أصبحت Interactions API (https://ai.google.dev/gemini-api/docs/interactions-overview?hl=ar) متاحة الآن للجميع. ننصحك باستخدام واجهة برمجة التطبيقات هذه للوصول إلى جميع أحدث الميزات والنماذج.

إعدادات الأمن الشخصي

توفّر Gemini API إعدادات أمان يمكنك تعديلها خلال مرحلة إنشاء النموذج الأولي لتحديد ما إذا كان تطبيقك يتطلّب إعدادات أمان أكثر أو أقل تقييدًا. يمكنك تعديل هذه الإعدادات في أربع فئات فلاتر لحظر أنواع معيّنة من المحتوى أو السماح بها.

يشرح هذا الدليل كيفية تعامل Gemini API مع إعدادات الأمان والفلترة، وكيف يمكنك تغيير إعدادات الأمان لتطبيقك.

ملاحظة: قد تخضع التطبيقات التي تستخدم إعدادات أمان أقل تقييدًا للمراجعة. لمزيد من المعلومات، يُرجى الاطّلاع على بنود الخدمة (https://ai.google.dev/gemini-api/terms?hl=ar#use-restrictions).

فلاتر الأمان

تغطّي فلاتر الأمان القابلة للضبط في Gemini API الفئات التالية:

الفئة	الوصف
التحرش	تعليقات سلبية أو ضارة تستهدف الهوية و/أو السمات المحمية
الكلام الذي يحضّ على الكراهية	محتوى وقح أو غير محترم أو بذيء
محتوى جنسي فاضح	محتوى يشير إلى أفعال جنسية أو محتوى فاحش آخر

الفئات الخطيرة	محتوى يروّج لأفعال ضارة أو يسهّلها أو يشجّع عليها

يتم تحديد هذه الفئات في `HarmCategory` (https://ai.google.dev/api/rest/v1/HarmCategory?hl=ar). يمكنك استخدام هذه الفلاتر لتعديل المحتوى المناسب لحالة الاستخدام. على سبيل المثال، إذا كنت تنشئ حوارًا للعبة فيديو، قد تعتبر أنّه من المقبول السماح بمزيد من المحتوى الذي تم تقييمه على أنّه *خطير* بسبب طبيعة اللعبة.

بالإضافة إلى فلاتر الأمان القابلة للضبط، تتضمّن Gemini API وسائل حماية مضمّنة ضد الأضرار الأساسية، مثل المحتوى الذي يعرّض سلامة الأطفال للخطر. يتم حظر هذه الأنواع من الأضرار دائمًا ولا يمكن تعديلها.

مستوى فلترة المحتوى من حيث الأمان

تصنّف Gemini API مستوى احتمالية أن يكون المحتوى غير آمن على أنّه `HIGH` أو `MEDIUM` أو `LOW` أو `NEGLIGIBLE`.

تحظر Gemini API المحتوى استنادًا إلى احتمالية أن يكون المحتوى غير آمن وليس استنادًا إلى مدى خطورته. من المهم أخذ ذلك في الاعتبار لأنّ بعض المحتوى قد يكون لديه احتمالية منخفضة بأن يكون غير آمن على الرغم من أنّ مدى الضرر قد يظل مرتفعًا. على سبيل المثال، عند مقارنة الجملتين:

1. ضربني الروبوت.
2. قطعني الروبوت.

قد تؤدي الجملة الأولى إلى احتمالية أعلى بأن تكون غير آمنة، ولكن قد تعتبر الجملة الثانية أكثر خطورة من حيث العنف. في ضوء ذلك، من المهم إجراء اختبار دقيق وتحديد مستوى الحظر المناسب المطلوب لدعم حالات الاستخدام الرئيسية مع تقليل الضرر الذي يلحق بالمستخدمين النهائيين.

فلترة الأمان لكل طلب

يمكنك تعديل إعدادات الأمان لكل طلب ترسله إلى واجهة برمجة التطبيقات. عند إرسال طلب، يتم تحليل المحتوى وتعيين تقييم أمان له. يتضمّن تقييم الأمان الفئة واحتمالية تصنيف الضرر. على سبيل المثال، إذا تم حظر المحتوى بسبب احتمالية عالية لفئة التحرش، سيكون تقييم الأمان الذي يتم عرضه من الفئة `HARASSMENT`، وسيتم ضبط احتمالية الضرر على `HIGH`.

بسبب الأمان المتأصّل في النموذج، تكون الفلاتر الإضافية **غير مفعّلة** تلقائيًا. إذا اخترت تفعيلها، يمكنك ضبط النظام لحظر المحتوى استنادًا إلى احتمالية أن يكون غير آمن. يغطّي السلوك التلقائي للنموذج معظم حالات الاستخدام، لذا عليك تعديل هذه الإعدادات فقط إذا كان ذلك مطلوبًا باستمرار لتطبيقك.

يصف الجدول التالي إعدادات الحظر التي يمكنك تعديلها لكل فئة. على سبيل المثال، إذا ضبطت إعداد الحظر على **حظر القليل** لفئة **الكلام الذي يحضّ على الكراهية**، يتم حظر كل المحتوى الذي لديه احتمالية عالية بأن يكون كلامًا يحضّ على الكراهية. ولكن يُسمح بأي محتوى لديه احتمالية أقل.

الحدّ (Google AI Studio)	الحدّ (واجهة برمجة التطبيقات)	الوصف
إيقاف	`OFF`	إيقاف فلتر الأمان
بدون حظر	`BLOCK_NONE`	العرض دائمًا بغض النظر عن احتمالية أن يكون المحتوى غير آمن
حظر القليل	`BLOCK_ONLY_HIGH`	الحظر عند وجود احتمالية عالية بأن يكون المحتوى غير آمن
حظر بعض المحتوى	`BLOCK_MEDIUM_AND_ABOVE`	الحظر عند وجود احتمالية متوسطة أو عالية بأن يكون المحتوى غير آمن
حظر معظم المحتوى	`BLOCK_LOW_AND_ABOVE`	الحظر عند وجود احتمالية منخفضة أو متوسطة أو عالية بأن يكون المحتوى غير آمن
لا ينطبق	`HARM_BLOCK_THRESHOLD_UNSPECIFIED`	لم يتم تحديد الحدّ، ويتم الحظر باستخدام الحدّ التلقائي

إذا لم يتم ضبط الحدّ، يكون الحدّ التلقائي للحظر **إيقاف** لطرازَي Gemini 2.5 و3.

يمكنك ضبط هذه الإعدادات لكل طلب ترسله إلى الخدمة التوليدية. لمزيد من التفاصيل، يُرجى الاطّلاع على مرجع واجهة برمجة التطبيقات `HarmBlockThreshold` (https://ai.google.dev/api/generate-content?hl=ar#harmblockthreshold).

ملاحظات الأمان

`generateContent` (https://ai.google.dev/api/generate-content?hl=ar#method:-models.generatecontent) تعرض الدالة `GenerateContentResponse` (https://ai.google.dev/api/generate-content?hl=ar#generatecontentresponse) يتضمّن ملاحظات الأمان.

يتم تضمين ملاحظات الطلب في `promptFeedback` (https://ai.google.dev/api/generate-content?hl=ar#promptfeedback). إذا تم ضبط `promptFeedback.blockReason`، يعني ذلك أنّه تم حظر محتوى الطلب.

يتم تضمين ملاحظات المرشّح للردّ في `Candidate.finishReason` (https://ai.google.dev/api/generate-content?hl=ar#candidate) و `Candidate.safetyRatings` (https://ai.google.dev/api/generate-content?hl=ar#candidate). إذا تم حظر محتوى الردّ وكان `finishReason` هو `SAFETY`، يمكنك فحص `safetyRatings` للحصول على مزيد من التفاصيل. لا يتم عرض المحتوى الذي تم حظره.

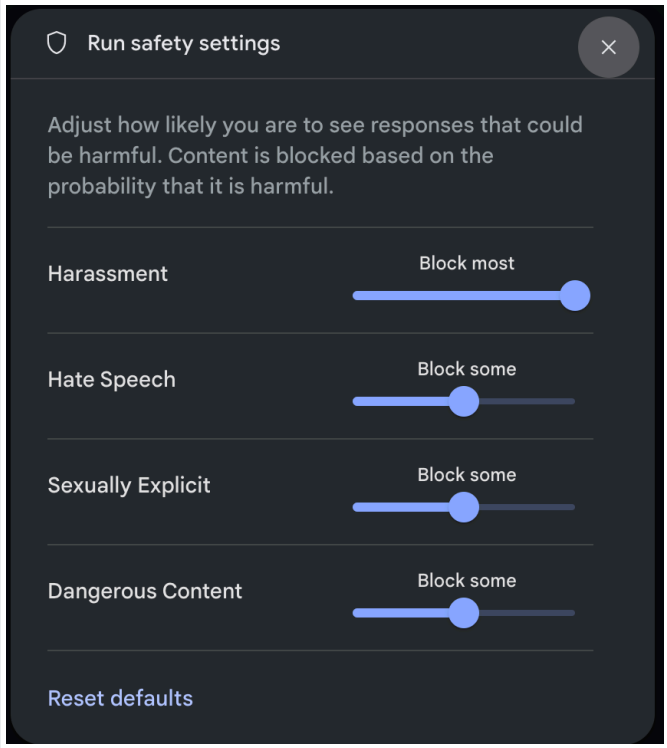
تعديل إعدادات الأمان

يشرح هذا القسم كيفية تعديل إعدادات الأمان في كل من Google AI Studio وفي الرمز البرمجي.

Google AI Studio

يمكنك تعديل إعدادات الأمان في Google AI Studio.

انقر على **إعدادات الأمان** ضمن **الإعدادات المتقدّمة** في لوحة **إعدادات التشغيل** لفتح النافذة المنبثقة **تشغيل إعدادات الأمان**. في النافذة المنبثقة، يمكنك استخدام أشرطة التمرير لتعديل مستوى فلترة المحتوى لكل فئة أمان:



ملاحظة: أنت مسؤول عن التأكّد من أنّ إعدادات الأمان لحالة الاستخدام المقصودة تتوافق مع بنود الخدمة (https://ai.google.dev/gemini-api/terms?hl=ar#use-restrictions).

عند إرسال طلب (على سبيل المثال، من خلال طرح سؤال على النموذج)، تظهر رسالة ⚠ **تم حظر المحتوى** إذا تم حظر محتوى الطلب. للاطّلاع على مزيد من التفاصيل، مرِّر المؤشر فوق النص **تم حظر المحتوى** للاطّلاع على الفئة واحتمالية تصنيف الضرر.

أمثلة للرموز البرمجية

يوضّح مقتطف الرمز البرمجي التالي كيفية ضبط إعدادات الأمان في طلب `GenerateContent`. يضبط هذا الرمز الحدّ لفئة الكلام الذي يحضّ على الكراهية (`HARM_CATEGORY_HATE_SPEECH`). يؤدي ضبط هذه الفئة على `BLOCK_LOW_AND_ABOVE` إلى حظر أي محتوى لديه احتمالية منخفضة أو أعلى بأن يكون كلامًا يحضّ على الكراهية. لفهم إعدادات الحدّ، يُرجى الاطّلاع على فلترة الأمان لكل طلب (#safety-filtering-per-request).

Python انتقال (#انتقال) JavaScript (#javascript) جافا (#جافا) REST (#rest)
(#python)

```
from google import genai
from google.genai import types

client = genai.Client()

response = client.models.generate_content(
    model="gemini-3.5-flash",
    contents="Some potentially unsafe prompt",
    config=types.GenerateContentConfig(
      safety_settings=[
        types.SafetySetting(
            category=types.HarmCategory.HARM_CATEGORY_HATE_SPEECH,
            threshold=types.HarmBlockThreshold.BLOCK_LOW_AND_ABOVE,
        ),
      ]
    )
)

print(response.text)
```

الخطوات التالية

- يمكنك الاطّلاع على مرجع واجهة برمجة التطبيقات (https://ai.google.dev/api?hl=ar) لمعرفة المزيد عن واجهة برمجة التطبيقات الكاملة.
- يمكنك مراجعة إرشادات الأمان (https://ai.google.dev/gemini-api/docs/safety-guidance?hl=ar) للحصول على نظرة عامة على اعتبارات الأمان عند التطوير باستخدام النماذج اللغوية الكبيرة.
- يمكنك التعرّف أكثر على تقييم الاحتمالية مقابل مدى الخطورة من فريق Jigsaw (https://developers.perspectiveapi.com/s/about-the-api-score)
- يمكنك التعرّف أكثر على المنتجات التي تساهم في حلول الأمان، مثل الـ Perspective API (https://medium.com/jigsaw/reducing-toxicity-in-large-language-models-with-perspective-api-c31c39b7a4d7)

. * يمكنك استخدام إعدادات الأمان هذه لإنشاء مصنِّف للمحتوى السام. للبدء، يُرجى الاطّلاع على مثال التصنيف (https://ai.google.dev/examples/train_text_classifier_embeddings?hl=ar).

تاريخ التعديل الأخير: 2026-06-01 (حسب التوقيت العالمي المتفَّق عليه)