The Team

AI ESTIMATOR · ROUND 2 · 14 EXTRACTION MODELS · 5 IMAGE TOOLS · 5 VIDEO TOOLS · 7 SEPT 2026

أربعة عشر نموذجاً وخمس أدوات صور وخمس فيديو
القراءة لم تتغيّر، والشكّ والصورة تغيّرا

وسّعنا الاختبار إلى ثلاث شركات وثلاث مراحل، بحكم أعمى من قاضيين مستقلّين على ثلاثمئة صورة. أرخص نموذج ما زال يقرأ الأبعاد المطبوعة بلا خطأ؛ نموذج أرخص منه صار الوحيد الذي يرفض إعطاء رقم لرسمة لا تُغلَق؛ وأداة صور واحدة فقط تُقنع القاضيين بأن البلاطة هي البلاطة.

نماذج الاستخراج: ١٤ **أدوات الصور:** ٥ × ٣٠٠ صورة **أدوات الفيديو:** ٥ **الأحكام:** +٦٠٠ **الإنفاق:** ≈٥٢ $ **التاريخ:** ٧ سبتمبر ٢٠٢٦

٠٠ الجواب أولاً: التوليفة لم تتغيّر في القراءة، وتغيّرت في الشكّ والصورة

الخلاصة

وسّعنا المصفوفة من ٧ نماذج Google إلى **١٤ نموذجاً من ثلاث شركات**، وأضفنا مرحلتي **الصورة** (٥ أدوات × ٢٠ بلاطة × ٣ غرف = ٣٠٠ صورة بحكم أعمى من قاضيين مستقلّين) و**الفيديو** (٥ أدوات × ٣ غرف). النتيجة: **أرخص نموذج في القائمة ما زال يقرأ الأبعاد المطبوعة بلا خطأ**. لكن ظهر نموذج أرخص منه يُحسن الشكّ: `gpt-5.6-sol` هو الوحيد الذي **رفض إعطاء رقم** للرسمة التي لا تُغلَق في التشغيلات الخمس كلها. وفي الصور، **Gemini Flash Image** وحده يُنتج ما يقبله القاضيان «البلاطة نفسها» في أكثر من نصف الحالات؛ Recraft وStability عند ٢%.

المرحلة	الموصى به	لماذا	ما تغيّر عن ٦ سبتمبر
قراءة الأبعاد (نصّ · قصاصة · مخطط مصوَّر)	`gemini-3.5-flash-lite`	صفر خطأ في كل الغرف المرجعية. ٢–٥ ثوانٍ، ٠٫١–٠٫٣ سنت — لم يتفوّق عليه أحد من ١٣ نموذجاً آخر	● ثابت
كشف التعارض في رسمة العميل	`gpt-5.6-sol`	٥/٥ على الرسمة الأصلية و٥/٥ على المصغَّرة. يعيد `null` بدل رقم مخمَّن. ٠٫١٨ سنت مقابل ٢٫٠ لـFlash-3.5	● تغيّر كان Flash-3.5 (٤/٥)
الـPDF الخام	● لا يُرسل	لم نرسله لأي نموذج جديد هذه الجولة — نحوّله بـ `pdftoppm -r 150` أولاً	● ثابت
صورة الغرفة بالبلاطة	**Gemini Flash Image** (`gemini-3.1-flash-image`)	«البلاطة نفسها» ٥٧% عند قاضي OpenAI و٩٢% عند قاضي Google · ١١٠ ثانية · ٦٫٧ سنت. Pro أغلى ×٢ بلا فرق، وLite أرخص لكنه يخطئ البلاطة في النصف	● جديد
فيديو الجولة	**ffmpeg حتمي** أولاً	انظر §٠٧ — القاضيان حكما على ثلاثة إطارات من كل مقطع	● جديد

01 ما اختُبر: ثلاث مراحل، ثلاث شركات، وقاضيان لا يعرفان من رسم

الاستخراج: ٧ نماذج جديدة — `gpt-5.6-luna` · `gpt-5.6-sol` · `gpt-5.6-terra` من OpenAI، و`claude-opus-5` · `claude-sonnet-5` · `claude-haiku-4-5` · `claude-opus-4-5` من Anthropic — على السيناريوهات الستّة (بلا الـPDF الخام) بخمس تشغيلات، بنفس البرومبت والحقيقة المرجعية. وأُعيد `gemini-3.1-pro-preview` و`gemini-3.8-flash` من خادم DigitalOcean لأن الشبكة المحلية كانت تقطعهما. المعرّفات كلها من قوائم الخدمات لا مخمَّنة. Opus-5 وSonnet-5 بتفكير تكيّفي وبلا `temperature` (الواجهة ترفضها)؛ الباقي بحرارة `1.0` كالجولة الأولى.

الصور: ٢٠ بلاطة حقيقية من `niceramics.com` (٩ منها عيّنة مسطّحة و١١ لقطة غرفة تعرض البلاطة — وهذا ما يجده العميل فعلاً في الكتالوج) × ٣ غرف فارغة بترخيص مشاع × ٥ أدوات: Gemini Lite/Flash/Pro Image (البلاطة والغرفة كمرجعين في طلب واحد) · Recraft V3 (يُصنع `style` من البلاطة ثم `imageToImage` على الغرفة) · Stability (`control/style-transfer`: الغرفة أساساً والبلاطة أسلوباً). الترقيم أعمى: والحكم بقاضيين من شركتين غير شركة الأداة قدر الإمكان — `gpt-5.6-sol` و`gemini-3.8-flash` — كلٌّ يرى البلاطة المرجعية والصورة المولَّدة ويجيب: هل هي البلاطة نفسها؟ أمانة النقشة واللون /٥؟ هل بقيت الغرفة؟

الفيديو: أفضل صورة لكل غرفة بمجموع درجات القاضيين ← ffmpeg حتمي · Veo 3.1 Fast · Veo 3.1 · Veo 3.1 Lite · Sora 2، ثمانية ثوانٍ 720p، ثم ثلاثة إطارات من كل مقطع إلى القاضيين نفسيهما.

ما لم يكتمل — وسببه

٣٠ خلية من Sonnet-5 وHaiku-4.5 وOpus-4.5 على المخطط الكامل (S5/S6) توقّفت لأن **رصيد Anthropic ($٥) نفد** في منتصفها — محفوظة للاستئناف بعد الشحن. **Mistral OCR 4.1** شُغّل بعد توثيق إطفاء التدريب (٤ مُدخَلات): على المخطط الكامل (S5 وS6) قرأ **جدول العنوان فقط** ولم يلتقط أي رقم بُعد داخل الرسم (٠/٩ في كليهما)، وعلى القصاصة قرأ ٣ من ٥ أبعاد مطبوعة (ضاع 600 وقُرئ 190 على أنه 210)، وعلى الرسمة اليدوية لا شيء (٠/٤) — أقلّ من ثانيتين للمُدخَل. **ليس طبقة قراءة أبعاد**: يعامل الرسم الهندسي كصورة ويُخرج النصوص الجدولية حوله؛ يسقط. القاضي الثاني كان مقرَّراً Sonnet-5 فحلّ محلّه `gpt-5.6-sol` للسبب نفسه، ثم أُضيف Sonnet-5 قاضياً ثالثاً بعد الشحن (عموده في جدول §٠٦). Document AI تُخطّي بقرار، وgpt-image-2 وFLUX وRunway مُقصاة من الدراسة الصفرية.

02 النصّ والقصاصة: أربعة عشر نموذجاً، جواب واحد

رسالة العميل بالأبعاد (S1): **١٤/١٤ نموذجاً أعطى ٤٦,٠٠ م في التشغيلات الخمس كلها**. الفرق الوحيد هو السعر والزمن: GPT-5.6 عند ٠,٠٦–٠,٠٨ سنت، Lite عند ٠,٠٩، Opus-5 عند ٢,٦. قصاصة المخطط (S3):

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحكم
`Lite-3.1` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٢,٣	٠,١٢	● صفر خط
`Lite-3.5` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٢,٣	٠,١٢	● صفر خط
`Flash-3.5` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	١٠,١	١,١٣	● صفر خط
`Flash-3.6` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٨,٠	٠,٨٣	● صفر خط
`Flash-3.7` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٤,٢	٠,٥٣	● صفر خط
`Flash-3.8` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٧,٨	٠,٥٨	● صفر خط
`Pro-3.1` Google	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	١٧,٢	٣,٧٥	● صفر خط
`GPT-5.6-luna` OpenAI	٥/٥	٣–٣	٣٦,٦	٢/٣	١,٨	٥,٨	٠,١١	● جزئي

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحكم
`GPT-5.6-sol` OpenAI	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٨,٩	٠,١١	● صفر خط
`GPT-5.6-terra` OpenAI	٥/٥	٣–٤	٣٦,٣	٣/٣	٠,٠	٤,٨	٠,٠٩	● صفر خط
`Opus-5` Anthropic	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	١٤,٢	٣,٧٤	● صفر خط
`Sonnet-5` Anthropic	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٢٤,٧	٢,٦١	● صفر خط
`Haiku-4.5` Anthropic	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٥,٨	٠,٥١	● صفر خط
`Opus-4.5` Anthropic	٥/٥	٣–٣	٣٦,٣	٣/٣	٠,٠	٧,٢	٢,٤٥	● صفر خط

الخطأ الوحيد في القصاصة كان `gpt-5.6-luna` الذي قرأ الحمّام ٦,٣ بدل ٦,٠ في التشغيلات الخمس (٣٦,٦ بدل ٣٦,٣) — و`gpt-5.6-terra` الذي أضاف فراغاً رابعاً غير موجود في تشغيلة. Haiku-4.5 صحيح في الوسيط لكنه الوحيد الذي تشتّت (٥% بين التشغيلات). ● مؤكد · ١٤٠ نموذجاً × ٥

03 الرسمة التي لا تُغلَق: من يقول «لا أعرف»؟

الرسمة نفسها من الجولة الأولى: ارتفاع اليسار ٣ م واليمين ١ + ٢,٥ = ٣,٥ م. فلها ثلاثة أجوبة «صحيحة» (١١,٠ · ١٣,٠ · ١٣,٢٥) ولا جواب واحد. المقياس هنا ليس الرقم. بل: **هل صرّح النموذج بالتعارض؟**

النموذج	كشف التعارض	القيم التي أعطاها	ثانية	سنت	الحكم
`Lite-3.1`	٠/٥	١٣,٠، ١٣,٢٥	٥,٦	٠,١٠	● صامت
`Lite-3.5`	٠/٥	١١,٠	٤,٨	٠,٠٩	● صامت
`Flash-3.5`	٤/٥	١١,٠، ١٣,٢٥	٢٠,٠	٢,٠١	● يكشف
`Flash-3.6`	٠/٥	١١,٠، ١٣,٠٠	٣٠,٧	٢,٠٧	● صامت
`Flash-3.7`	٠/٥	١١,٥، ١٣,٢٥، ١٥,٠٠، ١٦,٠٠	١٥,٤	١,٤٠	● صامت
`Flash-3.8`	٢/٥	١١,٠، ١١,٥٠، ١٣,٢٥	٦٣,١	٨,٢٣	● أحياناً
`Pro-3.1`	٤/٥	١١,٥، ١٣,٢٥، ١٤,٢٥	٣٥,٩	٦,٨١	● يكشف
`GPT-5.6-luna`	٣/٥	١١,٠، ١١,٥٠	١٧,٠	٠,٢٠	● أحياناً
`GPT-5.6-sol`	٥/٥	لا رقم	١٦,٣	٠,١٨	● يكشف
`GPT-5.6-terra`	٠/٥	١١,٠، ١٣,٠٠	١٨,١	٠,٢٣	● صامت
`Opus-5`	٠/٥	١١,٠، ١٣,٠٠	٢٣,٦	٧,٣٥	● صامت
`Sonnet-5`	٠/٥	١١,٠، ١٣,٠٠	٢٩,٥	٣,٨٦	● صامت
`Haiku-4.5`	٠/٥	١٤,٧٧٥، ١٦,٠٢٥، ١٦,٢٥، ٢٢,٥٠	٥,٧	٠,٥٣	● صامت
`Opus-4.5`	٠/٥	١٢,٢٥، ١٣,٠٠	١٤,١	٣,٣٨	● صامت

النتيجة الجديدة

`gpt-5.6-sol` كشف التعارض في ٥/٥ — وفعل ما لم يفعله أحد: **أعاد `floor_area_m2: null` وكتب في الدليل «الارتفاع الكلي المكتوب ٣ م بينما ٢,٥ + ١ = ٣,٥ م».**

Flash-3.5 وPro-3.1 يكشفان ٤/٥ لكنهما يعطيان رقماً مع ذلك. **عائلة Claude الأربعة صامتة ٠/٥** وتعطي رقماً ثابتاً (١١,٠ أو ١٣,٠) بثقة — وHaiku-4.5 أسوأ: أرقام من ١٤,٨ إلى ٢٢,٥ لا يفسّرها أحد. على النسخة المصغَّرة (S2b) ١٦٠٠px تكرّر الترتيب نفسه: sol ٥/٥، luna ٤/٥، Pro ٣/٥.

ما يعنيه هذا للتصميم: طبقة «الرأي الثاني» لا تحتاج نموذجاً غالياً مفكّراً: تحتاج نموذجاً **مسموحاً له أن يرفض**. sol يكلّف ٠,١٨ سنت للاستدعاء — عُشر Flash-3.5 وثلاثون ضعف أرخص من Pro.

● رسمة واحدة — يُعمَّم بعد ١٠ رسمات

04 المخطط الكامل: من يقرأ الرقم المطبوع، ومن «يقيس» بدلاً منه

الدور الأرضي كصورة راستر (S5) والـPDF نفسه بعد تصييرنا (S6). خمس غرف مرجعية أبعادها مطبوعة على الرسم بالسنتيمتر. المقياس: كم غرفة أصابها النموذج بخطأ ≥ ٢%.

S5 — صورة راستر للدور الأرضي

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحكم
Lite-3.1 Google	٥/٥	١٢–١٤	٢٧١,٧٤	٥/٥	٠,٠	٦,٥	٠,٢٨	● صفر
Lite-3.5 Google	٥/٥	١٣–١٣	٢٨٣,٦٧	٥/٥	٠,٠	٥,١	٠,٢٩	● صفر
Flash-3.5 Google	٥/٥	١٣–١٥	٣٣٨,٧٩	٤/٥	٤,٠	٢٤,٢	٣,١١	● جزئي
Flash-3.6 Google	٥/٥	١٤–١٦	٣٢٤,٣٦	٥/٥	٠,٠	٢٠,٦	٢,٢٩	● صفر

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحكم
`Flash-3.7` Google	٥/٥	١١–١٣	٣٢٥,٢١	٥/٥	٠,٠	٨,٣	١,٢٠	● صفر
`Flash-3.8` Google	٥/٥	١١–١٣	٢٥٣,٧٢	٥/٥	٠,٠	٩,٥	١,٣٩	● صفر
`Pro-3.1` Google	٥/٥	١٣–١٤	٢٣٥,٨٤	٥/٥	٠,٠	٦٣,٠	١٢,١٩	● صفر
`GPT-5.6-luna` OpenAI	٥/٥	١٤–١٦	٢٨٢,٤٦	١/٥	٨,٥	٤٩,٠	٠,٦٠	● ضعيـ يقرأ الأر المطبو. خطأً (١٠ بدل ٥٦٠ ٥٠٠ بدل
`GPT-5.6-sol` OpenAI	٥/٥	١٤–١٨	٢٢٧,١١	٢/٥	٧,٣	٨٠,٨	٠,٧٢	● ضعيـ يقرأ الأر المطبو. خطأً (١٠ بدل ٥٦٠ ٥٠٠ بدل
`GPT-5.6-terra` OpenAI	٥/٥	١٢–١٥	٢١٦,٦٩	٢/٥	١٣,٤	٤٠,٨	٠,٥٤	● ضعيـ يقرأ الأر المطبو. خطأً (١٠ بدل ٥٦٠ ٥٠٠ بدل

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحكم
`Opus-5` Anthropic	٥/٥	١٥–١٨	٣٣٧,٤٢	٥/٤	٤,٠	٧٦,٣	١٧,٩٠	● جزئي المطبخ مقيس الرسم لا من الرقـ المطبو
`Sonnet-5` Anthropic	٥/١	١٥–١٥	٣٩٥,٨٦	٥/٤	٤,٠	٢٠١,٩	١٦,٧٥	● جزئي
`Haiku-4.5` Anthropic	٥/٥	١٢–١٣	٣١٢,١٣	٥/٠	٧٩,١	١٥,٣	١,١٩	● ضعيـ
`Opus-4.5` Anthropic	٥/٥	١٤–١٤	٢١٣,٢٤	٥/٠	٧٢,٤	٢٠,٨	٦,١١	● ضعيـ

S6 — الـPDF نفسه بعد `pdftoppm -r 150`

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحـ
`Lite-3.1` Google	٥/٥	١٣–١٤	٢٥٩,٨٢	٥/٥	٠,٠	٥,٩	٠,٢٩	● ـ
`Lite-3.5` Google	٥/٥	١٥–١٥	٣٦٣,٨١	٥/٤	٤,٠	٥,٣	٠,٣٣	● جـ
`Flash-3.8` Google	٥/٥	١٢–١٤	٢٥٦,٨٨	٥/٥	٠,٠	١١,٣	١,٤٤	● ـ

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحـ
`Pro-3.1` Google	٥/٥	١٢–١٨	٢٦٦,٦٧	٥/٥	٠,٠	٥٢,٧	١٠,٧٠	● ح
`GPT-5.6-luna` OpenAI	٥/٥	١٣–١٦	٢٩٦,٣٢	٣/٥	١٠,٢	٣٥,٣	٠,٥٦	● ح يقرأ المـ خطـ بدل ٥٠٠ ب
`GPT-5.6-sol` OpenAI	٥/٥	١٥–٢٠	٢٢١,٢٥	٤/٥	٢,٢	٩٨,٣	٠,٧٥	● ج الأرق المـ خطـ بدل ٥٠٠ ب
`GPT-5.6-terra` OpenAI	٥/٥	١٢–١٧	٢٨١,٤٨	٣/٥	٣,٧	٣٨,٠	٠,٥٣	● ح يقرأ المـ خطـ بدل ٥٠٠ ب
`Opus-5` Anthropic	٥/٥	١٥–١٦	٣٧١,٧٨	٥/٥	٠,٠	٦١,١	١٦,٥٢	● ح المـ مقي الرس من المـ

النموذج	ناجح	فراغات	المجموع م²	مطابق	خطأ الغرف%	ثانية	سنت	الحكم
`Sonnet-5` Anthropic	٤/٥	١٥–١٥	٣٨٨,٣٦	٤,٠/٥	٧,٥	١٣٨,٧	١٣,٨١	جيد
`Haiku-4.5` Anthropic	٥/٥	١٢–١٤	٢١٩,٠	٠/٥	٥٤,٥	١٤,٧	١,٢١	ضعيف
`Opus-4.5` Anthropic	٥/٥	١٤–١٦	٢٣٥,٧١	١/٥	٢٦,٧	٢٣,٠	٦,٤٢	ضعيف

تفسير

عائلة GPT-5.6 تجد الغرف الخمس كلها لكنها **تقرأ الأرقام المطبوعة خطأً** — `875×590` بدل ٥٦٠، `430×500` بدل ٥٥٠، `250×200` بدل ٢٢٠ — وهي الأرقام نفسها التي أخطأتها Gemini على الـPDF الخام في الجولة الأولى. الفرق أن Gemini تُصيبها على الصورة وGPT لا. وتستغرق ٤٠–١٠٠ ثانية للمخطط (تفكير طويل) مقابل ٥ ثوانٍ لـOpus-5 .Lite يقرأ أربعاً من خمس صحيحة، لكنه في المطبخ ترك الرقم المطبوع و**قاس من الرسم** فأعطى ١٥,٠٥ بدل ٢٣,٦٥ — ثم أصاب الخمس على S6 بتكلفة ١٦,٥ سنت و٦١ ثانية. Lite-3.5 وFlash-3.8 يصيبان الخمس بأقلّ من سنت ونصف. بقية عائلة Claude بعد الشحن: **Sonnet-5** دقيق حين يُكمل (٤/٥ على S6 بخطأ ٧,٥%) لكنه **استنفد سقف ١٦ ألف رمز** في ٤ من ٥ تشغيلات على S5 (تفكير تكيّفي يلتهم الميزانية) عند ١٤٠–٢٠٠ ثانية و١٤–١٧ سنتاً؛ **Haiku-4.5** يجد الغرف الخمس لكن بخطأ ٥٥–٧٩% في مساحاتها (٠/٥ مطابق) — يقرأ الأرقام ولا يربطها بالغرفة الصحيحة؛ **Opus-4.5** ١/٥ مطابق بخطأ ٢٧%. لا واحد منها يقترب من Lite.

05 Pro-3.1 وFlash-3.8 من شبكة الإنتاج: الحكم المعلّق حُسم

في الجولة الأولى قُطع Pro-3.1 ١١ مرة من ٢٥ وعُلّق الحكم عليه. أُعيد النموذجان من droplet في DigitalOcean: **صفر قطيعة شبكية** في ٦٣ خلية (ثلاث `503` من خادم Google نفسه أُعيدت فنجحت). فتبيّن أن Pro-3.1 يصيب الغرف الخمس ٥/٥ على S5 وS6 ويكشف تعارض الرسمة ٤/٥ — لكنه يستغرق ٣٥–

٨٣ ثانية ويكلّف ٣–١٥ سنتاً للاستدعاء، أي **لا يضيف شيئاً على Lite-3.5 في القراءة ولا على sol في الشكّ**. وبثلاثين ضعف سعرهما. Flash-3.8 ثابت كما كان، والقطيعة التي أصابته على صور الجوال الكبيرة كانت شبكية أيضاً.

الدرس الذي يهمّ مسار الإنتاج: **الشبكة المحلية تُسقط أي اتصال يطول نحو دقيقتين** — أصابت OpenAI أيضاً على S5/S6 (٦/٦ قبل الانتقال إلى الخادم). أي قياس زمن أو قطيعة يُجرى من الإنتاج لا من المكتب.

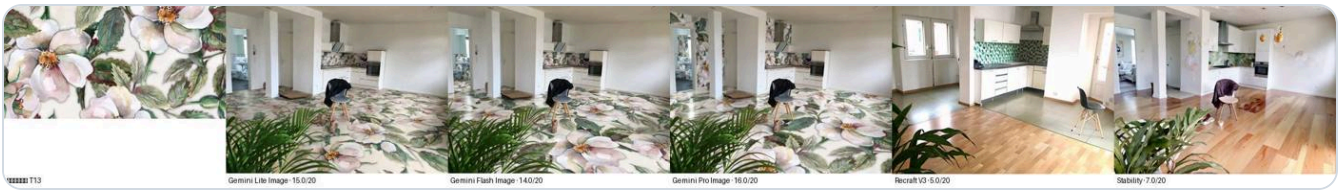
● مؤكد · ٦٣ + ٧٠ خلية من الخادم

06 الصور: أداة واحدة تُقنع القاضيين، واثنتان لا تُقنعان أحداً

٣٠٠ صورة، كلٌّ منها حُكمت مرتين بلا معرفة الأداة. السؤال الذي يهمّ العميل واحد: **هل هذه هي البلاطة التي في الكتالوج؟**

الأداة	صور	«البلاطة نفسها» — قاضي OpenAI	— قاضي Anthropic	— قاضي Google
Gemini Lite Image	٦٠/٦٠	٣٥٪	٣٢٪	٦٨٪
Gemini Flash Image	٦٠/٦٠	٥٧٪	٣٠٪	٩٢٪
Gemini Pro Image	٦٠/٦٠	٥٣٪	٢٧٪	٩٢٪
Recraft V3	٦٠/٦٠	٢٪	٠٪	٢٪
Stability	٦٠/٦٠	٢٪	٣٪	٣٪

ثلاثة قضاة من ثلاث شركات، وثلاث صرامات: قاضي Google أكثر تساهلاً مع صور Google (٩٢%)، قاضي OpenAI في الوسط (٥٧%)، وقاضي Anthropic (Sonnet-5) الأشدّ — يعطي الثلاثة من Gemini نحو ٣٠% بلا فرق بينها (٣٢% · ٣٠% · ٢٧%) ويُبقي Recraft وStability عند ٠% و٣%. ما يتّفق عليه الثلاثة بلا استثناء: **Recraft وStability يسقطان** (٩٣–٩٨% اتفاق على «ليست البلاطة»). ما يختلفون فيه: هل Flash أفضل من Lite وPro — قاضيان يقولان نعم بفارق واضح، والثالث لا يرى فرقاً. لذلك الترتيب المعتمَد **Gemini » Recraft ≈ Stability** مؤكَّد، وتفضيل Flash داخل عائلة Gemini ● مرجَّح لا مؤكَّد — يُحسم بعيّنات مسطّحة وبشر يحكمون.



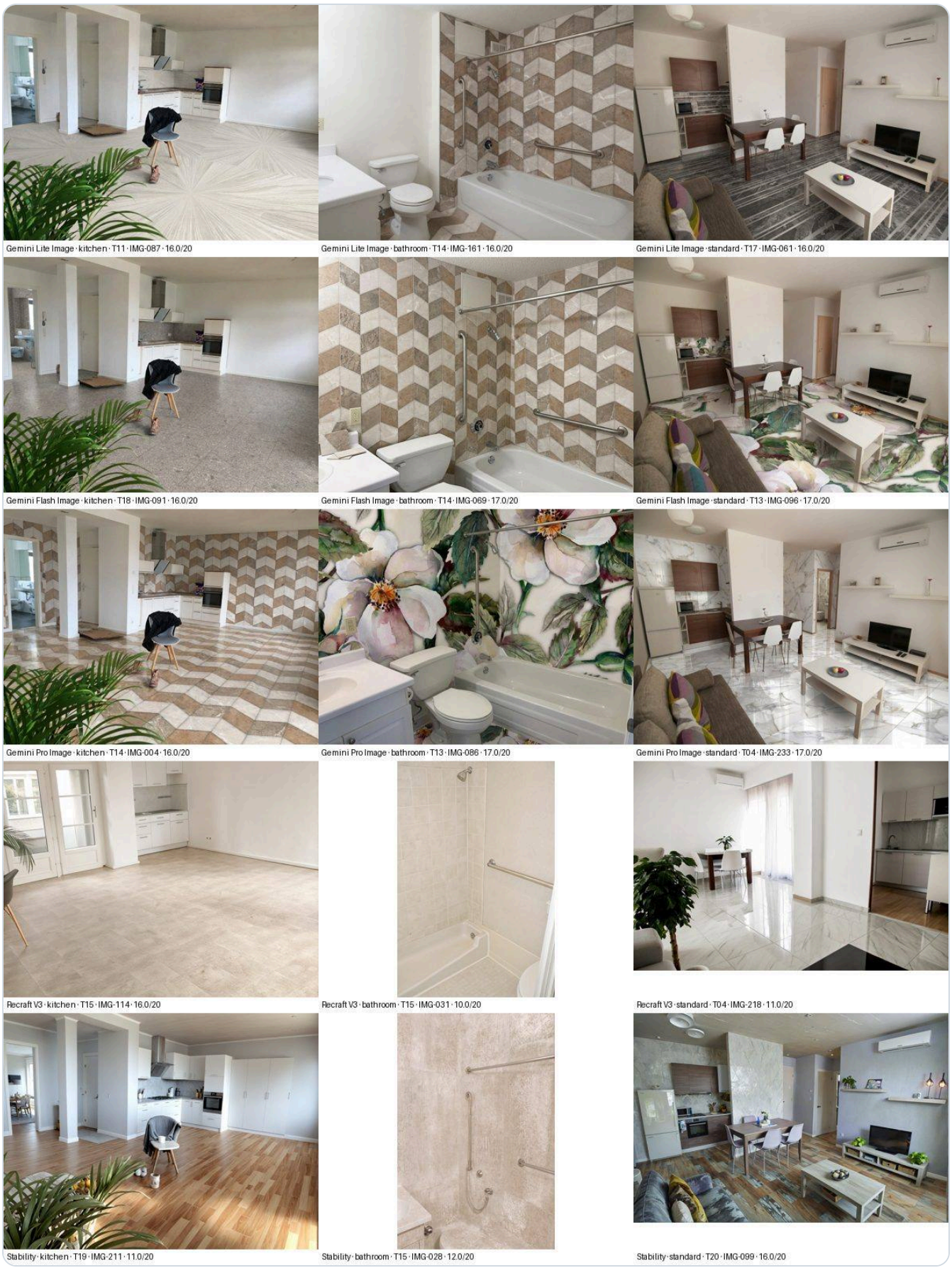
بلاطة زهور مائية (T13) في المطبخ نفسه: المرجع ثم Gemini Lite · Flash · Pro · Recraft · Stability. الثلاثة الأولى تنقل النقشة؛ Recraft يضع بلاطة أخرى؛ Stability يترك الأرضية خشباً ويرسم الزهور على الجدار.

ما الذي يُفشل الأداة؟

الأداة	مرجع عيّنة مسطّحة (٩ بلاطات) نفسها% · نقشة/٥	مرجع لقطة غرفة (١١ بلاطة)	مطبخ نفسها% · غرفة/٥	حمّام
Gemini Lite Image	٢٦٪ · ٢,٨	٤٢٪ · ٣,٠	٣٥٪ · ٣,٥	٤٠٪ · ٣,٤
Gemini Flash Image	٦٧٪ · ٣,٧	٤٨٪ · ٣,١	٧٠٪ · ٤,٠	٥٠٪ · ٤,٠
Gemini Pro Image	٦٣٪ · ٣,٦	٤٥٪ · ٣,٢	٦٠٪ · ٤,٢	٤٥٪ · ٣,٠

الأداة	مرجع عيّنة مسطّحة (٩ بلاطات) نفسها% · نقشة/٥	مرجع لقطة غرفة (١١ بلاطة)	مطبخ نفسها% · غرفة/٥	حمّام
Recraft V3	٤٪ · ٠,٦	٠٪ · ٠,٥	٥٪ · ٣,٩	٠٪ · ٣,٩
Stability	٠٪ · ١,١	٣٪ · ١,٢	٠٪ · ٢,٥	٠٪ · ١,٦

- **نوع المرجع يحسم:** Flash وPro يصيبان «البلاطة نفسها» في ثلثي الحالات حين يكون المرجع عيّنة مسطّحة. وفي أقلّ من النصف حين يكون لقطة غرفة من الكتالوج — لأن الأداة لا تعرف أي سطح في اللقطة هو البلاطة. **الكتالوج الحالي أغلبه لقطات غرف** (١١/٢٠) — وهذا بند تجهيز بيانات، لا بند نموذج.
- **Lite يخطئ البلاطة لا الغرفة:** يحفظ الغرفة (٣,٧/٥) لكنه يضع نقشة قريبة لا مطابقة (٣٥%). أرخص بالنصف من Flash، ولا يساوي الفرق.
- **Pro لا يفوق Flash** (٥٣% مقابل ٥٧%) وهو أبطأ ٢× وأغلى ٢×، وأخرج في بعض الحمّامات صورة مزدوجة جنباً إلى جنب.
- **Recraft** يحفظ الغرفة جيداً (٣,٨/٥) لكن `style` المستخرَج من البلاطة يُنتج «بلاطة بروح المرجع» لا المرجع — ٢% فقط. **Stability** بـstyle-transfer إمّا يُتلف الغرفة (الإعداد الافتراضي) أو يطلي الجدران والسقف بالنقشة ولا يفهم «الأرضية فقط» — حفظ الغرفة ١,٩/٥. كلاهما يسقط بغضّ النظر عن السعر.



أفضل صورة لكل أداة (صفوف) في كل غرفة (أعمدة) بمجموع درجات القاضيين — حتى «أفضل» Recraft وStability ليست البلاطة المطلوبة.

لكل غرفة الصورة الأعلى درجةً عند القاضيين (bathroom: `IMG-244`, kitchen: `IMG-126`, standard: `IMG-096`). المقطع ٨ ثوانٍ 720p، والحكم على ثلاثة إطارات (البداية · المنتصف · النهاية) مقابل الصورة المصدر.

الأداة	مقاطع	النقشة ثابتة — قاضي OpenAI عدد · درجة/٥	— قاضي Google	ثانية	$ للثلاثة	ملاحظة
ffmpeg (حتمي) `ffmpeg zoompan`	٣/٣	٣/٣ · ٤,٣	٣/٣ · ٤	١	٠,٠٠	
Veo 3.1 `veo-3.1-generate-preview`	٣/٣	٢/٣ · ٣	٢/٣ · ٣,٧	٦٨	٩,٦٠	
Veo 3.1 Fast `veo-3.1-fast-generate-preview`	٣/٣	٠/٣ · ١,٧	١/٣ · ٣	٥٧	٢,٤٠	
Veo 3.1 Lite `veo-3.1-lite-generate-preview`	٣/٣	١/٣ · ٣	١/٣ · ٣,٣	٤٦	١,٢٠	
Sora 2 `sora-2`	٣/٣	٢/٣ · ٣,٣	١/٣ · ٣	١٢٦	٢,٤٩	

ما يحدث للنقشة

الإطار الأول في كل مقطع توليدي يطابق المصدر (٤–٥/٥ عند القاضيين) لأنه الصورة نفسها تقريباً: **التدهور يبدأ في المنتصف ويكتمل في الإطار الأخير**: بلاطة الحمّام الشيفرونية تحوّلت عند Veo 3.1 Fast إلى بلاطة أخرى (٥ ← ١)، وعند Sora في الحمّام اختفت من الإطار الأخير (٤ ← ٠). الغرفة الأسهل هي المطبخ (بلاطة خشبية النقشة، حركة أمامية بسيطة) — هناك ثبت Veo 3.1 وSora عند القاضيين. الترتيب: **ffmpeg » Veo 3.1 ≈ Sora 2 > Veo Lite > Veo Fast**. وVeo 3.1 يكلّف ٣,٢ $ للمقطع مقابل ٠,٨٣ $ لـSora وصفر لـffmpeg.

الحمّام: الإطار الأول والأوسط والأخير لأربع أدوات. ffmpeg (يسار) لا يتغيّر؛ الثلاثة الأخرى تبدأ صحيحة وتنتهي ببلاطة أخرى أو بلا بلاطة.

ffmpeg لا يعيد رسم بكسل واحد — النقشة ثابتة **بالتعريف** وبتكلفة حوسبة تُقاس بأجزاء السنت (٠,٦ ثانية للمقطع). التوليدي يُقاس بما يفعله بالنقشة، لا بجمال الحركة.

08 التوليفة الموصى بها بعد الجولتين

القراءة

`gemini-3.5-flash-lite`

كل مُدخَل يبدأ هنا (نصّ · قصاصة · مخطط مصوَّر · PDF بعد تصييرنا). ١٣ نموذجاً آخر لم يتفوّق عليه في غرفة واحدة، وأغلبها أبطأ وأغلى بعشرات الأضعاف.

التحقّق الحتمي

كود، لا نموذج

إغلاق المضلّع، مطابقة العدّ، رفض التشغيلة الشاذّة، و**محلّل JSON متسامح** — ثلاث خلايا هذه الجولة (Opus-5 ×٢، Flash-3.5) أخرجت JSON صحيحاً بفاصلة زائدة أو محرف تحكّم، والمحلّل الصارم يراها فشلاً.

الشكّ — عند العلم فقط

`gpt-5.6-sol`

حين ترفع الطبقة الحتمية علماً: يُستدعى ليقول إن كانت الرسمة لا تُغلَق — ويقولها ٥/٥ ويرفض الرقم. ٠,١٨ سنت و١٦ ثانية. مزوّد ثانٍ من شركة أخرى = احتياط تشغيلي أيضاً.

الصورة

`gemini-3.1-flash-image` + عيّنات مسطّحة

البلاطة والغرفة في طلب واحد، ٦,٧ سنت، ١١ ثانية. الشرط الذي يرفع الإصابة من النصف إلى الثلثين ليس النموذج بل **تصوير كل موديل عيّنةً مسطّحة** بدل لقطة الغرفة.

الفيديو

ffmpeg على الصورة المعتمَدة

الحكم في §٠٧: التوليدي يُضاف فقط حيث يثبت القاضيان أن النقشة لم تتغيّر.

09 ما انقُتل، وما لم يُتحقَّق، وما مصدره واحد

انقُتل في هذه الجولة

- «Flash-3.5 هو أفضل كاشف للتعارض» — sol يفوقه (٥/٥ مقابل ٤/٥) ويرفض الرقم بدل أن يعطيه، بعُشر التكلفة.

- «Pro-3.1 محجوب شبكياً فحكمه معلَّق» — من الخادم لا قطيعة، ولا ميزة.
- «النموذج الأغلى يقرأ المخطط أفضل» — Opus-5 قاس المطبخ من الرسم متجاهلاً الرقم المطبوع؛ GPT-5.6 أخطأ الأرقام المطبوعة نفسها التي أخطأتها Gemini على الـPDF الخام.
- «Recraft مرشّح امتثالياً فيُجرَّب» — امتثاله لا يشفع له: ٢% «البلاطة نفسها».

غير متحقَّق

- أسعار GPT-5.6 (٠,٢٠/١,٢٠ $ للمليون) مفترَضة من الدراسة الصفرية — الرموز مقيسة، السنتات لا. أسعار Anthropic وGemini الصور وVeo منشورة.
- ٣٠ خلية Anthropic على المخطط الكامل لم تكتمل (نفاد الرصيد) — الجدولان S5/S6 يخلوان من Sonnet-5 وHaiku وOpus-4.5.
- القاضيان نموذجان لا بشر. اتفاقهما على الفاشل تامّ، واختلافهما على الناجح واضح — فالنسب المطلقة تُقرأ كترتيب لا كقياس.
- Sora 2 بسعر مفترَض (٠,١٠ $/ث) من الدراسة الصفرية.

مصدر واحد

- رسمة يد واحدة تحكم كل استنتاج §٠٣. عشر رسمات من عملاء حقيقيين قبل تثبيت sol في التصميم.
- مخطط واحد (الدور الأرضي نفسه) لكل نتائج §٠٤.
- ثلاث غرف فقط، ومصوَّرة بترخيص مشاع لا بجوال العميل. غرفة العميل الحقيقية (إضاءة رديئة، زاوية عشوائية) لم تُختبر.

10 الخطوات التالية

01 **شحن رصيد Anthropic (~١٠ $) واستئناف الـ٣٠ خلية + إعادة حكم Sonnet-5 على الصور كقاضٍ ثالث.**

يُكمل الجدولين ويُثبّت أن ترتيب الصور لا يتغيّر بقاضٍ من شركة ثالثة.

---

02 **إسقاط «OCR متخصّص» من بدائل طبقة القراءة في `ai-estimator-spec.md`.**

Mistral OCR 4.1 لا يقرأ أرقام الأبعاد داخل الرسم (٠/٩ على المخطط الكامل) — نماذج الرؤية العامة تفعل، وأرخصها بصفر خطأ.

---

03 **إضافة `gpt-5.6-sol` إلى `ai-estimator-spec.md` بوصفه طبقة الشكّ، مع سياسة «null أفضل من رقم مخمَّن».**

تغيير في العقد بين الطبقات: الطبقة الحتمية تقبل null وتسأل العميل بدل أن تُقدّر.

---

04 **بند تجهيز بيانات في BRD v1.1.0: عيّنة مسطّحة واحدة لكل موديل (٧٠–١٠٠ موديل).**

يرفع إصابة الصورة من النصف إلى الثلثين بلا تغيير نموذج — أرخص تحسين في الجولة كلها.

---

05 **محلّل JSON متسامح في طبقة الاستخراج (فاصلة زائدة · محرف تحكّم · شظيّة بعد الكائن).**

ثلاث خلايا هذه الجولة كانت ستُسجَّل فشلاً وهي صحيحة.

---

06 **عشر رسمات وعشر لقطات غرف من عملاء حقيقيين قبل أي قرار نهائي.**

كل استنتاج في §٠٣ و§٠٦ مبنيّ على مُدخَل واحد أو ثلاثة.

الاستخراج: ٤١٥ خلية لـ١٤ نموذجاً (٢٠٥ Google · ٩٠ OpenAI · ١٢٠ Anthropic) · ١,٠٣٦,٢١٥ رمز دخل · ١,٢٣٣,٢١٢ رمز خرج وتفكير · ٣ خلايا أُصلحت بمحلّل متسامح · ٣٠ خلية معلَّقة على الرصيد. الصور: ٣٠٠ صورة و٦٠٠ حكم. الفيديو: ١٥ مقطعاً. الإنفاق التقديري: استخراج ١١,٥٦ $ (Google ٤,٩٧ · OpenAI ٠,٢٨ · Anthropic ٦,٣٢) · صور ٢١,٥٩ $ · فيديو ١٥,٦٩ $ · حكم ٣,٥≈ $ — **٥٢≈ $ للجولة**.
كل مخرَج خام محفوظ في `out/matrix_raw/` و`out/images2/` و`out/video2/` ويُعاد التحليل بـ `report_matrix.py` و`report_images2.py`.

الحقيقة المرجعية كما في الجولة الأولى: الأبعاد المطبوعة على `plan1.pdf` بتكبير 300dpi، ورسمة يد وقصاصة من المستخدم. البلاط من الموقع العام للشركة؛ الغرف من Wikimedia Commons. لم تُقَس غرفة على الأرض ولم يحكم بشر على صورة.