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.

إرشادات السلامة والدقة

تعدّ نماذج الذكاء الاصطناعي التوليدي أدوات قوية، ولكنّها ليست خالية من القيود. فقد يؤدي تنوّعها وقابليتها للتطبيق أحيانًا إلى نتائج غير متوقّعة، مثل النتائج غير الدقيقة أو المتحيّزة أو المسيئة. لذلك، من الضروري إجراء معالجة لاحقة وتقييم يدوي دقيق للحدّ من خطر حدوث ضرر من هذه النتائج.

يمكن استخدام النماذج التي توفّرها Gemini API في مجموعة كبيرة من تطبيقات الذكاء الاصطناعي التوليدي ومعالجة اللغة الطبيعية (NLP). ولا يمكن استخدام هذه الوظائف إلا من خلال Gemini API أو تطبيق الويب Google AI Studio. يخضع استخدامك لـ Gemini API أيضًا لـ سياسة الاستخدام المحظور للذكاء الاصطناعي التوليدي (https://policies.google.com/terms/generative-ai/use-policy?hl=ar) و بنود خدمة Gemini API (https://ai.google.dev/terms?hl=ar).

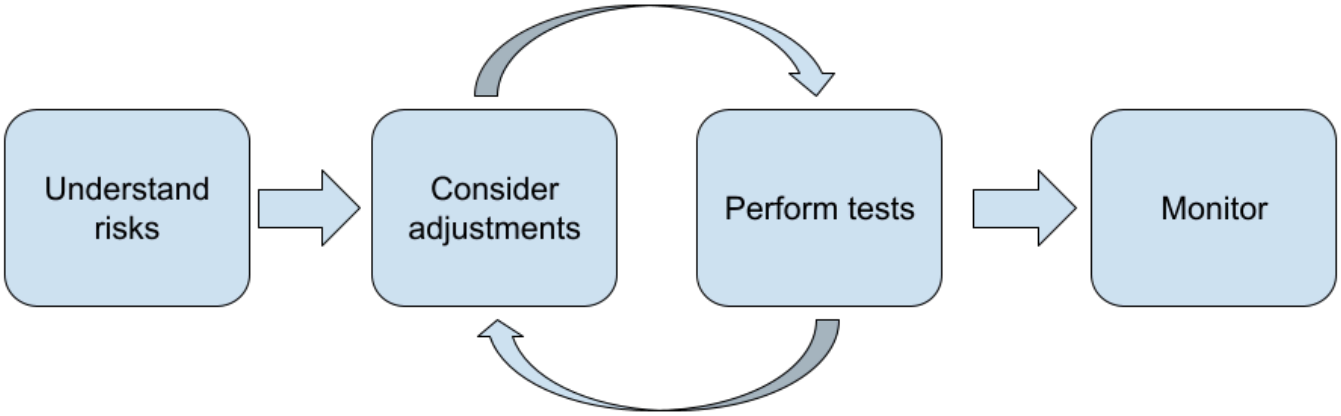
من بين الأسباب التي تجعل النماذج اللغوية الكبيرة (LLM) مفيدة جدًا أنّها أدوات إبداعية يمكنها معالجة العديد من مهام اللغة المختلفة. ولكن هذا يعني أيضًا أنّ النماذج اللغوية الكبيرة يمكنها إنشاء نتائج غير متوقّعة، بما في ذلك نصوص مسيئة أو غير مراعية أو غير صحيحة من الناحية الواقعية. علاوةً على ذلك، فإنّ التنوّع المذهل لهذه النماذج هو أيضًا ما يجعل من الصعب توقّع أنواع النتائج غير المرغوب فيها التي قد تنتجها بالضبط. على الرغم من أنّ Gemini API تم تصميمه مع مراعاة مبادئ الذكاء الاصطناعي من Google (https://ai.google/principles/?hl=ar)، فإنّ مسؤولية تطبيق هذه النماذج بشكل مسؤول تقع على عاتق المطوّرين. لمساعدة المطوّرين في إنشاء تطبيقات آمنة ومسؤولة، يتضمّن Gemini API بعض ميزات فلترة المحتوى المضمّنة بالإضافة إلى إعدادات الأمان القابلة للتعديل على مستوى 4 جوانب من الضرر. يُرجى الرجوع إلى دليل إعدادات الأمان لمزيد من المعلومات. (https://ai.google.dev/gemini-api/docs/safety-settings?hl=ar) توفّر واجهة برمجة التطبيقات أيضًا ميزة "تحديد المصدر من خلال بحث Google" المحسّنة للدقة، ولكن يمكن إيقاف هذه الميزة للمطوّرين الذين تكون حالات استخدامهم أكثر إبداعًا ولا تهدف إلى البحث عن المعلومات.

يهدف هذا المستند إلى تعريفك ببعض المخاطر المتعلّقة بالأمان التي يمكن أن تنشأ عند استخدام النماذج اللغوية الكبيرة، وتقديم توصيات ناشئة بشأن تصميم الأمان وتطويره. (يُرجى العِلم أنّ القوانين واللوائح قد تفرض أيضًا قيودًا، ولكنّ هذه الاعتبارات خارج نطاق هذا الدليل).

يُنصح باتّباع الخطوات التالية عند إنشاء تطبيقات باستخدام النماذج اللغوية الكبيرة:

- فهم المخاطر المتعلّقة بالأمان في تطبيقك
- مراعاة التعديلات اللازمة للتخفيف من المخاطر المتعلّقة بالأمان
- إجراء اختبار الأمان المناسب لحالة استخدامك
- طلب ملاحظات من المستخدمين ومراقبة الاستخدام

يجب أن تكون مرحلتا التعديل والاختبار متكرّرتَين إلى أن تحقّق الأداء المناسب لتطبيقك.



فهم المخاطر المتعلّقة بالأمان في تطبيقك

في هذا السياق، يُعرَّف الأمان بأنّه قدرة النموذج اللغوي الكبير على تجنُّب إلحاق الضرر بمستخدميه، مثلاً من خلال تجنُّب إنشاء لغة سامة أو محتوى يروّج للصور النمطية. تم تصميم النماذج المتاحة من خلال Gemini API مع مراعاة مبادئ الذكاء الاصطناعي من Google (https://ai.google/principles/?hl=ar)، ويخضع استخدامك لها لـ سياسة الاستخدام المحظور للذكاء الاصطناعي التوليدي (https://policies.google.com/terms/generative-ai/use-policy?hl=ar). توفّر واجهة برمجة التطبيقات فلاتر أمان مضمّنة للمساعدة في معالجة بعض المشاكل الشائعة في النماذج اللغوية، مثل اللغة السامة وخطاب الكراهية، والسعي إلى تحقيق الشمولية وتجنُّب الصور النمطية. ومع ذلك، يمكن أن يطرح كل تطبيق مجموعة مختلفة من المخاطر على مستخدميه. لذلك، بصفتك مالك التطبيق، أنت مسؤول عن معرفة مستخدميك والأضرار المحتملة التي قد يسببها تطبيقك، وعن ضمان استخدام تطبيقك للنماذج اللغوية الكبيرة بأمان ومسؤولية.

كجزء من هذا التقييم، عليك مراعاة احتمالية حدوث الضرر وتحديد مدى خطورته وخطوات التخفيف منه. على سبيل المثال، يجب أن يكون التطبيق الذي ينشئ مقالات استنادًا إلى أحداث واقعية أكثر حذرًا بشأن تجنُّب المعلومات المضلِّلة، مقارنةً بتطبيق ينشئ قصصًا خيالية لأغراض الترفيه. من الطرق الجيدة لبدء استكشاف المخاطر المحتملة المتعلّقة بالأمان إجراء بحث عن المستخدمين النهائيين والمستخدمين الآخرين الذين قد يتأثرون بنتائج تطبيقك. يمكن أن يتخذ ذلك أشكالاً عديدة، بما في ذلك البحث عن أحدث الدراسات في مجال تطبيقك، أو مراقبة كيفية استخدام المستخدمين لتطبيقات مماثلة، أو إجراء دراسة أو استطلاع للمستخدمين، أو إجراء مقابلات غير رسمية مع المستخدمين المحتملين.

نصائح متقدمة

- تحدَّث مع مجموعة متنوعة من المستخدمين المحتملين ضمن الفئة المستهدَفة عن تطبيقك والغرض المقصود منه للحصول على منظور أوسع بشأن المخاطر المحتملة وتعديل معايير التنوّع حسب الحاجة.
- يوفّر "إطار عمل إدارة المخاطر في الذكاء الاصطناعي" (https://www.nist.gov/itl/ai-risk-management-framework) الذي أصدره المعهد الوطني للمقاييس والتكنولوجيا (NIST) التابع للحكومة الأمريكية إرشادات أكثر تفصيلاً وموارد تعليمية إضافية لإدارة المخاطر في الذكاء الاصطناعي.
- يصف منشور DeepMind حول الـ مخاطر الأخلاقية والاجتماعية للضرر الناتج عن النماذج اللغوية (https://arxiv.org/abs/2112.04359) بالتفصيل الطرق التي يمكن أن تسبب بها تطبيقات النماذج اللغوية الضرر.

مراعاة التعديلات اللازمة للتخفيف من المخاطر المتعلّقة بالأمان والدقة

بعد فهم المخاطر، يمكنك تحديد كيفية التخفيف منها. يُعدّ تحديد المخاطر التي يجب منحها الأولوية ومقدار ما يجب فعله لمحاولة منعها قرارًا بالغ الأهمية، على غرار تحديد أولويات الأخطاء في مشروع برمجي. بعد تحديد الأولويات، يمكنك البدء في التفكير في أنواع إجراءات التخفيف الأكثر ملاءمة. غالبًا ما يمكن أن تحدث تغييرات بسيطة فرقًا وتحدّ من المخاطر.

على سبيل المثال، عند تصميم تطبيق، ننصحك بمراعاة ما يلي:

- **ضبط مخرجات النموذج** لتعكس بشكل أفضل ما هو مقبول في سياق تطبيقك. يمكن أن يجعل الضبط نتائج النموذج أكثر قابلية للتوقّع والاتساق، وبالتالي يمكن أن يساعد في التخفيف من مخاطر معيّنة.
- **توفير طريقة إدخال تسهّل الحصول على نتائج أكثر أمانًا.** يمكن أن يؤدي الإدخال الدقيق الذي تقدّمه إلى النموذج اللغوي الكبير إلى إحداث فرق في جودة النتائج. إنّ تجربة طلبات الإدخال للعثور على ما يحقّق أفضل النتائج بأمان في حالة استخدامك يستحق الجهد، لأنّه يمكنك بعد ذلك توفير تجربة مستخدم تسهّل ذلك. على سبيل المثال، يمكنك منع المستخدمين من الاختيار إلا من قائمة منسدلة لطلبات الإدخال، أو تقديم اقتراحات منبثقة تتضمّن عبارات وصفية وجدتها آمنة في سياق تطبيقك.
- **حظر الإدخالات غير الآمنة وفلترة النتائج قبل عرضها للمستخدم.** في الحالات البسيطة، يمكن استخدام قوائم الحظر لتحديد الكلمات أو العبارات غير الآمنة في الطلبات أو الردود وحظرها، أو مطالبة المراجعين البشريين بتعديل هذا المحتوى أو حظره يدويًا.

★ **ملاحظة:** يمكن أن يؤدي الحظر التلقائي استنادًا إلى قائمة ثابتة إلى نتائج غير مقصودة، مثل استهداف مجموعة معيّنة تستخدم عادةً المفردات في قائمة الحظر.

- **استخدام المصنّفات المدرَّبة لتصنيف كل طلب على أنّه يتضمّن أضرارًا محتملة أو إشارات مخالفة.** يمكن بعد ذلك استخدام استراتيجيات مختلفة حول كيفية معالجة الطلب استنادًا إلى نوع الضرر الذي تم رصده. على سبيل المثال، إذا كان الإدخال مخالفًا أو مسيئًا بشكل واضح، يمكن حظره وعرض ردّ مكتوب مسبقًا بدلاً منه. **نصيحة متقدمة:** إذا أشارت الإشارات إلى أنّ الناتج ضار، يمكن أن يستخدم التطبيق الخيارات التالية:
  - عرض رسالة خطأ أو ناتج مكتوب مسبقًا
  - إعادة تجربة الطلب، في حال تم إنشاء ناتج آمن بديل، لأنّ الطلب نفسه سيؤدي أحيانًا إلى نتائج مختلفة
- **وضع إجراءات وقائية ضد إساءة الاستخدام المتعمدة** ، مثل تخصيص معرّف فريد لكل مستخدم وفرض حدّ على عدد طلبات المستخدمين التي يمكن إرسالها في فترة معيّنة. من إجراءات الوقاية الأخرى محاولة الحماية من إمكانية حقن الطلبات. إنّ حقن الطلبات، على غرار SQL Injection، هو طريقة يمكن للمستخدمين الضارين من خلالها تصميم طلب إدخال يعالج نتائج النموذج، مثلاً عن طريق إرسال طلب إدخال يطلب من النموذج تجاهُل أي أمثلة سابقة. يُرجى الاطّلاع على سياسة الاستخدام المحظور للذكاء الاصطناعي التوليدي (https://policies.google.com/terms/generative-ai/use-policy?hl=ar) للحصول على تفاصيل حول إساءة الاستخدام المتعمدة.
- **تعديل الوظائف لتصبح أقل خطورة بطبيعتها.** غالبًا ما تكون المهام ذات النطاق الأضيق (مثل استخراج الكلمات الرئيسية من فقرات نصية) أو التي تخضع لإشراف بشري أكبر (مثل إنشاء محتوى قصير سيراجعه مستخدم) أقل خطورة. على سبيل المثال، بدلاً من إنشاء تطبيق لكتابة ردّ على رسالة إلكترونية من البداية، يمكنك بدلاً من ذلك قصر استخدامه على توسيع مخطط أو اقتراح عبارات بديلة.
- **تعديل إعدادات الأمان للمحتوى الضار لتقليل احتمالية ظهور ردود قد تكون ضارة.** توفّر Gemini API إعدادات أمان يمكنك تعديلها خلال مرحلة إنشاء النموذج الأولي لتحديد ما إذا كان تطبيقك يتطلب إعدادات أمان أكثر أو أقل تقييدًا. يمكنك تعديل هذه الإعدادات على مستوى خمس فئات من الفلاتر لتقييد أنواع معيّنة من المحتوى أو السماح بها. يُرجى الرجوع إلى دليل إعدادات الأمان (https://ai.google.dev/gemini-api/docs/safety-settings?hl=ar) للتعرّف على إعدادات الأمان القابلة للتعديل المتاحة من خلال Gemini API.
- **تقليل الأخطاء الواقعية المحتملة أو التخيّلات من خلال تفعيل ميزة "تحديد المصدر من خلال بحث Google"**. يُرجى العِلم أنّ العديد من نماذج الذكاء الاصطناعي تجريبية وقد تعرض معلومات غير دقيقة من الناحية الواقعية أو تتخيّل أو تنتج نتائج أخرى غير مرغوب فيها. تتيح ميزة "تحديد المصدر من خلال بحث Google" ربط نموذج Gemini بمحتوى الويب في الوقت الفعلي، وهي تعمل بجميع اللغات المتاحة. يتيح ذلك لـ Gemini تقديم إجابات أكثر دقة والإشارة إلى مصادر يمكن التحقّق منها بعد تاريخ آخر تحديث للبيانات.

إجراء اختبار الأمان المناسب لحالة استخدامك

يُعدّ الاختبار جزءًا أساسيًا من إنشاء تطبيقات قوية وآمنة، ولكن سيختلف مدى الاختبار ونطاقه واستراتيجياته. على سبيل المثال، من المرجّح أن يطرح مولّد قصائد الهايكو المخصّص للمرح فقط مخاطر أقل خطورة من تطبيق مصمّم مثلاً لاستخدامه من قِبل مكاتب المحاماة لتلخيص المستندات القانونية والمساعدة في صياغة العقود. ولكن قد يستخدم مولّد قصائد الهايكو مجموعة أوسع من المستخدمين، ما يعني أنّ احتمالية محاولات إساءة الاستخدام أو حتى الإدخالات الضارة غير المقصودة يمكن أن تكون أكبر. يُعدّ سياق التنفيذ مهمًا أيضًا. على سبيل المثال، قد يُعتبر التطبيق الذي يراجع نتائجه خبراء بشريون قبل اتخاذ أي إجراء أقل عرضة لإنتاج نتائج ضارة من التطبيق المطابق بدون هذا الإشراف.

من الشائع إجراء عدة تكرارات لإجراء تغييرات واختبارها قبل التأكّد من أنّك مستعد للإطلاق، حتى بالنسبة إلى التطبيقات التي تكون منخفضة المخاطر نسبيًا. هناك نوعان من الاختبارات مفيدان بشكل خاص لتطبيقات الذكاء الاصطناعي:

- **قياس الأداء في ما يتعلق بالأمان** : يتضمّن تصميم مقاييس الأمان التي تعكس الطرق التي يمكن أن يكون بها تطبيقك غير آمن في سياق كيفية استخدامه المحتمل، ثم اختبار مدى جودة أداء تطبيقك على مستوى المقاييس باستخدام مجموعات بيانات التقييم. من الممارسات الجيدة التفكير في الحد الأدنى من المستويات المقبولة لمقاييس الأمان قبل الاختبار حتى 1) تتمكّن من تقييم نتائج الاختبار مقارنةً بهذه التوقعات و2) تتمكّن من جمع مجموعة بيانات التقييم استنادًا إلى الاختبارات التي تقيِّم المقاييس التي تهمّك أكثر.

  نصائح متقدمة:

  - يُرجى الحذر من الإفراط في الاعتماد على الأساليب الجاهزة، لأنّه من المرجّح أن تحتاج إلى إنشاء مجموعات بيانات الاختبار الخاصة بك باستخدام مقيِّمين بشريين لتناسب سياق تطبيقك بالكامل.
  - إذا كان لديك أكثر من مقياس واحد، عليك تحديد كيفية الموازنة إذا أدّى تغيير إلى تحسينات في مقياس واحد على حساب مقياس آخر. كما هو الحال مع هندسة الأداء الأخرى، قد ترغب في التركيز على أسوأ أداء في مجموعة التقييم بدلاً من الأداء المتوسّط.
- **الاختبارات المخالفة** : تتضمّن محاولة إيقاف تطبيقك بشكل استباقي. الهدف هو تحديد نقاط الضعف حتى تتمكّن من اتخاذ خطوات لمعالجتها حسب الاقتضاء. يمكن أن تستغرق الاختبارات المخالفة وقتًا وجهدًا كبيرَين من المقيِّمين الذين لديهم خبرة في تطبيقك، ولكن كلما زاد عدد الاختبارات، زادت فرصتك في رصد المشاكل، لا سيّما المشاكل التي تحدث نادرًا أو بعد عمليات تشغيل متكرّرة للتطبيق فقط.
  - الاختبارات المخالفة هي طريقة لتقييم نموذج تعلُّم الآلة بشكل منهجي بهدف معرفة كيفية أدائه عند تزويده بإدخال ضار أو غير مقصود:
    - قد يكون الإدخال ضارًا عندما يكون مصمّمًا بوضوح لإنتاج ناتج غير آمن أو ضار، مثلاً عن طريق الطلب من نموذج إنشاء نص إنشاء خطاب كراهية عن دين معيّن.

- يكون الإدخال ضارًا غير مقصود عندما يكون الإدخال نفسه غير ضار، ولكنّه ينتج عنه ناتج ضار، مثلاً عن طريق الطلب من نموذج إنشاء نص وصف شخص من عرق معيّن وتلقّي ناتج عنصري.

- ما يميّز الاختبار المخالف عن التقييم العادي هو تكوين البيانات المستخدَمة للاختبار. بالنسبة إلى الاختبارات المخالفة، اختَر بيانات الاختبار التي من المرجّح أن تؤدي إلى نتائج غير مرغوب فيها من النموذج. يعني ذلك فحص سلوك النموذج لجميع أنواع الأضرار المحتملة، بما في ذلك الأمثلة النادرة أو غير العادية والحالات القصوى ذات الصلة بسياسات الأمان. يجب أن يشمل ذلك أيضًا التنوّع في الجوانب المختلفة للجملة، مثل البنية والمعنى والطول. يمكنك الرجوع إلى ممارسات الذكاء الاصطناعي المسؤول من Google في ما يتعلق بالإنصاف (https://ai.google/responsibilities/responsible-ai-practices/?category=fairness&hl=ar) لمزيد من التفاصيل حول ما يجب مراعاته عند إنشاء مجموعة بيانات اختبار. **نصائح متقدمة:**

- استخدِم الاختبارات الآلية (https://www.deepmind.com/blog/red-teaming-language-models-with-language-models?hl=ar) بدلاً من الطريقة التقليدية المتمثلة في الاستعانة بمستخدمين في 'فرق حمراء' لمحاولة إيقاف تطبيقك. في الاختبارات الآلية، يكون "الفريق الأحمر" نموذجًا لغويًا آخر يعثر على نص إدخال يؤدي إلى نتائج ضارة من النموذج الذي يتم اختباره.

★ **ملاحظة:** من المعروف أنّ النماذج اللغوية الكبيرة تنتج أحيانًا نتائج مختلفة لطلب الإدخال نفسه. قد تكون هناك حاجة إلى إجراء جولات اختبار متعددة لرصد المزيد من النتائج غير المرغوب فيها.

مراقبة المشاكل

مهما كان مقدار الاختبار والتخفيف، لا يمكنك ضمان تحقيق الكمال أبدًا، لذا عليك التخطيط مسبقًا لكيفية رصد المشاكل التي تنشأ والتعامل معها. تشمل الطرق الشائعة إعداد قناة مراقَبة ليشارك المستخدمون ملاحظاتهم (مثل التقييم بإبهام لأعلى/لأسفل) وإجراء دراسة للمستخدمين لطلب الملاحظات بشكل استباقي من مجموعة متنوعة من المستخدمين، وهو أمر قيّم بشكل خاص إذا كانت أنماط الاستخدام مختلفة عن التوقعات.

نصائح متقدمة

- عندما يقدّم المستخدمون ملاحظات عن منتجات الذكاء الاصطناعي، يمكن أن يؤدي ذلك إلى تحسين أداء الذكاء الاصطناعي وتجربة المستخدم بشكل كبير بمرور الوقت، مثلاً من خلال مساعدتك في اختيار أمثلة أفضل لضبط الطلبات. يسلّط فصل "الملاحظات والتحكّم" (https://pair.withgoogle.com/chapter/feedback-controls/) في دليل "الأشخاص والذكاء الاصطناعي"

من Google (https://pair.withgoogle.com/guidebook/chapters) الضوء على الاعتبارات الرئيسية التي يجب أخذها في الاعتبار عند تصميم آليات تقديم الملاحظات.

الخطوات التالية

- يُرجى الرجوع إلى دليل إعدادات الأمان (https://ai.google.dev/gemini-api/docs/safety-settings?hl=ar) للتعرّف على إعدادات الأمان القابلة للتعديل المتاحة من خلال Gemini API.
- يُرجى الاطّلاع على مقدّمة عن كتابة الطلبات (https://ai.google.dev/gemini-api/docs/prompting-intro?hl=ar) للبدء في كتابة طلباتك الأولى.

تاريخ التعديل الأخير: 2026-06-05 (حسب التوقيت العالمي المتفَّق عليه)