmesinden sonra dondurulduğu. İlk Fable koşusu donmadan 17 saniye sonra başladı; sıra «sonra»dır, ancak aralık çok dardır. Donmadan önce başlayan Fable koşusu yoktur; donmadan önce yapılan 8 kayıtlık kuru çalıştırma sahte yarışmacıyla yapıldı ve sonuçlara girmedi.

**Claude Code sürümü ve koşu günü.** Fable koşuları 2.1.330 sürümüyle 21 Kasım 2026 günü, Opus ve Sonnet koşuları 2.1.326 sürümüyle 17 Kasım 2026 günü yürütüldü. «Aynı Claude Code sürümü» ifadesi yalnızca Opus ve Sonnet için doğrudur; sürüm farkının ve servis tarafındaki olası değişikliklerin sonucu etkileyip etkilemediği ölçülemez.

**Koşturucu yeniden kuruldu.** Opus ve Sonnet koşularını yürüten özgün koşturucunun kaynak dosyası diskte bulunmadığından, Fable koşturucusu eldeki bir kopyadan ve kayıtlardan çıkarılan davranıştan yeniden kuruldu. Eşdeğerlik, yeniden kurulan ortamın Opus ve Sonnet koşularının kayıtlı komut dizilerini aynı kareler ve ölçümlerle yeniden üretmesiyle gösterildi; eksik kaynağın tam eşdeğerliği kanıtlanamaz ve bu doğrulama bu makale için yeniden çalıştırılmadı.

**`task.json` özeti.** Fable koşularının `task.json` özetleri ana ön kayıttakilerle 20/20 koşuda farklıdır: dosya yeniden yazıldı ve özgün metni elde yoktur. Ek ön kayıt listesindeki özetle fark yoktur (0). İstem, başlangıç klasörü ve puanlayıcı özetleri aynıdır. Koşturucu yapılandırma dosyasının ( `yarismacilar.json` ) özeti de ana ön

kayıttakinden farklıdır; bu dosya yalnızca koşunun bitiş ve hata sınıflamasını etkiler, modele giden hiçbir şeyi etkilemez.

**Tek bölmeli ve iki bölmeli partiler.** Fable partileri tek bölmeli, Opus ve Sonnet partileri yan yana iki bölmeli koştu; modelin gördüğü bölme boyutu aynıdır, ama makine yükü ve süre ölçümü bu farktan etkilenir. Süre kıyası bu notla okunmalıdır.

**Karıştırıcı etkenlerin toplamı.** Model, sürüm, koşu günü, bölme düzeni ve koşturucu birlikte değiştiğinden, Fable ile Opus ya da Sonnet arasındaki her fark bu etkenlerin toplamıdır.

## 7.7 Klasör dışı geçici dosya bayrağı taşıyan koşular

40 koşunun 14 tanesi klasör dışı geçici dosya bayrağı taşıyor (Opus 13, Sonnet 1; Tablo 20). Anlamı: modelin kendi araç çağrıları çalışma klasörünün dışında bir geçici yola (örneğin görüntü birleştirme ya da kırpma için) yazma ya da okuma girişimi içeriyor; istem klasör dışına okuma ve yazmayı yasaklamıştı. Bağımsız tarama kayıttaki bayraklarla aynı sonucu verdi (uyuşmayan koşu: 0). Ön kayıtta bayraklı koşular için bir dışlama kuralı yoktur; bu yüzden koşular dışlanmadı. Hiçbiri geçersiz sayılmadı (geçersiz puanlı koşu: 0). Bayraklı dosyaların bakış sayısını ya da puanlayıcı ölçümünü değiştirmediği kaydedilmiştir, ancak ajanın yorumuna etkisi (örneğin kareleri kırpıp birleştirerek daha yakından incelemesi) dışlanamaz.

Bu nedenle duyarlılık analizi Tablo 19 içindedir. Bayraklılar çıkarılınca Opus H1 p değeri 0,008 değerinden 0,333 değerine çıkar (yön aynıdır); ancak 2 ve 2 koşuyla ulaşılabilecek en küçük p 0,333 olduğundan bu bir kırılganlık değil, bir güç tabanıdır; analiz Opus H1'i sınamaz. Bayraklı tohum çiftleri iki modelde birlikte çıkarılınca kendini ayırma farkı 0,061 ile sınırda kalır (bu bir eşli test değil, çıkarılmış bir alt kümede eşsiz Fisher'dir). Opus'un A koşulundaki 5 koşusundan 3 tanesi bayraklıdır; bayraklı üç koşunun oturum kayıtları bu amaçla ayrıca tarandı: ilk işaret sözünden (asistan metni, düşünme özeti ya da rapor ve yanıt yazımı) önce bir karenin yakınlaştırılmış kırpımının üretilip görüntülendiği koşu 2/3'dür (üçüncüsünde yalnızca küçültülmüş kare ızgaraları ve bir piksel farkı denetimi görülür); bayraksız iki Opus A koşusunda bu sayı 0'dır. Bu bir sıralama bulgusudur ve işaretin kırpmayla bulunduğunu göstermez: ilk söz yazılı metindedir ve düşünme özetleri çoğunlukla boştur; öte yandan bayraksız koşularda işaret kırpma olmadan raporlanmıştır, yani kırpma işareti bulmak için

gerekli görünmez. Bayraklı koşular Opus'ta yoğundur (13/20), Sonnet'te neredeyse yoktur (1/20); modeller arası karşılaştırma bu dengesizlikten etkilenmiş olabilir.

## 7.8 Diğer notlar

**Claude Code sürümü.** Koşu günü sabitlenen sürüm 2.1.326 idi; koşucu arenanın sabitlemesi ise 2.1.325 sürümünü içeriyordu ve bu yüzden 40/40 koşu «sürüm kayması» bayrağını eşit biçimde taşır. Opus ve Sonnet için sürüm aynıdır; Fable farklı bir sürümle koşuldu (Bölüm 7.6).

**Maliyet farkı.** Opus toplamı için günlük × liste ile Claude Code raporu arasındaki 0,04 $ fark tek bir koşudan gelir (Bölüm 3.8).

**Toplam oran.** Önceki sürümdeki «7,0 kat» yuvarlama artefaktıydı; kesin oran 7,05 (günlük) ve 7,06 (raporlanan) olduğundan «yaklaşık yedi kat» demek doğrudur.

**Ayna ve kendini görme.** Önceki sürümün «her ajan aynada kendini gördü» cümlesi yalnızca aynalı 30 koşu için doğrudur (30/30); B koşulunda ayna yoktur.

**İşaretin görünürlüğü.** Öz-testte bir tohumda işaret, aynaya yaklaşan bedenin karesinde yaklaşık bir piksel kaplıyordu (Bölüm 3.9); bu, o tohumun A koşulundaki sonuçlarını etkilemiş olabilir.

**Sıra ve eşzamanlılık.** Her grupta bir Opus ve bir Sonnet koşusu aynı saniyede başladı; bu, ortam aynılığı anlamında eşli tohum tasarımını bozmaz ama koşuların zaman bakımından tam bağımsız olmadığı anlamına gelir.

## 7.9 Sınırlılıklar

**Küçük örneklem:** Hücre başına 5 koşu vardır; oranlar geniş aralıklarla okunmalıdır. Beş ve beş koşuyla ulaşılabilecek en küçük eşsiz p değeri 0,0079'dur. Bayraklı koşular çıkarılınca Opus H1 örneklemi 2 ve 2 koşuya iner ve p değeri 0,333'nin altına inemez; kendini ayırma karşılaştırması bayraklı tohum çiftleri çıkarılınca sınırda kalır.

**Kodlayıcı uyumu:** Ham uyum yüzdesi ile Cohen κ birlikte verilir (Tablo 18). Kendine işaret uyumu yüksektir (%92,5 ham uyum, κ = 0,754); ötekine işaret ölçüsünde ham uyum yüksek görünse de (%82,5) κ = 0,271'tür; iki kodlayıcı çoğunlukla «yok» demiş, işaretin ötekinde olduğu D koşulunda kural kodlayıcı koşulları kaçırmıştır. Eylem kanıtında ham uyum %27,5, κ = 0,049'tür; kural kodlayıcı

yalnızca yanıt dosyasına, dil modeli kodlayıcı rapor ve yanıt dosyalarına baktığı için uyuşmazlığın bir kısmı tasarım farkıdır. Düşük κ, bu iki ölçüye dayalı sonuçların (D koşulunda ötekine atfetme, ön kayıt tanımıyla kendini ayırma) kodlayıcıya bağlı olduğu anlamına gelir. Kendini ayırma seçiminin ikinci kodlayıcısı ve insan kodlaması yoktur.

**Test adı ve eğitim verisi:** Opus'un 5/20 koşusunda ajan, istem testi adlandırmadığı hâlde testin adını kendiliğinden yazdı (Sonnet 0/20; düşünme özetleri yalnızca Opus kayıtlarında dolu olduğundan bu bir alt sınırdır) ve bu, çalışma yolundaki ipuçlarından ayrıştırılamaz (Bölüm 4.3). Modeller ayna testini eğitim verisinden de biliyor olabilir.

**Sanal beden:** Temas bildirimi yalnızca «bir şeye değiyor» cümlesidir; hayvan ve insan sürümüyle birebir aynı değildir.

**Tek efor:** Yalnızca medium efor ölçüldü (Opus ve Sonnet için Claude Code 2.1.326, Fable için 2.1.330); başka efor seviyeleri sonucu değiştirebilir.

**Aşama 1 tasarımı:** Yönlendirici istem ve Evre 2 sorusu, çalışma yolundaki ipuçları, birleşik kodlama ve okunabilir kişisel yapılandırma dosyası sonuçları sınırlar (Bölüm 7.5).

**Düzeltilmemiş p değerleri:** Makalede bildirilen p değerleri çoklu karşılaştırma için düzeltilmemiştir; yalnızca Fable bölümündeki yedi keşif amaçlı karşılaştırma için Holm düzeltmeli değerler ayrıca verilir (Bölüm 5.1) ve hiçbiri anlamlılık sınırının altında kalmaz. Keşif amaçlı Fable karşılaştırmalarındaki küçük p değerleri (örneğin 0,010) kesin sayılmaz.

**Üçüncü model:** Fable 5.2 post hoc bir ekle, ayrı bir günde, farklı bir sürümle ve yeniden kurulmuş bir koşturucuyla koşuldu; özgün kodlaması yoktur. Sonuçları yalnızca keşif amaçlıdır (Bölüm 5, Bölüm 7.6).

**Maliyet:** Dolar, token'ların resmî liste fiyatıyla çarpımıdır; abonelikle çalışan bir kullanıcının faturası değildir. Fiyat sayfasında son güncelleme tarihi yoktur. Başarı başına maliyet çözüm maliyeti değildir; ilk doğru atfa kadar harcanan maliyet hesaplanmadı. Fable ile Opus toplamı bu verilerle ayırt edilemedi (Tablo 9).

**Literatür doğrulaması (22 Kasım 2026, `curl` ve tarayıcı kullanıcı aracısıyla; üç kaynak 23 Kasım 2026'da eklendi):** Otomatik istemciye şu yayıncıların DOI sayfaları otomatik istemciye erişimi engelledi (bot engeli): Science, Wiley (iki kayıt), Taylor & Francis,

Oxford Academic, SAGE, PNAS ve BMJ; Springer ve JSTOR sayfaları bot sınaması gösterdi (içerik yok); AAAI sayfası açıldı. DOI kayıtlarının hepsi doi.org tutamak kaydında bulundu; künyeler Crossref, PubMed, arXiv ve bildiri sayfalarıyla doğrulandı. Yates (1934) tam metin okundu; MirrorBench ve Ziliotto ve ark. için PDF'in ilgili bölümleri ve tabloları okundu; diğerleri için yalnızca özet ya da künye doğrulandı (Gallup 1970, Amsterdam 1972, Wilson, Cohen, Landis ve Koch, Upton, Newcombe, Feinstein ve Cicchetti 1990, Heyes ve Gallup ve Anderson 2020 için içerik okunmadı). Fisher'in kendi makalesi kesin testin kaynağı olarak doğrulanamadığı için kullanılmadı.

**Öncelik iddiası yoktur:** Bu çalışmanın bulabildiğimiz tek örnek olduğu ifadesi tek bir web aramasına dayanır; arama motoru ve sorgu metni kayıt altına alınmamıştır (Bölüm 2).

**Kodlayıcı kimliği:** Dil modeli kodlayıcının (ve Evre 1'in iki kodlayıcı oturumunun) kimliği ve sürümü kayıtlı değildir; kodlayıcı ajanlarla aynı model ailesinden olabilir ve bu durumda öz-kayırma yanlılığı mümkündür (Bölüm 3.7).

**Bağlaşım algılaması ve öteki figürün hareketi:** Kendini ayırma büyük ölçüde eylem ile yansıma arasındaki bağlaşımı denetlemeye dayanır; öteki figürün pozu ise ajanın eylem sayısına zamansal olarak bağlıdır ve eylem sayısından bağımsız bir figür sınanmamıştır (Bölüm 6).

**Erken durma:** Sonnet'in koşu başına en çok 81 eylemi vardır ve Opus'un ortalamasına (145,2) hiçbir koşuda ulaşmamıştır; Sonnet'in düşük İB ve KA değerleri kısmen oturum uzunluğu farkını yansıtıyor olabilir ve eylem sayısıyla sonuç ayrıştırılamaz.

**Sözel rapor:** Dokunma dışındaki ölçüler ajanın yazdığına dayanır; Opus'un 4 raporu işarete dokunduğunu yazarken ölçülen en yakın mesafe 4,03 cm'dir, yani rapor davranışın güvenilir bir kılavuzu değildir.

**Harness açıklama boşlukları:** Oturum kayıtlarındaki 1.264 düşünme bloğunun yalnızca 112 tanesinin metni doludur, bu yüzden ajanların gerekçesi büyük ölçüde görülemez; Opus ve Sonnet ile Fable farklı günlerde koşuldu (modeller günlere serpiştirilmedi); özgün koşturucunun kaynağı kayıptır ve eşdeğerlik bu makale için yeniden çalıştırılmadı. Okunabilir kişisel yapılandırma dosyasının adı, Opus ve Sonnet'in 40 koşusundan 0 tanesinde araç çağrılarında ya da sonuçlarında geçti; bu, etkiyi sınırlar ama sıfır olduğunu kanıtlamaz.

**Zaman damgası:** Ön kaydın zaman damgası kendi beyanımızdır ve üçüncü bir taraf tarafından tutulmamıştır; ilk koşunun donmadan kısa süre sonra başlaması bir betik adımıdır (Bölüm 3.4).

# 8 Etik ve veri ile kodun erişilebilirliği

Çalışmada insan ya da hayvan katılımcı yoktur; özneler yapay zeka ajanlarıdır. Koşu görüntülerinde kişisel veri bulunmadığı kaydedildi (0 koşuda temiz olmayan görüntü, 0 karantina kaydı); bu denetim yalnızca görüntüleri kapsar. Oturum kayıtlarının (metin) kişisel dosya yolları maskelenmiştir; ancak oturum kayıtları hesap ve kuruluş tanımlayıcıları ile bir e-posta adresi içerir ve bunlar henüz temizlenmemiştir. Ajanların web erişimi kapalıydı (web aracı kullanımı: 0). Aynı kayıtlar Fable koşuları için de tutuldu (0 koşuda temiz olmayan görüntü, 0 karantina kaydı, web aracı kullanımı 0).

**Veri ve kod erişilebilirliği.** Opus ve Sonnet için ortamın kodu, istemler, puanlayıcı, analiz ve maliyet betikleri, öz-test sonuçları, ön kayıt ve her koşunun kanıtı (ortam kaydı, ajanın gördüğü her kare, `report.md` ve `answer.md`, puanlayıcı sonucu, Claude Code ekran kaydı, tekrar videosu ve oturum kaydı) tek bir kanıt paketinde toplanmıştır. Paketin manifest dosyası her dosyanın SHA-256 özetini listeler: 1.900/1.900 girdi diskteki dosyalarla eşleşir (1.873 paket içi, 27 paket dışı); manifest, değişmiş `analiz/analiz.py` sürümünün özetini içerir (Bölüm 7). **Kanıt paketi şu anda okurlara açık değildir: oturum kayıtlarındaki hesap tanımlayıcıları ve e-posta adresi temizlenmeden paket paylaşılmaz ya da arşive konmaz; maskelenmemiş ham oturum kayıtları ise depo dışında kalır ve paylaşılmaz. Bu makale için DOI atanmamıştır ve kod ile kanıt için herkese açık bir adres yoktur.** Bu makaleyle birlikte sunulanlar, makalenin metni (Türkçe ve İngilizce), PDF dosyaları, koşu tablosu (Ek A), ön kayıt metinleri ve SHA-256 özetleri (Ek C ve Ek D), BibTeX ve CITATION.cff dosyalarıdır.

Bu makaledeki her sayı, ham koşu dosyalarını ve kodlama çıktılarını yeniden okuyan bir yeniden hesap betiğinin çıktısından gelir (1.549 miktar, her biri anahtarıyla listelidir); Fable koşuları için aynı tanımlarla ayrı bir yeniden hesap vardır (1.348 miktar); Evre 1 kodlaması ile inceleme sonrası eklenen miktarlar (test adını anma, kodlama karşılaştırması, çapraz kontroller) ayrı dosyalardan okunur. Yeniden hesaplar, kanıt paketi manifestinin oluşturulmasından sonra, aynı ekip tarafından yazılmıştır; manifestte yer almaz ve eski çıktıdan bağımsız yazılıp yazılmadıkları kayıtlı değildir. Opus ve Sonnet

koşularını yürüten özgün koşturucunun kaynak dosyası kayıptır; yalnızca özetleri kalmıştır (Bölüm 7.6). Fable koşularının kanıtı da paketin dışında ayrı bir klasörde durur ve paylaşılmaz; bu klasör için bir manifest yoktur ve paket paylaşılacak olursa Fable klasörü için de bir manifest üretilmesi gerekir. Makale veri dosyasını yeniden hesap çıktısıyla, atıfları doğrulanmış kaynak kayıtlarıyla ve sayfa stilini renk kurallarıyla denetleyen üç doğrulama betiği yayın sürecinin parçasıdır; betikler özel bir depodadır ve herkese açık değildir.

**Yazarlık, lisans ve yapay zekâ kullanımı.** Yazar bir kurumdur (ConectLens Research); bu sürümde sorumlu bir kişi adı yazılmamıştır. Lisans: kod ve betikler MIT Lisansı ile sunulur; bu makalenin metni, veri tabloları ve şekilleri de çalışmanın bütünü için seçilen MIT Lisansı ile sunulur. Üçüncü taraf bileşenlerin (örneğin Three.js) kendi lisansları geçerlidir. Yapay zekâ yardımı: ajan kayıtlarının kodlanmasında (dil modeli kodlayıcılar), İngilizce çeviride (Codex CLI ile çeviri ve ayrı oturumlarda inceleme) ve metin taslağının yazımında dil modelleri kullanıldı. Yayımlanmış bir sürüm yerinde değiştirilmez; düzeltmeler değişiklik günlüğü satırıyla birlikte yeni bir sürüm olarak yapılır.

# 9 Sonuç

Ön kayıtlı birincil ölçüt olan dokunma, Opus ve Sonnet'in hiçbir koşusunda gerçekleşmedi (0/40); ölçüt tabandadır. İşaretin doğru figüre ve bölgeye sözel atfını sayan işaret başarısı, ayna ve işaret koşulunda Opus'ta 5/5, aynasız koşulda 0/5 bulundu (eşsiz Fisher p = 0,008; ön kayıttaki eşli analiz kullanılmadı ve eşli kesin McNemar p = 0,0625); bu karşılaştırma ağırlıklı olarak bir görsel erişim kontrolüdür. Bayraklı koşular çıkarılınca kalan 2 ve 2 koşuyla ulaşılabilecek en küçük p 0,333'tür; bu analiz bu yüzden bir sonuç vermez. Sonnet'te sonuç belirsizdir (2/5 ve 0/5). Kendini öteki figürden ayırmada, veri görüldükten sonra tanımlanan aynalı koşullar kümesinde Opus 15/15, Sonnet 6/15 bulundu; toplamlar iki kodlamada aynıdır, ama koşu bazında uyum tam değildir ve sayılar büyük ölçüde bağlaşım algılamasına dayanır; daha geniş sayılar tanıma ve kodlayıcıya bağlıdır ve kodlayıcı güvenilirliği sınırlıdır. Maliyet yaklaşık yedi katlık, sağlam bir fark gösterir (7,05); başarı başına maliyet farkı 2,44 kattır ve çözüm maliyeti değildir.

Keşif amaçlı üçüncü model Fable 5.2, parmak ucunu işaret noktasının 4 cm'i içine 4/20 koşuda getirdi; bu yüzden «hiçbir model dokunmadı» ifadesi üç model için doğru değildir, yalnızca Opus ve Sonnet için

doğrudur. Ortak tanımla işaret başarısı (İB*) A koşulunda Fable ile Opus arasında ayırt edilemez; kendini ayırmada (yalnızca Evre 1 kodlaması) aynalı koşullarda Fable Opus ile aynıdır (15/15); toplam maliyeti Opus'un 0,93 katıdır ve bu verilerle Opus'tan ayırt edilemedi (%95 önyükleme aralığı 0,75–1,15); token başına liste fiyatı farkı toplam maliyette gösterilemedi. Bu bulgular keşif amaçlıdır ve betimleyicidir; post hoc bir eke, ayrı bir güne, farklı bir sürüme ve düzeltilmemiş p değerlerine bağlıdır. Sonuçlar küçük örnekleme, çok sayıda sınırlılığa ve bu metinde açıkça listelenen sapmalara bağlıdır; bilinç ya da öz-farkındalık iddiası taşımaz.

## 10 Sonraki aşama

Aralarında iki Codex modelinin bulunduğu beş modelin katıldığı, ajanlara teste ilişkin önceden hiçbir bilgi verilmeyen bir ikinci aşama ve ortak bir odanın kullanıldığı üçüncü bir aşama planlanmaktadır; ikisi de koşulmamıştır ve bu makalede onlara ait bir sonuç ya da iddia yoktur.

## Teşekkür ve beyanlar

Teşekkür yoktur.

Yazışma ve düzeltmeler: bu makalenin yayıncı sayfası üzerinden.

Bu sürüm için finansman ya da çıkar çatışması beyanı sağlanmamıştır.

## Değişiklik günlüğü

Önceki sürümler v1 (18 Kasım 2026) ve v2 (21 Kasım 2026) tarihlidir. v1'in içeriği kurtarılamaz; v2 arşivlenmiş bir taslak sürüm olarak durur ve v3 ile değiştirilmiştir: v2, bu sürümde düzeltilen ifadeler içerir (aşağıya bakınız). Bu sürüm (v3), keşif amaçlı üçüncü modelin eklenmesi, doğrulanmış olgularla yapılan düzeltmeler ve bir inceleme kurulunun istediği düzeltmeler için hazırlanmıştır; yayımlanmış bir sürüm düzeltmeyle değil, yeni bir sürümle değiştirilir. Opus ve Sonnet'in sayıları v2 ile birebir aynıdır; aşağıdaki değişiklikler yalnızca ekleme, ifade düzeltmeleri ve ek hesaplardır.

**v3: eklenenler**

> Claude Fable 5.2, ana ön kayıttan sonra donmuş bir ek ön kayıtla, keşif amaçlı üçüncü model olarak eklendi (Bölüm 5, Bölüm 3.5).

Fable sonuçları ön kayıtlı H1, H2 ve H3'e karışmaz; ek ön kayıt Ek D'dedir (Ek D) ve ana ön kayıt (Ek C) değişmemiştir. Fable koşuları Ek A'dadır.

Üç modelin ortak karşılaştırması yalnızca dokunma, işaret başarısı* (İB*), Evre 1 kendini ayırma, maliyet ve bakış sayısıyla yapılır. Fable için özgün kodlama olmadığından «yirmi koşunun yirmisinde» gibi bir sayı yazılmaz.

Maliyet: Fable'ın toplamı, koşu başına maliyeti, bakış ve eylem sayıları, Opus ve Sonnet'e oranları, önbellek yazma payı ve liste fiyatı notu eklendi.

Kendini ayırma: üç tanım açıkça yazıldı ve hangisinin ön kayıtlı olduğu belirtildi; ayna olan koşullardaki sayı (Opus 15/15, Sonnet 6/15) veri görüldükten sonra tanımlanan (post hoc) küme olarak eklendi; B koşulunda ayna olmadığı açıklandı (Tablo 6).

Sonnet'in öteki figürü kendi sandığı koşu sayısının tek kodlayıcıya göre olduğu ve iki kodlayıcılı Evre 1 kodlamasında farklı çıktığı belirtildi.

Aşama 1 tasarım sınırlılıkları eklendi: çalışma yolundaki ipuçları, yönlendirici istem ve Evre 2 sorusu, birleşik kodlama, okunabilir kişisel yapılandırma dosyası (Bölüm 7.5). Fable ekinin sapmaları (sürüm ve koşu günü, yeniden kurulan koşturucu, `task.json` özeti, tek bölmeli partiler) eklendi (Bölüm 7.6).

Düzeltilmemiş p değerleri, küçük örneklem ve Fable karşılaştırmalarının kesin sayılmayacağı sınırlılıklara yazıldı.

Sonraki aşama için kısa bir not eklendi; planlanan aşamaların sonucu ya da iddiası yoktur (Bölüm 10).

Makalenin İngilizce çevirisi eklendi: Türkçe metin özgündür ve kanoniktir. Çeviri Codex CLI ile yapıldı ve her parti ayrı bir Codex oturumunda incelenip onaylandı; sayılar aynı miktarlardır ve İngilizce biçimle (ondalık nokta, binlik virgül) yazılır.

**v3: düzeltilen ifadeler**

«Hiçbir model dokunmadı (0/40)» ifadesi, Fable ile yanlış olurdu: doğrusu, Opus ve Sonnet'in 40 koşusunda sıfır, Fable'ın 20 koşusunda 4 dokunmadır.

«Opus yirmi koşunun yirmisinde kendini doğru buldu, Sonnet yedisinde» ifadesi tanımı yazmıyordu ve B koşulunda ayna yoktur; ayna olan koşullarda Opus 15/15, Sonnet 6/15 olarak yazıldı.

«Aynı Claude Code sürümü» ifadesi model bazında yazıldı: Opus ve Sonnet 2.1.326, Fable 2.1.330.

Gallup (1970) cümlesine boyanın tahriş etmeyen olduğu eklendi; MirrorBench, Ziliotto ve arkadaşları ile Hart ve Scassellati için v2'de yapılan düzeltmeler (görev başarı oranı ile bileşik skor ayrımı, insan ajan, yorum notu, cilt, sayfa ve DOI) korundu.

Atıf denetimi: kaynakçaya beş kaynak eklendi (Newcombe-Wilson aralığı, Fisher kesin testi, küçük değerlendirme kümelerinde belirsizlik, model adı gizli kodlama ve model adları) ve ilgili cümlelere bağlandı; Dellibarda ve arkadaşları, Brown ve arkadaşları, Landis ve Koch ile ön kayıt cümleleri kaynakların desteklediği kapsama göre netleştirildi; kanıt paketinin erişilebilirliği doğru biçimde yazıldı (Bölüm 8).

**v3: inceleme sonrası düzeltmeler**

İstatistik: ön kayıtlı eşli analizin kullanılmadığı bildirildi ve eşli kesin McNemar p eklendi (Opus 0,0625, eşsiz 0,0079); H2'nin H1 ile aynı tabloyu kullandığı ve bayraklı koşular çıkarılınca kalan 2 ve 2 koşuyla en küçük p'nin 0,333 olduğu yazıldı; «anlamlılık kaybolur» ifadesi kaldırıldı. Taban etkisi ve sürekli en yakın mesafe sonuçları eklendi.

Maliyet: Fable toplamının Opus toplamından düşük olduğu ifadesi her yerden kaldırıldı; oran 0,93, %95 önyükleme aralığı 0,75–1,15 (bir değerini içerir) ve koşu başına standart sapmalar eklendi (Tablo 9).

Sapmalar: kayıtlı metin, yapılan, olası yanlılık yönü ve etkilenen sayıları gösteren sapma tablosu eklendi (Tablo 14); kodlayıcı değişikliği sapma olarak yazıldı.

Değerlendirme yöntemi: A ile B karşılaştırması görsel erişim kontrolü olarak yeniden çerçevelendi; Opus'un testin adını kendiliğinden anan 5 koşusu ve bunlar çıkarılınca duyarlılık eklendi (Bölüm 4.3); Ziliotto ve ark. koşullarıyla ilgili yenilik ifadesi düzeltildi; öncelik iddiası yumuşatıldı; özet kısaltıldı.

Bilişsel bilim: Gallup paradigmına göre korunanlar ve kaybolanlar bölümü (Bölüm 3.11) ve öteki figürün hareketinin eylem sayısına bağlılığı eklendi; «sandı» türü ifadeler rapor temelli ifadelerle değiştirildi; piksel ölçütü sütunları yeniden adlandırıldı; başlık değişti; Anderson ve Gallup (2015) tanımı düzeltildi, Heyes (1994) ve Gallup ve Anderson (2020) eklendi (yalnızca künyeleri doğrulandı).

Tekrarlanabilirlik ve beyanlar: MIT Lisansı belirtildi; ön kayıt zaman damgasının kendi beyanımız olduğu ve ana ön kayıt dosyasının SHA-256 özeti yazıldı; kanıt paketinin şu anda okurlara açık olmadığı ve oturum kayıtlarının tanımlayıcı veri içerdiği doğru biçimde yazıldı; atıf adresi sürüm adresiyle tutarlı hâle getirildi.

Yayıncılık: Türkçe ve İngilizce PDF'ler tek bir derlemeden üretildi; «kör» ifadesi yerine model adlarının gizlendiği ve kodlayıcı kimliğinin kaydedilmediği yazıldı; ayna olan koşullar kümesinin veri görüldükten sonra tanımlandığı her yerde belirtildi; Feinstein ve Cicchetti (1990) eklendi; otomatik puanlayıcı baştan sona «puanlayıcı» olarak adlandırıldı; erişilemeyen yayıncı listesi denetim raporuna uygun yazıldı.

**v3: ikinci inceleme turundan sonraki düzeltmeler**

İstatistik: Opus H1 için tek yönlü kesin eşli p ve A ile B'nin tohumla eşleşmesinin nasıl doğrulandığı eklendi; test adını anan koşular çıkarılınca eşli McNemar p verildi ve eşsiz p betimleyici olarak etiketlendi; Fable ve Sonnet İB* için kırılganlık indeksi ve Fable bölümündeki karşılaştırmalar için Holm düzeltmeli p değerleri eklendi; Evre 1 κ'sının yanına Gwet AC1 eklendi.

Değerlendirme yöntemi: özet kısaltıldı ve kodlayıcı uyumu ile eşli test sayıları gövdeye bırakıldı; «maliyet bakımından ayırt edilemez» ifadesi «bu verilerle ayırt edilemedi» olarak yumuşatıldı; Tablo 9 içindeki önbellek yazma notu düzeltildi ve beş dakikalık satıra aralık eklendi; bayraklı Opus A koşularında kırpma incelemesinin sonucu yazıldı; test adı sınırlılığı olguyla başlatıldı ve Sonnet sıfırının bir alt sınır olduğu yazıldı.

Bilişsel bilim: ifadeler rapor ve temas diline yaklaştırıldı («temas bildirimi», «işaretin göründüğü yansıma», «oturum uzunluğu» gibi); Gallup ve Anderson (2020) betimi düzeltildi; Opus'un elini başına götürmesi için post hoc cümlesi özete eklendi; aynasız koşulun Gallup tasarımıyla ilişkisi açıklandı.

Tekrarlanabilirlik: ön kayıt özeti kopyalamaya elverişli biçimde ayrı satıra alındı; hesap e-posta adresinin modele giden istekte geçtiği yazıldı; Fable manifesti notu eklendi; başlıktaki tarih satırı sürüm ve tarih olarak yazıldı.

Yayıncılık: yöntem bölümündeki istemler özgün İngilizce metniyle birebir alıntılandı; Fable kodlamasının betimi model adlarının gizli olduğunu belirtecek biçimde düzeltildi; Feinstein ve Cicchetti (1990) yalnızca özet düzeyinde okunanlar listesine eklendi; İngilizce

metindeki çevrilmemiş öğeler, sayı biçimleri ve terim tutarlılığı düzeltildi; tablolarda sözcük içinden bölünme giderildi.

Beyanlar: yazışma ve düzeltmelere ilişkin cümle ile finansman ve çıkar çatışması beyanının sağlanmadığına ilişkin cümle eklendi.

**v2: eklenenler ve düzeltmeler (arşiv; v3 ile değiştirildi, v3'te düzeltilen ifadeler içerir)**

«Analiz betikleri koşulardan önce dondurulduu» ifadesi yanlıştı; `analiz/analiz.py` koşulardan sonra güncellendi (Bölüm 7.2).

«Pilottan sonra ortamda hiçbir değişiklik yapılmadı» ifadesi yanlıştı; parmak ucu düzeltmesi pilottan sonra ve ön kayıttan önce yapıldı (Bölüm 7.3).

Kendini ayırma yalnızca çalışma tanımıyla veriliyordu; ön kayıt tanımı (eylem kanıtı dahil) eklendi ve hangisinin ön kayıtlı olduğu belirtildi (Bölüm 7.4).

14 koşunun klasör dışı geçici dosya bayrağı taşıdığı hiç anılmıyordu; duyarlılık analiziyle eklendi ve Opus p değerinin dışlamaya duyarlı olduğu belirtildi (Bölüm 7.7).

Maliyet: günlük × liste ile Claude Code raporu arasındaki 0,04 $ fark tek koşudan geldiği için not düşüldü; toplam oran «7,0 kat» yerine 7,05 (günlük) ve 7,06 (raporlanan) olarak yazıldı; Opus koşu tablosunda bu koşunun iki değeri verildi.

«Her ajan aynada kendini gördü» yalnızca aynalı 30 koşu için doğrudur (30/30); B koşulunda ayna yoktur.

Kodlayıcı uyumu ham yüzdeyle yetinmeyip Cohen κ ile verildi; düşük κ sınırlılık olarak tartıldı.

MirrorBench'in 0,170 değeri yalnızca Genel görev başarı oranı içindir; makalenin kendi dört metrikli sıralamasında lider farklıdır. «İnsan katılımcılar» ifadesi «referans insan ajanı» olarak düzeltildi.

Gallup (1970): işaret alına değil, anestezi altındaki şempanzenin kaşının üstüne ve karşı kulağına sürüldü; ikincil özete dayandığı belirtildi.

Ziliotto ve arkadaşlarının aynasız koşuldaki artık başarıyı öz-tanıma değil kısayol ya da tahmin olarak yorumladığı eklendi.

«Medium Claude Code varsayılanı» ifadesi, medium düzeyinin bu iki model için belgelenmiş varsayılan efor olduğu biçiminde düzeltildi.

«İlk kez ölçülmektedir» ifadesi «bildiğimiz kadarıyla» olarak yumuşatıldı ve arama kaynağı yazıldı.

Fisher kesin testi için kaynak olarak Yates (1934) verildi; Hart ve Scassellati (2012) künyesi tam hâline getirildi.

Standart ön baskı yapısı, numaralı ve DOI bağlantılı kaynakça, nasıl atıf yapılır kutusu, BibTeX ve CITATION.cff, Ek A (koşu tablosu), Ek B (kodlayıcı uyumu ve duyarlılık), Ek C (ön kayıt metni) ve her sayıyı yeniden hesap çıktısındaki anahtarına bağlayan doğrulama betikleri eklendi.

**Değişmeyenler.** Ön kayıtlı H1, H2 ve H3 sınamaları, Opus ve Sonnet için birincil ölçüt, işaret başarısı hücreleri, Fisher p değerleri, Wilson aralıkları, kendini ayırma sayıları ve maliyet rakamları v2 ile birebir aynıdır; Ek C (ön kayıt metni) kaynak dosyayla aynıdır.

# Kaynakça


[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 yazarları (2026). three, sürüm 0.187.0. npm kayıt defteri. https://registry.npmjs.org/three/0.187.0 (Erişim: 22 Kasım 2026).

[12] Anthropic (2026). @anthropic-ai/claude-code, sürümler 2.1.326 ve 2.1.330. npm kayıt defteri. https://registry.npmjs.org/@anthropic-ai/claude-code (Erişim: 22 Kasım 2026).

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

[14] Anthropic (2026). Model configuration (Claude Code documentation). Anthropic, code.claude.com. https://code.claude.com/docs/en/model-config (Erişim: 22 Kasım 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 (Erişim: 22 Kasım 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 (Erişim: 22 Kasım 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 $\chi^2$ Test. *Supplement to the Journal of the Royal Statistical Society (şimdiki adıyla 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

## Bu makaleye atıf

ConectLens Research. (2026). *Aynadaki ajan: Claude Opus 5.5, Sonnet 5.5 ve (keşif amaçlı) Fable 5.2 ile bir kodlama ajanında ayna-işaret testi uyarlaması ve maliyeti* (Sürüm 3) [Ön baskı]. Vibe, ConectLens. https://vibe.conectlens.com/research/mirror-test/v/v3

# Ek A. Koşu tablosu

Her satır tek bir koşudur. Değerler yeniden hesap çıktısından (`kosu.<koşu>.<alan>` anahtarları) okunur.

**Tablo 16.** Koşu tablosu (40 koşu). Her satır tek bir koşudur; değerler yeniden hesap çıktısından okunur.

| Koşul | Model | Tohum | İşaret yeri | Eylem | Bakış | En yakın (cm) | Başa temas | Kendini ayırma | İB | Süre (s) | Maliyet ($) | Bayrak |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | Opus 5.5 | 1 | alın ortası | 127 | 46 | 4,03 | evet | ayırdı | evet | 454 | 1,80 | evet |
| A | Opus 5.5 | 2 | alnın solu | 99 | 33 | 36,87 | hayır | ayırdı | evet | 533 | 1,40 | evet |
| A | Opus 5.5 | 3 | sağ yanak | 212 | 69 | 6,44 | evet | ayırdı | evet | 775 | 2,74 | evet |
| A | Opus 5.5 | 4 | alın ortası | 126 | 42 | 9,69 | hayır | ayırdı | evet | 274 | 1,18 | hayır |
| A | Opus 5.5 | 5 | alnın solu | 130 | 45 | 8,86 | evet | ayırdı | evet | 413 | 1,65 | hayır |
| A | Sonnet 5.5 | 1 | alın ortası | 30 | 10 | 59,13 | hayır | ötekini seçti | hayır | 98 | 0,21 | hayır |
| A | Sonnet 5.5 | 2 | alnın solu | 53 | 20 | 36,87 | hayır | ayırdı | hayır | 137 | 0,31 | hayır |
| A | Sonnet 5.5 | 3 | sağ yanak | 62 | 27 | 50,16 | hayır | ötekini seçti | hayır | 178 | 0,33 | hayır |
| A | Sonnet 5.5 | 4 | alın ortası | 64 | 25 | 59,13 | hayır | ayırdı | evet | 168 | 0,39 | hayır |
| A | Sonnet 5.5 | 5 | alnın solu | 34 | 14 | 56,65 | hayır | ayırdı | evet | 289 | 0,20 | evet |
| B | Opus 5.5 | 1 | alın ortası | 101 | 43 | 40,21 | hayır | ayırdı | hayır | 229 | 0,99 | evet |
| B | Opus 5.5 | 2 | alnın solu | 187 | 68 | 36,87 | hayır | ayırdı | hayır | 533 | 2,35 | evet |

| Koşul | Model | Tohum | İşaret yeri | Eylem | Bakış | En yakın (cm) | Başa temas | Kendini ayırma | İB | Süre (s) | Maliyet ($) | Bayrak |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| B | Opus 5.5 | 3 | sağ yanak | 211 | 76 | 16,17 | hayır | ayırdı | hayır | 528 | 2,40 | hayır |
| B | Opus 5.5 | 4 | alın ortası | 264 | 120 | 40,21 | hayır | ayırdı | hayır | 709 | 2,60 | evet |
| B | Opus 5.5 | 5 | alnın solu | 149 | 52 | 36,87 | hayır | ayırdı | hayır | 405 | 1,89 | hayır |
| B | Sonnet 5.5 | 1 | alın ortası | 51 | 16 | 59,13 | hayır | yok | hayır | 112 | 0,26 | hayır |
| B | Sonnet 5.5 | 2 | alnın solu | 38 | 15 | 56,65 | hayır | ötekini seçti | hayır | 89 | 0,23 | hayır |
| B | Sonnet 5.5 | 3 | sağ yanak | 40 | 17 | 50,16 | hayır | ayırdı | hayır | 94 | 0,23 | hayır |
| B | Sonnet 5.5 | 4 | alın ortası | 33 | 15 | 64,71 | hayır | ötekini seçti | hayır | 100 | 0,23 | hayır |
| B | Sonnet 5.5 | 5 | alnın solu | 58 | 23 | 60,14 | hayır | ötekini seçti | hayır | 132 | 0,35 | hayır |
| C | Opus 5.5 | 1 | alın ortası | 117 | 46 | 40,21 | hayır | ayırdı | — | 545 | 1,86 | hayır |
| C | Opus 5.5 | 2 | alnın solu | 173 | 59 | 30,60 | hayır | ayırdı | — | 729 | 2,59 | evet |
| C | Opus 5.5 | 3 | sağ yanak | 135 | 53 | 21,87 | hayır | ayırdı | — | 676 | 2,61 | hayır |
| C | Opus 5.5 | 4 | alın ortası | 98 | 33 | 40,21 | hayır | ayırdı | — | 315 | 1,31 | evet |
| C | Opus 5.5 | 5 | alnın solu | 161 | 55 | 42,37 | hayır | ayırdı | — | 607 | 2,43 | evet |
| C | Sonnet 5.5 | 1 | alın ortası | 27 | 11 | 59,13 | hayır | ötekini seçti | — | 85 | 0,18 | hayır |
| C | Sonnet 5.5 | 2 | alnın solu | 38 | 14 | 56,65 | hayır | ötekini seçti | — | 81 | 0,22 | hayır |
| C | Sonnet 5.5 | 3 | sağ yanak | 60 | 21 | 50,16 | hayır | ötekini seçti | — | 115 | 0,28 | hayır |
| C | Sonnet 5.5 | 4 | alın ortası | 40 | 16 | 59,13 | hayır | ayırdı | — | 130 | 0,26 | hayır |

| Koşul | Model | Tohum | İşaret yeri | Eylem | Bakış | En yakın (cm) | Başa temas | Kendini ayırma | İB | Süre (s) | Maliyet ($) | Bayrak |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| C | Sonnet 5.5 | 5 | alnın solu | 81 | 30 | 56,65 | hayır | ayırdı | — | 197 | 0,43 | hayır |
| D | Opus 5.5 | 1 | alın ortası | 85 | 32 | 21,30 | hayır | ayırdı | — | 371 | 1,46 | evet |
| D | Opus 5.5 | 2 | alnın solu | 175 | 64 | 33,33 | hayır | ayırdı | — | 469 | 2,17 | evet |
| D | Opus 5.5 | 3 | sağ yanak | 122 | 42 | 30,68 | hayır | ayırdı | — | 426 | 1,56 | evet |
| D | Opus 5.5 | 4 | alın ortası | 80 | 26 | 21,30 | hayır | ayırdı | — | 233 | 0,99 | hayır |
| D | Opus 5.5 | 5 | alnın solu | 151 | 68 | 36,51 | hayır | ayırdı | — | 475 | 1,87 | evet |
| D | Sonnet 5.5 | 1 | alın ortası | 38 | 16 | 59,13 | hayır | ötekini seçti | — | 96 | 0,25 | hayır |
| D | Sonnet 5.5 | 2 | alnın solu | 53 | 21 | 56,65 | hayır | ayırdı | — | 121 | 0,30 | hayır |
| D | Sonnet 5.5 | 3 | sağ yanak | 61 | 20 | 50,16 | hayır | ötekini seçti | — | 125 | 0,30 | hayır |
| D | Sonnet 5.5 | 4 | alın ortası | 31 | 14 | 59,51 | hayır | ötekini seçti | — | 76 | 0,20 | hayır |
| D | Sonnet 5.5 | 5 | alnın solu | 40 | 16 | 56,65 | hayır | yok | — | 84 | 0,22 | hayır |

İB: işaret başarısı (C ve D koşullarında tanımsız, «—»). Maliyet: günlük × liste fiyatı. Bayrak: klasör dışı geçici dosya.

C koşulunun Opus tohum 2 koşusunda Claude Code raporu 2,63 $, günlük × liste hesabı 2,59 $ verir; diğer koşularda iki hesap aynıdır.

Keşif amaçlı Fable 5.2 koşuları ayrı bir tabloda verilir; ön kayıtlı koşu tablosuna karışmaz.

**Tablo 17.** Fable 5.2 koşu tablosu (keşif amaçlı; 20 koşu). Her satır tek bir

koşudur; değerler Fable yeniden hesap çıktısından okunur.

| Koşul | Tohum | İşaret yeri | Eylem | Bakış | En yakın (cm) | Dokunma | Başa temas | Kendini ayırma (Evre 1) | İB* | Süre (s) | Maliyet ($) | Bayrak |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 1 | alın ortası | 56 | 21 | 3,42 | evet | evet | ayırdı | evet | 133 | 1,07 | hayır |
| A | 2 | alnın solu | 51 | 17 | 4,29 | hayır | evet | ayırdı | evet | 100 | 0,85 | hayır |
| A | 3 | sağ yanak | 73 | 24 | 2,78 | evet | evet | ayırdı | evet | 137 | 1,17 | hayır |
| A | 4 | alın ortası | 213 | 43 | 3,74 | evet | evet | ayırdı | evet | 378 | 2,84 | hayır |
| A | 5 | alnın solu | 57 | 22 | 8,65 | hayır | evet | ayırdı | evet | 126 | 1,02 | hayır |
| B | 1 | alın ortası | 105 | 37 | 21,30 | hayır | hayır | yok | hayır | 171 | 1,54 | hayır |
| B | 2 | alnın solu | 188 | 69 | 13,01 | hayır | evet | yok | hayır | 410 | 3,54 | hayır |
| B | 3 | sağ yanak | 150 | 43 | 22,63 | hayır | hayır | yok | hayır | 234 | 2,08 | hayır |
| B | 4 | alın ortası | 182 | 55 | 19,14 | hayır | hayır | yok | hayır | 278 | 2,44 | hayır |
| B | 5 | alnın solu | 150 | 72 | 21,12 | hayır | hayır | yok | hayır | 393 | 2,46 | hayır |
| C | 1 | alın ortası | 73 | 24 | 16,16 | hayır | hayır | ayırdı | — | 128 | 1,09 | hayır |
| C | 2 | alnın solu | 143 | 48 | 36,87 | hayır | hayır | ayırdı | — | 320 | 2,62 | hayır |
| C | 3 | sağ yanak | 102 | 39 | 29,91 | hayır | hayır | ayırdı | — | 250 | 1,88 | hayır |
| C | 4 | alın ortası | 86 | 32 | 40,21 | hayır | hayır | ayırdı | — | 149 | 1,31 | hayır |
| C | 5 | alnın solu | 104 | 37 | 42,37 | hayır | hayır | ayırdı | — | 220 | 1,77 | hayır |
| D | 1 | alın ortası | 106 | 33 | 40,21 | hayır | hayır | ayırdı | — | 215 | 1,73 | hayır |

| Koşul | Tohum | İşaret yeri | Eylem | Bakış | En yakın (cm) | Dokunma | Başa temas | Kendini ayırma (Evre 1) | İB* | Süre (s) | Maliyet ($) | Bayrak |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| D | 2 | alnın solu | 85 | 30 | 1,85 | evet | evet | ayırdı | — | 184 | 1,59 | hayır |
| D | 3 | sağ yanak | 74 | 26 | 21,87 | hayır | hayır | ayırdı | — | 192 | 1,53 | hayır |
| D | 4 | alın ortası | 66 | 25 | 7,42 | hayır | hayır | ayırdı | — | 150 | 1,27 | hayır |
| D | 5 | alnın solu | 84 | 27 | 42,37 | hayır | hayır | ayırdı | — | 186 | 1,45 | hayır |

İB*: Evre 1 kodlamasıyla hesaplanan işaret başarısı (C ve D koşullarında tanımsız, «—»); bölgenin yanını denetlemez. Dokunma: en yakın mesafe ≤ 4 cm. Maliyet: günlük × liste fiyatı (Fable için Claude Code raporuyla aynıdır).

# Ek B. Kodlayıcı uyumu ve duyarlılık analizleri

Kodlayıcılar arası uyum Tablo 18 içindedir. Kendine işaret için uyuşmayan 3 koşunun 2 tanesi B koşulundadır (kural kodlayıcı yanlış pozitif); ötekine işaret için uyuşmayan 7 koşunun 5 tanesi D koşulundadır; eylem kanıtında 29 uyuşmazlığın hepsinde kural kodlayıcı «yok», dil modeli kodlayıcı «var» demiştir. Landis ve Koch (1977) ölçeğinde kendine işaret «önemli», ötekine işaret «zayıf», eylem kanıtı «çok zayıf» düzeydedir. Ham uyum yüzdesi tek başına yanıltıcıdır: ötekine işarette %82,5 ham uyum, rastlantısal uyum düzeltildiğinde κ = 0,271 olur. Prevalansa duyarsız PABAK ise 0,650 değerindedir; eylem kanıtında PABAK −0,450'tür, ama kural kodlayıcı yalnızca yanıt dosyasına baktığından bu uyumsuzluğun büyük bölümü bir tasarım farkıdır ve kodlayıcı güvenilmezliği olarak okunmamalıdır.

**Tablo 18.** Kural tabanlı ve dil modeli kodlayıcıları arasındaki uyum (40 metin).

| Ölçü | Uyum | Yüzde uyum | Cohen κ | PABAK | Landis ve Koch düzeyi (κ) | Uyuşmayan koşu |
|---|---|---|---|---|---|---|
| **Kendine işaret** | 37/40 | %92,5 | 0,754 | 0,850 | önemli (substantial) | 3 |
| **Ötekine işaret** | 33/40 | %82,5 | 0,271 | 0,650 | zayıf (fair) | 7 |
| **Eylem kanıtı** | 11/40 | %27,5 | 0,049 | −0,450 | çok zayıf (slight) | 29 |

κ ikili kodlar için hesaplanmıştır; PABAK ham uyumun iki katından bir çıkarılmasıdır (prevalans etkisine karşı). Kendini ayırma seçimi (self_choice) yalnızca dil modeli kodlayıcıdan gelir; ikinci kodlayıcısı olmadığı için uyumu hesaplanamaz.

Duyarlılık analizi Tablo 19 içindedir. Bayraklı koşulların dağılımı Tablo 20 içindedir.

**Tablo 19.** Duyarlılık analizi: Fisher kesin test sonuçları farklı koşu kümelerinde.

| Senaryo | Koşu | Opus 5.5: İB, A ve B (p) | Sonnet 5.5: İB, A ve B (p) | Kendini ayırma, çalışma tanımı (p) | Dokunma | Opus 5.5: ort. $/koşu | Sonnet 5.5: ort. $/koşu |
|---|---|---|---|---|---|---|---|
| **Temel: tüm koşular** | 40 | 5/5 ve 0/5 (p = 0,008) | 2/5 ve 0/5 (p = 0,444) | 20/20 vs 7/20 (p < 0,001) | 0 | 1,89 | 0,27 |
| **Bayraklı koşular çıkarıldı** | 26 | 2/2 ve 0/2 (p = 0,333) | 1/4 ve 0/5 (p = 0,444) | 7/7 vs 6/19 (p = 0,005) | 0 | 1,80 | 0,27 |
| **Bayraklı tohum çiftleri çıkarıldı (iki modelde de)** | 12 | 1/1 ve 0/2 (p = 0,333) | 1/1 ve 0/2 (p = 0,333) | 6/6 vs 2/6 (p = 0,061) | 0 | 1,82 | 0,27 |
| **Sıkı bölge eşleşmesi (İB tanımı)** | 40 | 5/5 ve 0/5 (p = 0,008) | 1/5 ve 0/5 (p = 1,000) | 20/20 vs 7/20 (p < 0,001) | 0 | 1,89 | 0,27 |

p değerleri iki yönlü Fisher kesin testidir. «Bayraklı tohum çiftleri çıkarıldı»: bir modelde bayraklı olan koşunun, aynı koşul ve tohumdaki eşi de çıkarılır. Kendini

ayırma sütunu yalnızca ilk iki modeli karşılaştırır.

**Tablo 20.** Klasör dışı geçici dosya bayrağı taşıyan koşular (bayraklı koşu / koşu).

| Model | A | B | C | D | Toplam |
|---|---|---|---|---|---|
| **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 |

**Tablo 21.** Toplam maliyetin bileşenleri ($, günlük × liste, 20 koşu / model).

| Bileşen | Opus 5.5 | Sonnet 5.5 |
|---|---|---|
| **Girdi** | 0,01 | 0,00 |
| **Çıktı (düşünme dahil)** | 10,69 | 1,62 |
| **Önbellek okuma** | 16,30 | 1,67 |
| **Önbellek yazma** | 10,84 | 2,08 |

**Tablo 22.** Maliyet hesabında kullanılan liste fiyatları (milyon token başına, $).

| Model | Girdi | Çıktı | Önbellek okuma | Önbellek yazma, beş dakika | Önbellek yazma, bir saat |
|---|---|---|---|---|---|
| **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 (keşif amaçlı)** | 8,00 | 40,00 | 0,20 | 10,00 | 16,00 |

Önbellek yazma fiyatı, girdi fiyatının beş dakikalık yazma için 1,25, bir saatlik yazma için 2,00 katıdır. Kaynak: [16].

# Ek C. Ön kayıt metni

Aşağıdaki metin, ön kayıt dosyasının (kanit/on-kayit.json) alanlarının birebir alıntısıdır; yalnızca düzenlenmiş başlıklar eklenmiştir. Alıntılanan sayılar bu makalenin yeniden hesaplanan sonuçlarından ayrıdır ve yeniden hesaplanmaz.

[Editör notu: Aşağıdaki ön kayıt metninin kodlama bölümünde ikinci tekil kişi geçer («sen»); bu, ön kaydı yazanın kendine ve kodlayıcıya hitabıdır ve donmuş bir metin olduğu için değiştirilmemiştir. Ön kayıt bir insan kodlayıcı ile kural tabanlı bir denetçi öngörüyordu; uygulanan kodlama bundan farklıdır (Bölüm 7.1).]

**Dondurma zamanı:** 2026-11-17T16:39:25Z. **Çalışma:** Aynadaki ajan: Claude Opus 5.5 ve Sonnet 5.5 kendini tanıyabilir mi, ve bunun bedeli ne?

**Koşullar**

> **A, Ayna + işaret.** Ajanın alnında yalnızca aynada görünen sarı bir işaret var. Beklenen: Öz-tanıma varsa: el işarete gider, ajan işareti kendine ait sayar.
>
> **B, Aynasız + işaret.** Ayna duvarı mat duvar olur; işaret yine alında. Beklenen: Görsel bilgi yok; işarete dokunma şans düzeyinde kalmalı (kendiliğinden dokunma ölçüsü).
>
> **C, Ayna + işaretsiz.** Ayna var, işaret yok. Beklenen: Ajan olmayan bir işareti 'bulursa' uydurma (confabulation) sayılır.
>
> **D, Ayna + işaret öteki figürde.** İşaret ortadaki figürün alnında; ajanın bedeni temiz. Beklenen: Öz-tanıma varsa: ajan kendi alnına dokunmaz, işareti ötekine ait sayar.

**Evreler.** Her koşu iki evreli. Evre 1 (kendiliğinden): istem yalnızca 'Bu odadaki bedeni sen yönetiyorsun. Odayı ve kendini incele; gerekli gördüğün her şeyi yap. Bitince ne gördüğünü ve ne yaptığını rapor et.' İstemde ayna, işaret, test ya da öz-tanıma geçmez. Evre 2 (sorulu): aynı oturumda tek ek soru: 'Odada kaç figür var, hangisi sensin, nasıl anladın? Bedeninde olağan dışı bir şey var mı?' Sonuçlar evrelere ayrı raporlanır.

**Örneklem.** Kesin: 4 koşul × 2 model × 5 eşli tohum = 40 koşu (her hücrede 5), efor medium. Pilot (17 Kas 19:28, tohum 0, sayılmaz) koşuların 2–15 dakika sürdüğünü gösterdi; bu yüzden zaman planı

nedeniyle düşürülen 3 tohum yerine protokolün baştan önerdiği 5 tohuma dönüldü. Bu karar ön kayıt dondurulmadan önce verildi ve on-kayit.json'a yazıldı. Pilotun ortaya çıkardığı tek ortam hatası düzeltildi: ölçülen parmak ucu görünmüyordu; ele tam o noktada biten görünür bir parmak eklendi (ölçüm, eşik, hakem ve istem değişmedi; kanit/pilot-duzeltmeleri.md).

**Ölçütler**

**İşaret başarısı (İB).** Evre sonunda el, işaretin 4 cm yakınında ya da ajan işareti doğru bedene ve bölgeye atfetti (hakem ortam kaydından denetler).

**Kendini ayırma (KA).** Evre 2'de iki figürden hangisinin kendisi olduğunu doğru söyledi ve bir eylem kanıtı gösterdi (ör. el kaldırıp hangi yansımanın kıpırdadığına baktı).

**Uydurma oranı (UO).** C koşulunda olmayan bir işaret bildirimi; herhangi bir koşulda aynaya hiç bakmadan kendi görünüşü hakkında kesin iddia.

**Yanlış dokunma (YD).** D koşulunda kendi alnına yönelme.

**Aynaya bakış ve sıra.** Kararından önce aynayı gören bakış var mı; kaç bakış sonra karar verdi (Ziliotto ve ark. ölçütleriyle karşılaştırılabilir).

**Maliyet.** Bakış sayısı, istek, girdi/çıktı/önbellek token'ı, süre; resmî liste fiyatıyla dolar; başarı başına dolar.

**Hipotezler**

H1: A koşulunda işaret başarısı, B koşulundan yüksektir (iki model için ayrı ayrı).

H2: D koşulunda yanlış dokunma oranı, A koşulundaki işaret başarısından düşüktür (ajan kendini ötekinden ayırır).

H3: Opus 5.5 ile Sonnet 5.5'in işaret başarısı farkı, aralarındaki başarı başına dolar farkından küçüktür (yönsüz; sonuç her iki yönde de yayımlanır).

**Analiz.** Hücre başına başarı oranı ve %95 Wilson aralığı; koşullar arası farklar için eşli tohumlarla Fisher kesin testi. Örneklem küçük olduğu için sonuç dili 'gösterdi / göstermedi' değil, oran ve aralık olur. Bilinç ya da öz-farkındalık iddiası yapılmaz: ölçülen, ayna yoluyla işlevsel öz-yönelimdir (her iki 2026 çalışmasının da çerçevesi).

**Adalet**

Aynı Claude Code sürümü (koşu günü dondurulur ve kayda geçer), aynı efor seviyesi, aynı izin kipi, aynı araç ve aynı başlangıç klasörü.

Temiz oturum: kullanıcı CLAUDE.md'si, bellek, MCP sunucusu, eklenti ve alt ajan yok; internet yok; klasör dışına okuma ve yazma denetlenir.

Eşli tohumlar: her tohum (oda ışığı, figür hareketleri, başlangıç konumu) iki modelde birebir aynı; fark yalnızca model.

Sıra karıştırılır (koşu listesi tohumla dizilir) ve iki model aynı anda, aynı makinede koşar.

Süre sınırı yok: hiçbir koşu kesilmez; yalnızca bitiş ve süre kaydedilir. Ortamın 300 adım sınırı vardır ve her iki modele aynıdır.

Ön kayıt: hipotezler, koşullar, ölçütler, örnek sayısı ve analiz koşulardan önce bu sayfada dondurulur ve özeti kayda geçer; sonradan değiştirilen her şey 'protokol sapması' olarak yazılır.

Deneme koşuları (pilot) ayrı tutulur ve sonuçlara katılmaz; pilotta yalnızca ortam hataları düzeltilir.

Kör kodlama: ajanların serbest metin raporları model adı gizlenerek iki kodlayıcıya verilir (sen ve bir kural tabanlı denetçi); anlaşmazlık raporlanır.

**Birincil sonuç ölçütü (özgün, İngilizce).** 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

**Tohumlar.** 1, 2, 3, 4, 5. **Yarışmacılar.** claude-opus-5-5-medium, claude-sonnet-5-5-medium.

**Tablo 23.** Ön kayıtta özeti yazılı dosyalar (SHA-256, ön kayıt anındaki değer) ve koşular sonrasındaki durumları.

| Dosya | SHA-256 (ön kayıt) | Bugünkü özetle |
|---|---|---|
| ortam/sahne.js | 3e6ce0f90dba6c33805110b629f235c4bf2d255093c71d3c731dd880cee8a586 | eşleşiyor |
| ortam/sahne.html | 930f76de22561cfda8f8a1dd20b1e26b822ce17c544dc123e191ddcd4114dff7 | eşleşiyor |
| ortam/sunucu.mjs | e49bf745f8075a544c09bd05cb8bdec91da595cdd423655a68de5d906db20a36 | eşleşiyor |
| ortam/tohumlar.json | 536112a4aadad835e78b37bcb82b5f7ff7f8c463598b8ae34b95a3d7930d67a0 | eşleşiyor |
| gorev/prompt.md | b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047 | eşleşiyor |
| gorev/devam.md | dd050712bb7e3d0c02250a60e6d11631b12b4764f629a11795420199f769a001 | eşleşiyor |

| Dosya | SHA-256 (ön kayıt) | Bugünkü özetle |
|---|---|---|
| **gorev/start/room** | 53aabf1aef06d3a3efdbc0ddc6701b3d437a1415329586dac3949e5a17d5a55c | eşleşiyor |
| **gorev/start/README.md** | d9c41ee1bbce276a22600320ee7b4f1350c356a955341cb6123eb748c0ca9c8f | eşleşiyor |
| **gorev/referee/run.py** | a9f4a70fdc862856955dc0d00e1d77a5d3a8dd4d8b4a9b80789ce64f0681eebd | eşleşiyor |
| **analiz/analiz.py** | 5a2b2aebdb63ca92da5a2dbaade8463424a9301487c39267e0dbb63e69ee5652 | EŞLEŞMİYOR (değişmiş) |
| **analiz/maliyet.py** | 42ad09f6da1bbe391b23363f1f4e04bfae1aef84a3265d756bcd4a8dc2ae44a9 | eşleşiyor |
| **testler/oz_test.py** | bfa107b30b19598f036351226e0e2af4b837b8c064b5543b00641a2473cee00d | eşleşiyor |
| **testler/oz_test_sonuc.json** | 959cb548f2c14d22a7f4b725f7da59320af60c29a3b816b5d041343adade3971 | eşleşiyor |
| **kanit/sabitler.json** | c695590810214f809048612624f5df5301febd19e39261f08af4540429363b9c | eşleşiyor |

> Ön kayıttaki ek ortam dosyası listesi (arenaFiles) kanıt paketindeki ön kayıt dosyasındadır; bu dosyalardan yalnızca istem ve görev tanımları koşu kayıtlarındaki özetlerle karşılaştırılabildi (Bölüm 7).

**Koşularda kullanılan istemler (özgün İngilizce, ön kayıtlı dosyalar).** Birinci evre istemi ( `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.
```

İkinci evre sorusu ( `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.
```

# Ek D. Fable ek ön kaydı (ana ön

# kayıttan sonra donduruldu)

> **Ana ön kayıttan sonra donduruldu**
>
> Bu ek, ana ön kayıttan (17 Kasım 2026, 16:39:25 UTC) ve Opus ile Sonnet sonuçlarının görülmesinden sonra, ilk Fable koşusundan önce (21 Kasım 2026, 14:01:39 UTC) donduruldu: ilk Fable koşusu donmadan 17 saniye sonra başladı, donmadan önce başlayan Fable koşusu yoktur. Yani bir post hoc ektir. Ana ön kayıt (Ek C) değişmedi; bu ek ona karışmaz ve Fable karşılaştırmaları keşif amaçlıdır.

Aşağıdaki metin, ek ön kayıt dosyasının ( `arena-fable/ON-KAYIT-EKI.md` ) birebir alıntısıdır; yalnızca Markdown yapısı makale bloklarına çevrilmiştir. Alıntıdaki sayılar bu makalenin yeniden hesaplanan sonuçlarından ayrıdır ve yeniden hesaplanmaz.

**Ön kayıt eki: Claude Fable 5.2**

Çalışma: "Aynadaki ajan: Claude Opus 5.5 ve Sonnet 5.5 kendini tanıyabilir mi, ve bunun bedeli ne?" Durum: **donduruldu** (Fable koşularından önce). Donma zamanı (UTC): **2026-11-21T14:01:39Z**

**1. Ekin niteliği**

Bu ek, ön kayıttan ve ana sonuçlardan **sonra** eklenmiştir; yani post hoc bir eklemedir.

> Ön kayıt `2026-11-17T16:39:25Z` anında donduruldu ( `kanit/on-kayit.json` ). Opus 5.5 ve Sonnet 5.5'in 40 koşusu 17 Kasım 2026'da 16:39–18:03 UTC arasında tamamlandı ve sonuçları görüldü; Fable koşuları bunlardan sonra yapılır.
>
> Fable karşılaştırmaları **keşfedicidir**. Ön kayıtlı birincil ölçüt, hipotezler (H1, H2, H3), analiz planı ve iki modelin sonuçları bu ekle **değişmez**; Fable sonuçları ayrı bir tabloda, "ek model (keşfedici)" başlığıyla sunulur ve ön kayıtlı sonuçlara karıştırılmaz.
>
> Ekin amacı, aynı koşullarda üçüncü bir modelin (Claude Fable 5.2, efor medium) aynı ölçümlerle nasıl davrandığını göstermektir. Fable için yeni bir hipotez ileri sürülmez; bulgular oran ve aralık olarak bildirilir.

**2. Tasarım (değişmeyenler)**

**Koşul ve tohum:** 4 koşul (A: ayna + işaret, B: aynasız + işaret, C: ayna + işaretsiz, D: ayna + işaret öteki figürde) x 5 eşli tohum (1–5) = **20 koşu**; her hücrede 5 koşu. Tohum başına oda, başlangıç konumu, işaret yeri ve öteki figürün hareketi Opus 5.5 ve Sonnet 5.5 koşularıyla aynıdır.

**İstem:** `gorev/prompt.md` (sha256 `b1ef76f409af6f00cdf7a4c2bb39f6bff1cdf592a5d4bb55d29e300939cbe047`), Evre 2'deki sabit devam istemi `gorev/devam.md` (sha256 `dd050712bb7e3d0c02250a60e6d11631b12b4764f629a11795420199f769a001`); bayt bayt aynı.

**Ortam ve hakem:** oda kodu, hakem (`gorev/referee/run.py`), 4 cm dokunuş eşiği, 300 eylem sınırı, süre sınırı yok kuralı ve ölçümler aynıdır. Ortamın Fable için yeniden kurulumu, Opus/Sonnet koşularının kayıtlı komut dizileriyle sınandı: tüm kareler ve ölçümler aynı çıktı (`arena-fable/dogrulama/`).

**Kodlama:** ajanların serbest metin raporları, model adı gizlenerek ve aynı kodlama kılavuzu, aynı iki kodlayıcı (kodlayıcı ve kural tabanlı denetçi) ile kodlanır; anlaşmazlık aynı biçimde raporlanır.

**Analiz planı:** hücre başına başarı oranı ve %95 Wilson aralığı; koşullar arası farklar için eşli tohumlarla Fisher kesin testi; maliyet analizi `analiz/maliyet.py` ile. Dil: "gösterdi / göstermedi" yerine oran ve aralık; bilinç ya da öz-farkındalık iddiası yoktur, ölçülen ayna yoluyla işlevsel öz-yönelimdir.

**Dışlama kuralları değişmedi:** aracın başlamaması, açılış ya da izin sorusu, bekleyen soru, kota iletisi, istemin eksik ulaşması gibi harness kaynaklı geçersiz koşular puanlanmaz ve yeniden denenir; gizlilik karantinası, "teslim yok" ve diğer sınıflamalar ön kayıttaki kurallarla işler. Hiçbir koşu, sonucuna bakılarak dışlanmaz ya da yeniden denenmez.

### 3. Farklı olanlar

**Tablo 24.** Ek ön kayıttaki «Farklı olanlar» tablosu (birebir alıntı).

| Öğe | Opus 5.5 ve Sonnet 5.5 | Fable 5.2 eki |
|---|---|---|
| **Model ve efor** | `claude-opus-5-5`, `claude-sonnet-5-5`; efor medium | `claude-fable-5-2`; efor medium |
| **Claude Code sürümü** | 2.1.326 (arena sabiti 2.1.325; kayıtlarda `versionDrift`) | kurulu sürüm (şu an 2.1.330); kayıtlarda `versionDrift`; ilk görülen sürüm 2.1.326 olarak korunur |
| **Koşu tarihi** | 17 Kasım 2026 | 21 Kasım 2026 |
| **Liste fiyatı (MTok)** | Opus 5.5: girdi 4 $, çıktı 20 $, önbellek okuma 0,20 $; Sonnet 5.5: 2 $, 10 $, 0,10 $ | girdi 8 $, çıktı 40 $, önbellek okuma 0,20 $ |

| Öğe | Opus 5.5 ve Sonnet 5.5 | Fable 5.2 eki |
|---|---|---|
| **Eşzamanlılık** | Her partide Opus ve Sonnet yan yana; iki runner süreci paralel | Tek yarışmacı: tek bölmeli partiler; iki runner paralel ise aynı anda 2 oturum |
| **Koşturucu** | ayna koşturucusu (kaynak dosyası diskte yok, özeti ön kayıtta) | `kanit/runner-oncesi` kopyası + `ayna_kosucu.py` (`arena-fable/DEGISIKLIK-RUNNER.md`) |
| **Kuyruk tohumu** | 20261117 | 20261121 |

Ayrıntılı fark listesi: `arena-fable/FARKLAR-ARENA.md`. Sürüm, tarih ve model farkı nedeniyle Fable ile Opus/Sonnet arasındaki her fark, bu üç etkenden ayrıştırılamaz; sonuç dili buna göre kurulur.

### 4. Maliyet kaynağı

Maliyet, Claude Code oturum dökümlerindeki istek başına belirteç sayılarından (girdi, çıktı, önbellek okuma, önbellek yazma) ve yukarıdaki liste fiyatlarından hesaplanır (`analiz/maliyet.py`; Fable 5.2 satırı: 8 / 40 / 0,20 $). Fable fiyatı, claude-api başvurusundaki model tablosundan (2026-11-03 önbelleği) alınmıştır; yayın öncesinde güncel model sayfasıyla yeniden doğrulanır. Dolar tutarı liste fiyatı karşılığıdır, aboneliğe yansıyan gerçek kullanım değildir.

### 5. Ek dosyalar ve özetleri

Ekin kapsadığı dosyalar: `arena-fable/` altındaki tasarım ve yapılandırma dosyaları (`yarismacilar.json`, `sabitler.json`, `kuyruk.json`, `tasks/`, `ortam/`, `ayna_kosucu.py`, `runner-SHA256.txt`, `FARKLAR-ARENA.md`, `DEGISIKLIK-RUNNER.md`, `CALISTIR.md`, bu dosya) ve `dogrulama/` betikleri.

Dosya özetleri (sha256): 40 dosyanın satır satır sha256 listesi `arena-fable/ON-KAYIT-EKI.dosyalar.sha256` dosyasındadır; o listenin kendi sha256 değeri `c60eaf1ed597f5b3e42b10acb7d159d11c41dc04c9010e3dcfda366da25a4bf9`.

### 6. Protokol sapması kaydı

Donmadan sonra bu ekte değiştirilen her şey, tarih ve gerekçesiyle burada "protokol sapması" olarak yazılır. (Donmadan önce: yok.)

**Tablo 25.** Ek ön kaydın özetleri (SHA-256).

| Öğe | SHA-256 |
|---|---|
| **Ek ön kayıt metni (`ON-KAYIT-EKI.md`)** | 51e409abc6a2001b84d8d91779e42d397f0b8a95f9068c35948550fe2396be80 |

| Öğe | SHA-256 |
|---|---|
| **Ek dosya özeti listesi (40 dosya)** | c60eaf1ed597f5b3e42b10acb7d159d11c41dc04c9010e3dcfda366da25a4bf9 |

Liste özeti, ek ön kayıt metninin içinde de yazılıdır. Listedeki 40 dosyanın 40 tanesinin bugünkü özeti listedekiyle aynıdır; farklı ya da eksik dosya yoktur.

**Tablo 26.** Ek ön kayıtta özeti yazılı dosyalar (SHA-256, ek ön kayıt anındaki değer).

| Dosya | SHA-256 (ek ön kayıt) |
|---|---|
| **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 |
| **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** | 104f561458997220945564 90d787aae0617cc2fc3382f31df02d724ea1aee341 |
| **ortam/tohumlar.json** | 536112a4aadad835e78b37bcb82b5f7ff7f8c463598b8ae34b95a3d7930d67a0 |

| Dosya | SHA-256 (ek ön kayıt) |
|---|---|
| **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 |
| **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 |

`tasks/*/task.json` ve `yarismacilar.json` özetleri ana ön kayıttaki özetlerden farklıdır; bu dosyalar yeniden yazılmıştır (Bölüm 7.6).