

Codex: zuiniger werken met behoud van kwaliteit

Datum: 14 september 2026. Meting rond 16:28-16:32 Nederlandse tijd.

Mijn advies: test Terra / Low voor dagelijks werk. Gebruik Astra gericht voor moeilijke opdrachten. Koop nu geen extra capaciteit op basis van deze momentopname. Een algemene modelwijziging vereist Jamals keuze.

Wat is daadwerkelijk gemeten?

ONDERDEEL	RESULTAAT
Algemeen accountvenster	1% gebruikt, circa 99% resterend; venster 7 dagen
Reset algemeen venster	21 september 2026 om 16:25:07, Europe/Amsterdam
Spark	0% gebruikt in zowel vijf-uursvenster als weekvenster; afgeronde meter
Extra credits / resetcredits	Saldo 0 / geen resetcredits beschikbaar
Accountlabel	Interne code <code>prolite</code> ; geen factuur of abonnementstarief gecontroleerd
Lokale aanmelding	<code>auth_mode = chatgpt</code>
Globale modelconfiguratie	<code>gpt-6-astra , reasoning low , service tier default</code>
Deze taak	Sessiegegevens bevestigen Astra / Low
Globale AGENTS.md	22.310 bytes; bytes zijn geen tokens
MCP-configuratie	22 serverdeclaraties; dit bewijst niet dat alle servers actief of geladen zijn
Eerste gemeten modelaanroep	44.772 invoertokens, waarvan 14.976 gecachet; 206 uitvoertokens

ONDERDEEL

RESULTAAT

Gemeld contextvenster in deze taak	258.400 tokens
------------------------------------	----------------

Bewijs: accounttool `get_usage_limits` , geselecteerde velden uit lokale configuratie en aanmeldmetadata, eerste `token_count` in de sessie van deze taak. Geen account-ID of sleutels in dit verslag opgenomen.

Deze 44.772 invoertokens zijn circa 17,3% van het gemelde contextvenster. Circa 33,4% van de invoer was gecachet. Het gaat om de eerste geregistreeerde aanroep, niet om alle tokens van vandaag of het hele account. De startcontext bevat onder meer instructies, gereedschappen, vaardigheden en voorgeschiedenis. Deze meting verdeelt het verbruik niet exact over die onderdelen.

Drie verschillende meters

Abonnementlimiet: je beschikbare werkruimte binnen het abonnement. De accountmeter is leidend; 1% is een afgeronde momentopname, geen bedrag.

Tokens: informatie die het model leest of schrijft. Dezelfde geschiedenis kan bij meerdere aanroepen meetellen. Opgetelde invoertokens zijn daarom niet hetzelfde als de omvang van één contextvenster.

API-kosten: gebruik met een eigen API-sleutel wordt apart tegen API-tarieven afgerekend. Deze lokale aanmelding gebruikt ChatGPT. Dat sluit afzonderlijke kosten van externe tools of andere programma's niet uit. API-facturatie en externe dienstabonnementen zijn hier niet gecontroleerd.

Officiële uitleg over limieten, tokens en credits

Modelkeuze

WERK	VOORSTEL	CONTROLE VOOR OPLEVERING
Duidelijke extractie, omzetting of korte samenvatting	Luna / Low	Kloppen alle gegevens met de bron?
Dagelijks werk en afgebakende codewijzigingen	Terra / Low, zo nodig Medium	Vereisten controleren en relevante test uitvoeren
Complex onderzoek of lastige code	Sol / Medium, zo nodig High	Bronnen, aannames en foutgevallen nalopen
Moeilijke taken met veel stappen en belangrijke afwegingen	Astra / Low of Medium, hoger indien nodig	Volledige keten en resultaat controleren

Dit volgt de officiële taakverdeling. Het is geen lokale kwaliteitsbenchmark. Astra blijft beschikbaar als een goedkoper model onvoldoende presteert. Gebruik Max of Ultra alleen als extra denkwerk of delegatie werkelijk nodig is.

Officiële modelkeuze en redeneerniveaus

De actuele standaardtarieven in credits per miljoen tokens:

MODEL	INVOER	GEACHTTE INVOER	UITVOER
Astra	250	25	1.250
Sol	100	10	500
Terra	50	5	300
Luna	5	0,5	30

Bij identieke tokenaantallen ligt Terra dus 80% lager voor invoer en 76% lager voor uitvoer dan Astra. Dit is geen bewezen besparing per afgeronde taak: herstelwerk kan het verschil verkleinen. Low maakt Astra niet automatisch goedkoper dan Terra.

Officiële tarieventabel

Laat Fast standaard uit: Astra Fast verbruikt 2,5 maal het standaardcredittarief. Dit is iets anders dan een lager redeneerniveau. `service_tier = default` staat lokaal al ingesteld; bestaande taken kunnen afwijken.

Context en werkwijze

1. Geef per opdracht doel, bronnen, grenzen en het vereiste bewijs. Houd optionele verbeteringen apart.
2. Lees gerichte passages. Vermijd complete logs, transcripties en documentatiesites in de chat.
3. Bewaar per project een compact STAND.md met besluiten, bewijs en volgende stap. Start een nieuwe taak bij een nieuw onderwerp of een grote faseovergang, met dit standbestand als overdracht.
4. Maak niet voor iedere kleine vervolgvraag een nieuwe taak: je moet dan opnieuw context opbouwen. Caching werkt aantoonbaar al in deze sessie.
5. Stel geen willekeurige kleinere contextlimiet in. Dat kan juist extra samenvattingen en herlezingen veroorzaken. Het hier gemelde venster is geen bewijs voor het maximale venster van elk model.
6. Laat Conductor een voorstel maken om algemene instructies te verkorten en domeinregels alleen in passende projecten te laden. De aangetroffen oude heartbeat en dubbele instructies zijn concrete kandidaten voor controle. Essentiële veiligheidsregels moeten behouden blijven.
7. Bekijk welke MCP-servers per project nodig zijn. Declaraties tellen is geen effectmeting; meet de eerste invoertokens vóór en na een goedgekeurde wijziging.

OpenAI adviseert kleinere prompts, gerichte bronnen, compacte AGENTS.md-bestanden en minder onnodige MCP-context.

Officiële bespaartips

Veilige verbetering en keuze

Voor deze opdracht is een compact overdrachtsbestand gemaakt. Gedeelde instellingen, modellen van andere taken, hooks en LaunchAgents zijn niet aangepast. Er zijn geen betaalde diensten geactiveerd.

Voorgestelde keuze: Terra / Low als standaard voor nieuwe dagelijkse taken, met Astra beschikbaar voor moeilijke taken. Eerst een beperkte vergelijking op drie representatieve taken: dezelfde bronnen, hetzelfde eindresultaat en dezelfde checks. Meet kwaliteit, herstelwerk, tijd en tokens. Pas daarna de standaard wijzigen. Er is nu geen bewijs voor een exact percentage besparing op Jamals totale werk.

Er is daarnaast een API-sleutel in plaintext aangetroffen in de lokale configuratie. De waarde staat niet in dit verslag. Laat Conductor die opslag en eventuele rotatie beoordelen; de sleutel is tijdens deze inspectie onbedoeld in tooluitvoer verschenen. Er is niets aan gewijzigd.

Open: toestemming voor gedeelde wijzigingen, een echte modelvergelijking, API-facturatiecontrole en publicatiecontrole indien Conductor het verslag publiceert.

Nameting aan het einde van het onderzoek: algemeen accountvenster 2% gebruikt, circa 98% resterend. Dit is accountbreed en afgerond; het verschil met 1% is niet uitsluitend aan dit onderzoek toe te schrijven.