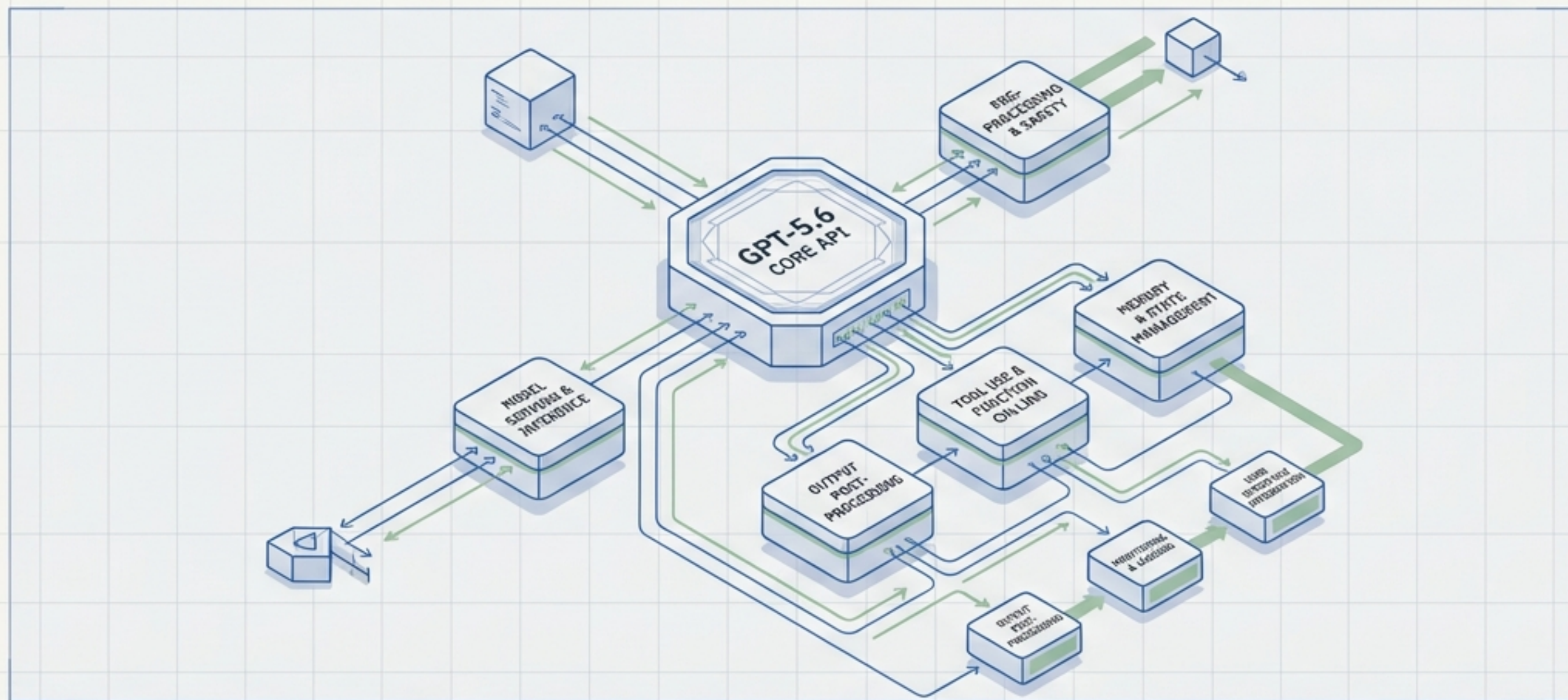


راهنمای جامع معماری و مهاجرت به GPT-5.6

استاندارد جدید کارایی، استدلال و ابزارسازی برای محیط پروداکشن

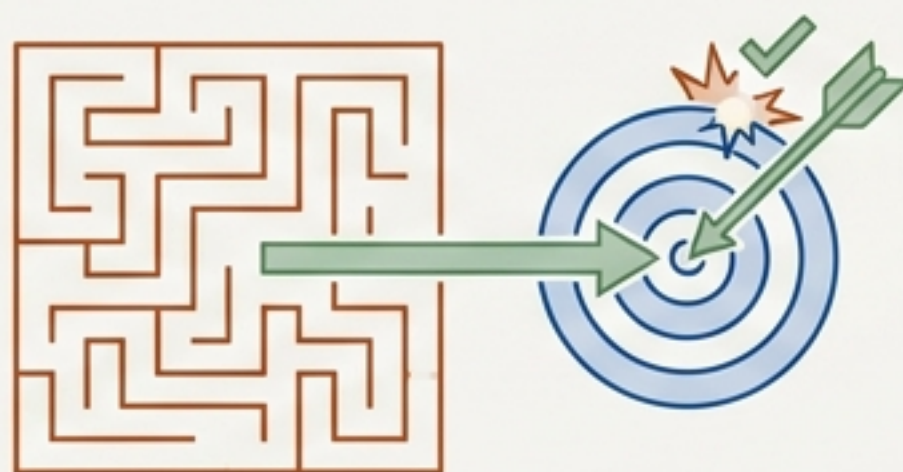


بر اساس مستندات رسمی OpenAI API

تغییر پارادایم در GPT-5.6



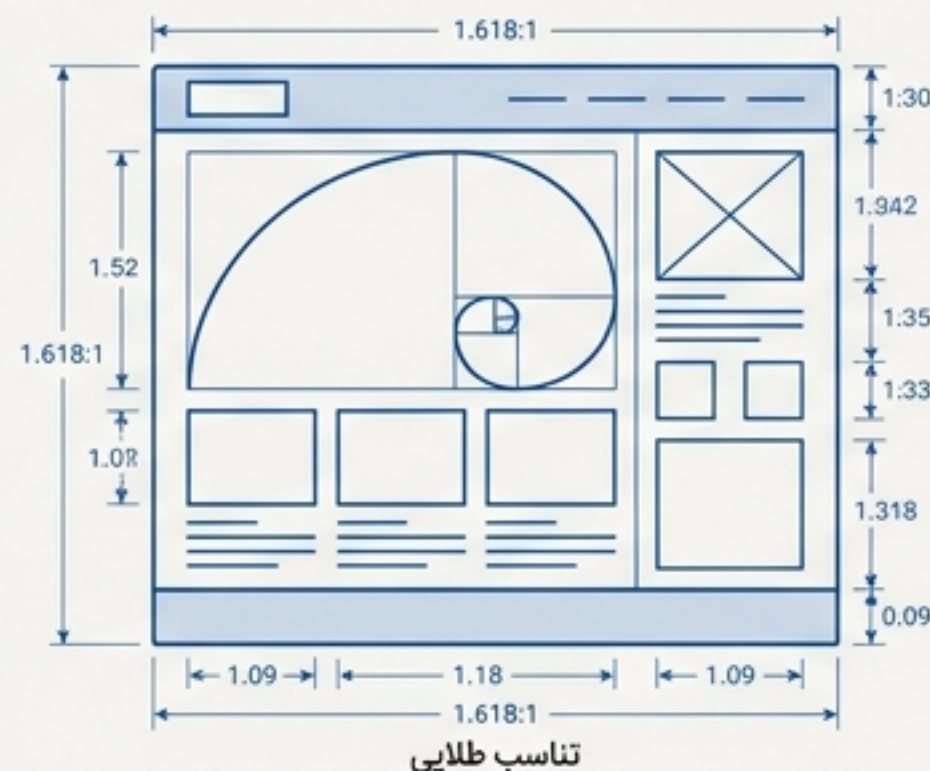
درک قصد کاربر (Intent)



کاهش نیاز به پرامپت‌های به شدت
دستوری؛ استنتاج بهتر هدف نهایی کاربر.



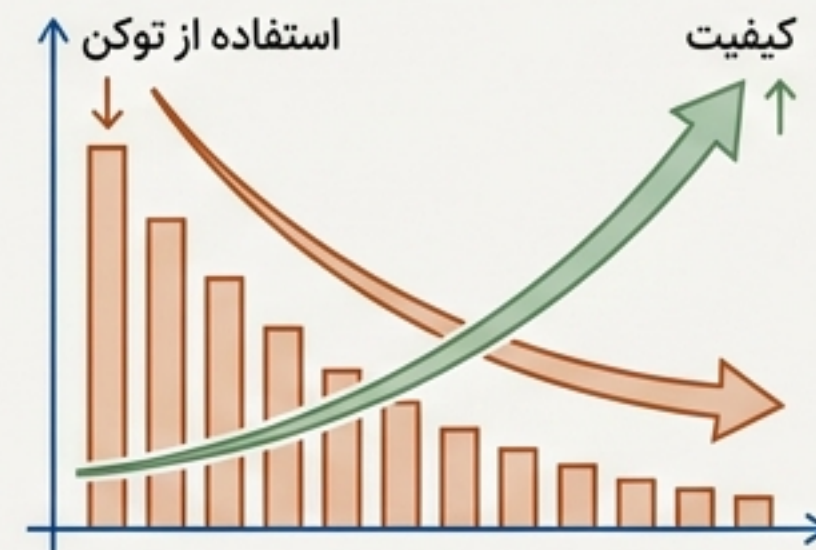
هوش طراحی (UI/UX)



درک عمیق‌تر از سلسله‌مراتب بصری و
قضاوت طراحی برای فرانت‌اند.



کارایی توکن



دستیابی به عملکرد مرزی با تولید
توکن‌های خروجی بسیار کمتر.

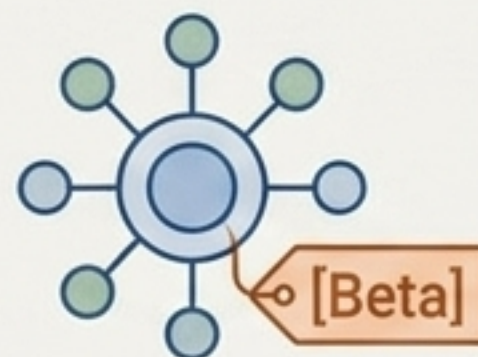
ماتریس خانواده مدل‌های GPT-5.6

gpt-5.6-luna (Volume)	gpt-5.6-terra (Balanced)	gpt-5.6-sol (Flagship)
• کاربرد: بارهای کاری با حجم بالا (High-volume). • بسیار کارآمد و حساس به سرعت و زمان پاسخ‌دهی.	• کاربرد: نقطه عطف بین هوش بالا و مدیریت هزینه‌ها. • عملکرد قدرتمند با قیمت پایین‌تر برای اکثر بارهای کاری.	• کاربرد: پیچیده‌ترین وظایف، استدلال‌های عمیق و بالاترین کیفیت ممکن. • مسيردهی: نام مستعار gpt-5.6 مستقیماً به این مدل اشاره می‌کند.

جعبه ابزار جدید توسعه دهندگان (API Innovations)



حالت حرفه‌ای
(Pro Mode)



ارکستراسیون چند-عاملی
(Multi-agent)



فراخوانی برنامه نویسی شده
ابزارها (PTC)



استدلال ماندگار
(Persisted Reasoning)

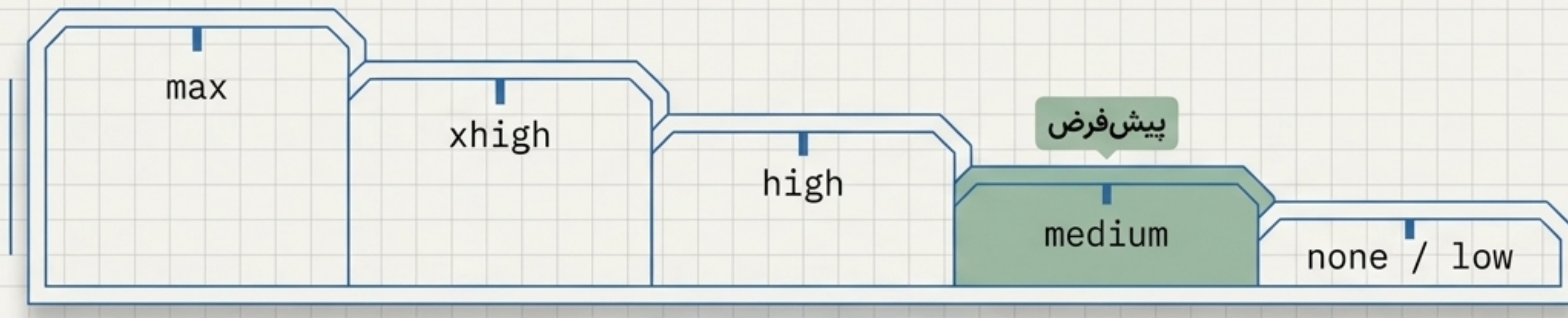


گش کردن صریح
(Explicit Caching)



تلاش استدلالی متغیر
(Max Reasoning)

کالیبراسیون طیف استدلال (reasoning.effort)



سطح مکس: فقط برای سخت‌ترین
پردازش‌های کیفیت-محور.

سطح متوسط و بالا: نقطه شروع
متعادل و ارتقای مستقیم کیفیت با
استدلال بیشتر.

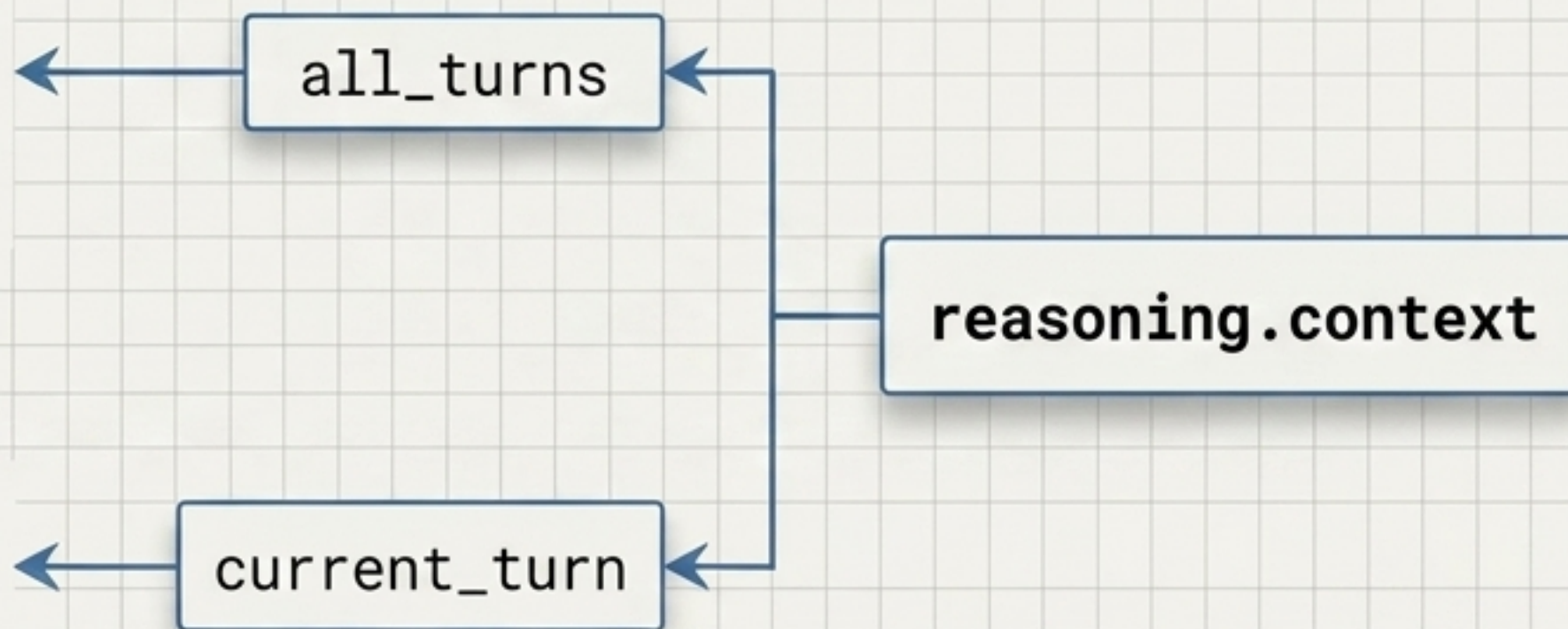
سطح پایین: مخصوص
پردازش‌های حساس به تاخیر
(Latency-sensitive).

⚠ **نکته مهاجرت:** هنگام ارتقا به GPT-5.6، تست خود را با
سطح فعلی و سپس یک سطح پایین‌تر آغاز کنید. مدل جدید با
توکن کمتر، کیفیت مشابهی تولید می‌کند.

مدیریت کانتکست و استدلال ماندگار در مکالمات

کی استفاده کنیم؟ زمانی که اهداف مکالمه در طول مسیر ثابت هستند.

از `previous_response_` برای اتصال مراحل استفاده کنید.

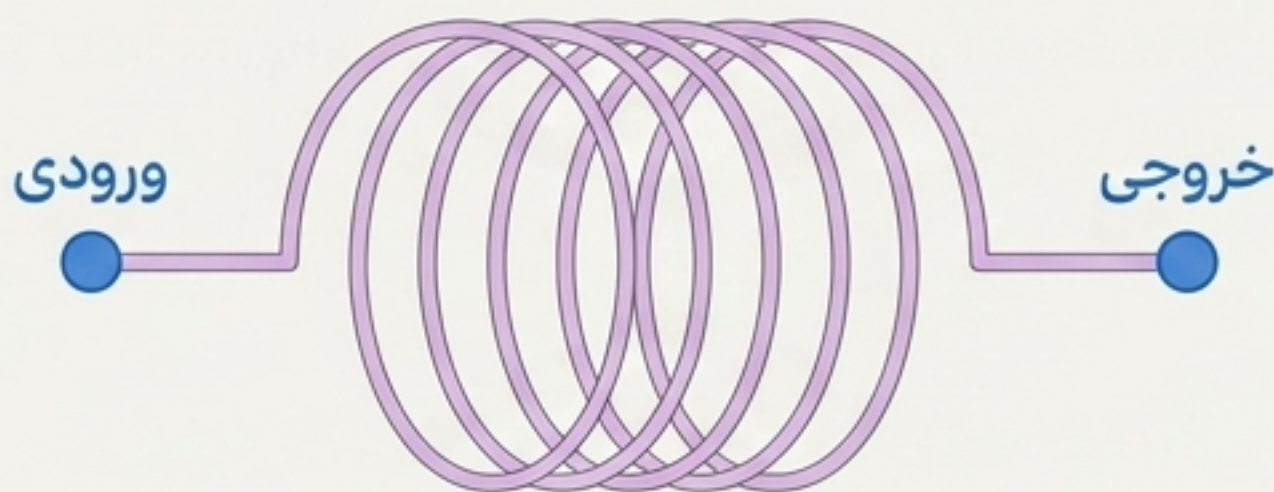


هشدار امنیتی در حالت (ZDR) Zero Data Retention

در حالت `store: false`، مدیریت تاریخچه دستی است. شما باید آیتم‌های استدلال رمزنگاری شده بازگشتی را برای تداوم کانتکست بازپخش (Replay) کنید.

کالبدشکافی حالت حرفه‌ای (reasoning.mode: pro)

حالت حرفه‌ای



تخصیص پردازش بیشتر به مدل قبل از پاسخ نهایی.
مناسب برای بهینه‌سازی‌های پیچیده، کدنویسی باارزش بالا و تحلیل‌های عمیق.

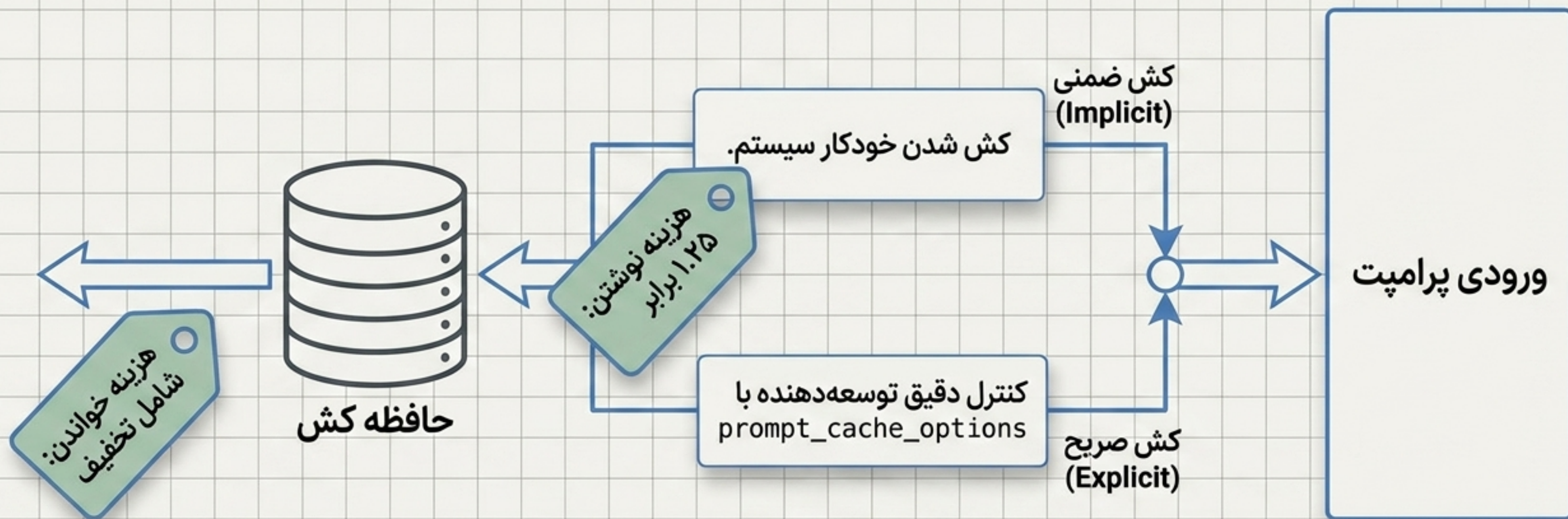
حالت استاندارد



مناسب برای کارهای روتین، بارهای کاری با حجم بالا و فرآیندهای حساس به تاخیر.

💡 نکته پرامپت‌نویسی: نیازی به گفتن عبارت "سخت‌تر فکر کن" در پرامپت نیست. فقط هدف، کانتکست و معیارهای موفقیت را صریح بیان کنید.

معماری و اقتصاد گش (Caching) در GPT-5.6



💡 اقدام ضروری پروداکشن: مانیتور کردن مداوم پارامتر `cached_tokens` برای درک هزینه خالص سیستم.

انتخاب معماری فراخوانی ابزارها

Direct Tool Calling (مستقیم)

استفاده شود اگر: خروجی‌های میانی کوچک هستند، تصمیم‌گیری بعدی به نتیجه ابزار بستگی دارد، یا نیاز به تایید کاربر هست.

Programmatic Tool Calling (PTC)

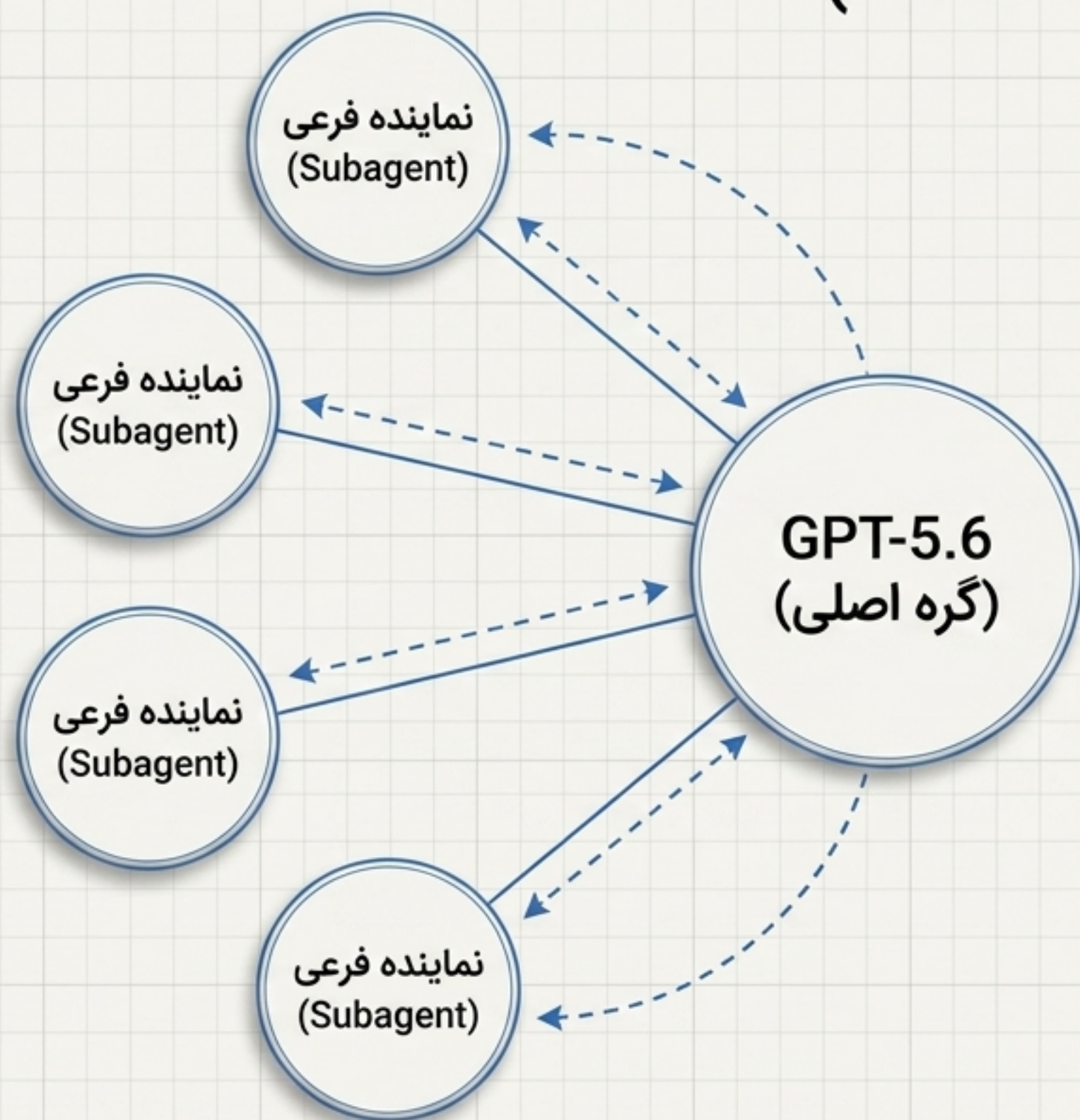
استفاده شود اگر: نیاز به فیلتر کردن، تجمیع، رتبه‌بندی یا اعتبارسنجی خروجی‌های میانی بسیار بزرگ است تا به یک خروجی ساختاریافته کوچک تبدیل شوند.

ساختار داده‌های ابزار
چگونه است؟

جزئیات فنی PTC: مدل، کدهای جاوااسکریپت را در یک محیط ایزوله اجرا می‌کند. این روش کاملاً با کدهای ZDR سازگار است و هیچ هزینه اضافی برای کانترینر ندارد.



ارکستراسیون چند-عاملی (Multi-Agent Beta)



- مکانیزم: یک اینستنس مرکزی، چندین نماینده فرعی را به صورت موازی هماهنگ کرده و نتایج آر‌ها را تجمیع می‌کند.
- مزیت اصلی: کاهش چشمگیر زمان اجرای کل برای فرآیندهای طولانی.
- کاربرد ایده‌آل: تسک‌های پیچیدای که به طور تمیز به جریان‌های کاری مستقل قابل تقسیم هستند.

تکامل پرامپت نویسی: مدل های لاغرتر (Leaner Prompts)

روش جدید (GPT-5.6)

پرامپت های لاغر و دقیق.
قانون طلایی: هر دستور را فقط یکبار بیان کنید و ابزارهای غیرمرتبط را حذف کنید.

روش قدیمی (GPT-5.5)

<code(">

پرامپت های متورم، پر از افزونگی و دستورات تکراری ("حتماً این کار را بکن"، "تاکید می کنم").
"تاکید می کنم".

<code_'1">

<wirt_agent'مُتورم">

افزایش کیفیت



۱۰% تا ۱۵%

کاهش هزینه ها



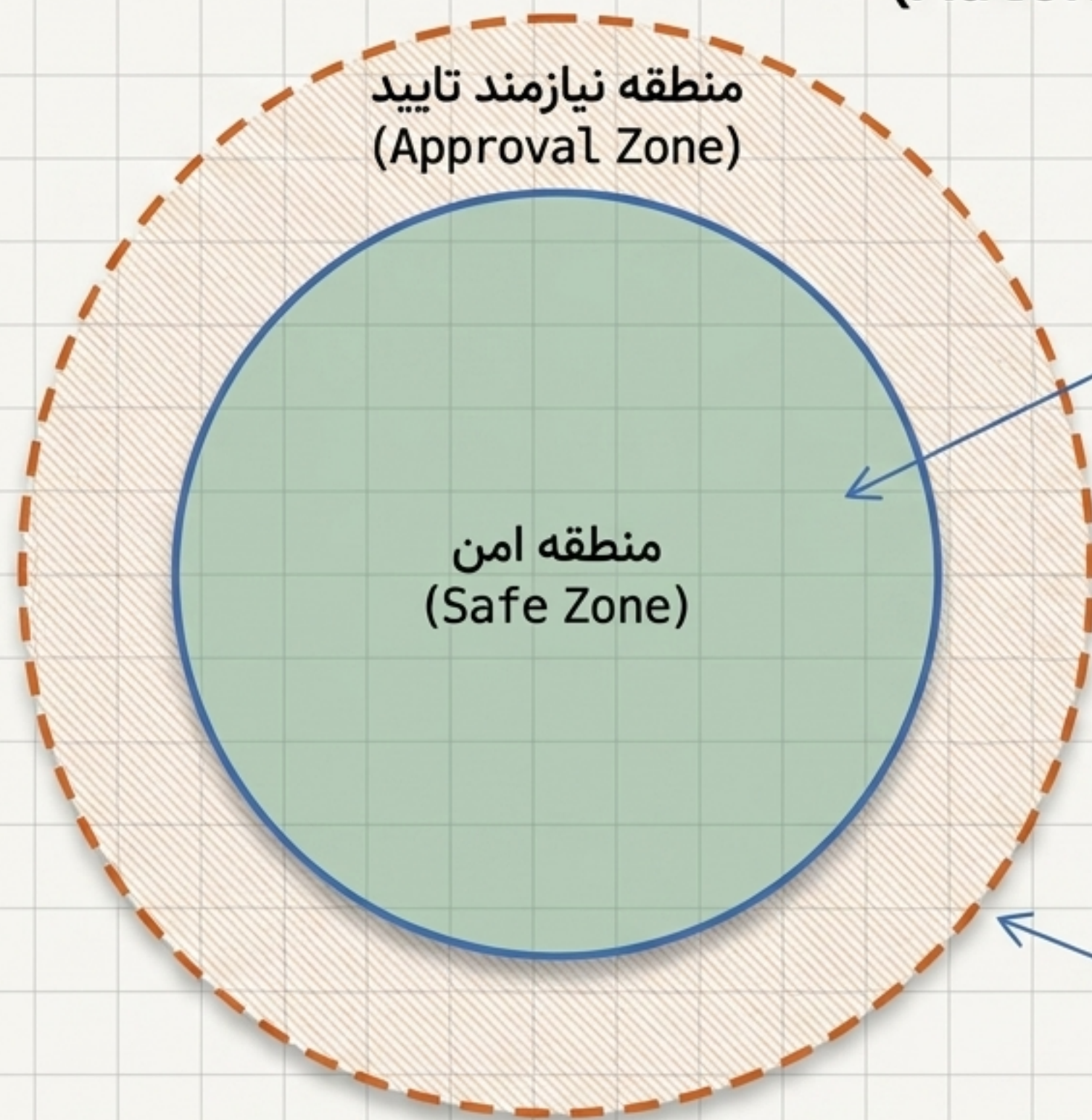
۳۳% تا ۶۷%

کاهش مصرف توکن



۴۱% تا ۶۶%

تعیین مرزها و استقلال مدل (Autonomy Boundaries)



اقدامات ایمن: کارهای محلی مانند خواندن فایل‌ها، بررسی لاگ‌ها و اجرای تست‌ها را صریحاً مجاز کنید تا مدل بی‌دلیل متوقف نشود.

خط قرمزها: مرزهای تخریبی، پرهزینه یا نیازمند تماس با خارج را یک‌بار و به صورت شفاف به عنوان یک خط‌مشی فشرده تعریف کنید.

نکته کلیدی: GPT-5.6 در کارهای چندمرحله‌ای بسیار پیگیر است؛ از تکرار دستوراتی مثل 'اول پرس' برای کارهای امن خودداری کنید.

کنترل طول و لحن پاسخ در سطح API



روش درست (تعریف رفتار)

دقیقاً مشخص کنید چه چیزی حفظ/حذف شود و رفتار را توصیف کنید.

جزئیات فنی را حذف کن و فقط مراحل نصب را نگه دار.

مستقیماً پاسخ بده و مشکل کاربر را تایید کن.

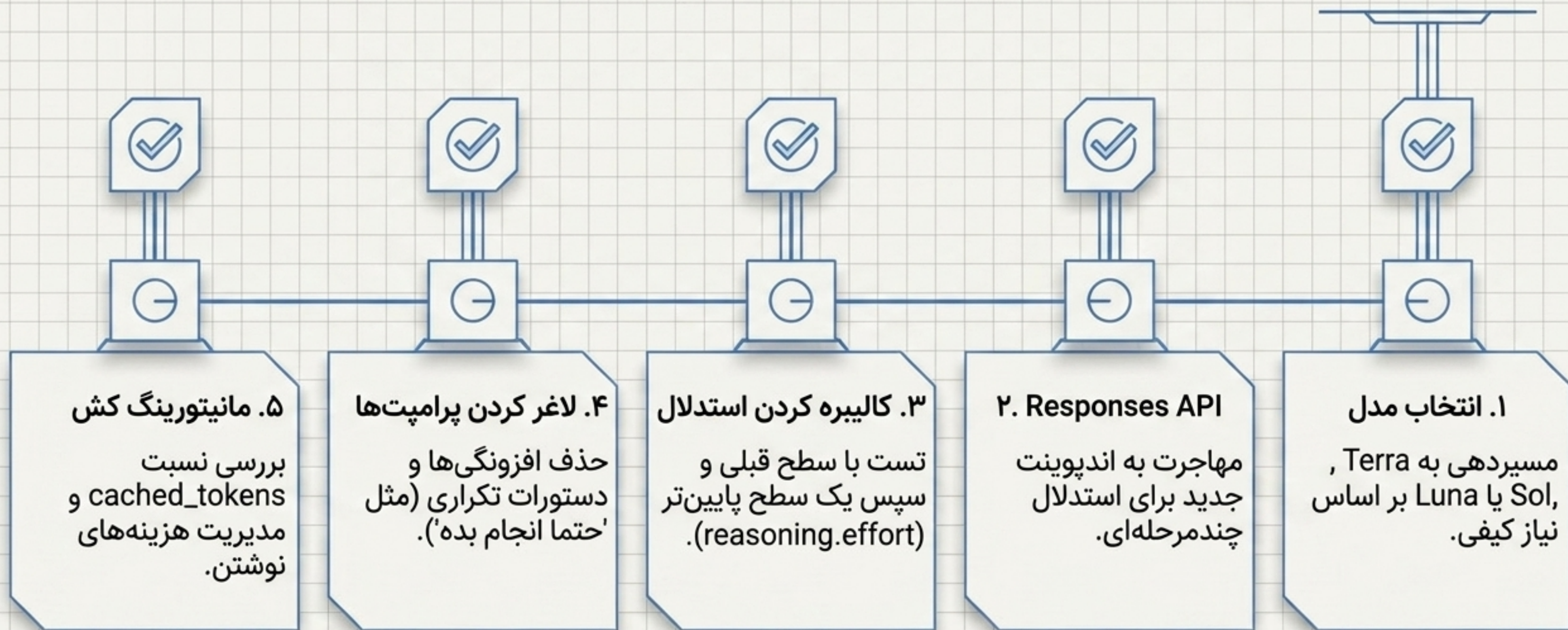
روش غلط (مبهم)

از صفتهای کیفی و کلی استفاده نکنید.

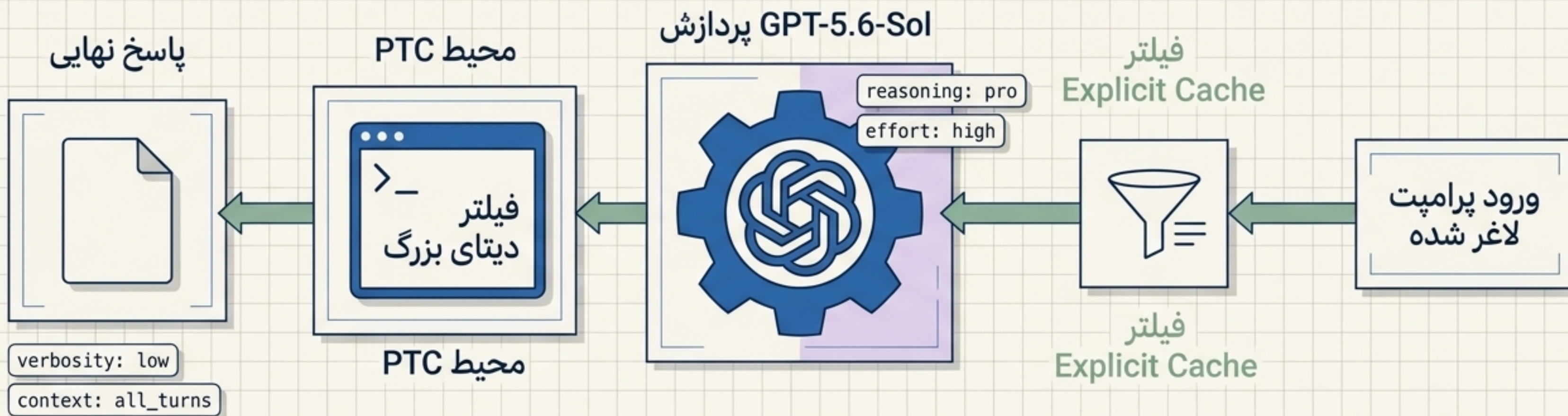
کوتاه بنویس.

همدلانه پاسخ بده.

چک‌لیست مهاجرت سریع (Migration Quickstart)



تصویر کلان: مهندسی دقیق ورودی‌ها = حداکثر قدرت، حداقل هزینه



GPT-5.6 فقط یک مدل هوشمندتر نیست؛ سیستمی است که شما ارکستر قدرت خام آن را بهینه‌سازی و رهبری می‌کنید.