



Penilaian beban kerja AI generatif

AWS Panduan Preskriptif



AWS Panduan Preskriptif: Penilaian beban kerja AI generatif

Copyright © 2026 Amazon Web Services, Inc. and/or its affiliates. All rights reserved.

Merek dagang dan tampilan dagang Amazon tidak boleh digunakan sehubungan dengan produk atau layanan apa pun yang bukan milik Amazon, dengan cara apa pun yang dapat menyebabkan kebingungan di antara pelanggan, atau dengan cara apa pun yang merendahkan atau mendiskreditkan Amazon. Semua merek dagang lain yang tidak dimiliki oleh Amazon merupakan hak milik masing-masing pemiliknya, yang mungkin atau mungkin tidak terafiliasi, terkait dengan, atau disponsori oleh Amazon.

Table of Contents

Pengantar	1
Tujuan dari panduan ini	2
Target audiens dan manfaatnya	2
Cakupan	2
Hasil bisnis yang ditargetkan	4
Pertimbangan dan prasyarat penilaian	7
Mulailah dengan kasus penggunaan yang jelas	7
Pastikan keselarasan bisnis	8
Menerapkan tata kelola dan pengawasan	8
Data alamat dan prasyarat teknis	8
Pertimbangkan persyaratan sumber daya komputasi	8
Mengatasi implikasi privasi dan keamanan	9
Libatkan pemangku kepentingan lebih awal	9
Iterasi dan pelajari	9
Kuesioner penilaian beban kerja AI generatif	10
Kesiapan	10
Kasus penggunaan	12
Arsitektur	16
Penyimpanan	17
Peraturan dan kepatuhan	18
Integrasi	19
Pengujian	21
Penerapan dan otomatisasi	22
Strategi data	25
Menerjemahkan wawasan penilaian menjadi hasil yang dapat ditindaklanjuti	28
Langkah selanjutnya	30
Pertanyaan yang Sering Diajukan	31
Apa tujuan utamanya?	31
Siapa yang harus menggunakan penilaian ini?	31
Apa komponen kuncinya?	31
Bagaimana ini membantu mendefinisikan arsitektur?	31
Apa manfaatnya?	31
Bagaimana kita bisa mengimplementasikan ini dengan sukses?	32
Apa tantangannya?	32

Apa persyaratan peraturan dan kepatuhan?	32
Apa peran pemangku kepentingan?	32
Bagaimana kita bisa mengukur kesuksesan?	33
Bagaimana pendekatannya berbeda berdasarkan ukuran organisasi?	33
Sumber daya	34
Riwayat dokumen	35
.....	xxxvi

Penilaian beban kerja AI generatif

Tabby Ward dan Deepak Dixit, Amazon Web Services (AWS)

November 2024 ([riwayat dokumen](#))

Penilaian beban kerja AI generatif adalah metode strategis yang bertujuan untuk mengevaluasi dan meningkatkan kesiapan organisasi untuk membuat atau memperbarui beban kerja AI generatifnya. Penilaian ini penting karena memasukkan AI generatif ke dalam operasi bisnis dapat sangat mengubah cara kerja, dan dapat memberikan efisiensi dan kemampuan baru. Namun, untuk mengadopsi AI generatif dengan sukses, penting untuk benar-benar memahami sistem saat ini dan memiliki rencana yang jelas untuk masa depan.

Beban kerja AI generatif mengacu pada tugas komputasi yang melibatkan penggunaan model kecerdasan buatan yang dapat membuat konten baru, seperti teks, gambar, kode, atau tipe data lainnya. Beban kerja ini biasanya membutuhkan daya komputasi yang besar, perangkat keras khusus seperti GPUs, dan kumpulan data besar untuk pelatihan dan inferensi. Mengintegrasikan beban kerja AI generatif ke dalam operasi menghadirkan beberapa tantangan:

- **Persyaratan infrastruktur:** Menyediakan sumber daya komputasi yang signifikan dan perangkat keras khusus yang dibutuhkan model AI generatif.
- **Manajemen data:** Memastikan kualitas data, privasi, dan kepatuhan saat menangani kumpulan data besar.
- **Kesenjangan keterampilan:** Kurangnya keahlian dalam teknologi AI dan penerapan model.
- **Pertimbangan etis:** Mengatasi bias, keadilan, dan transparansi dalam konten yang dihasilkan AI.
- **Kompleksitas integrasi:** Menggabungkan AI generatif dengan mulus ke dalam alur kerja dan sistem lama yang ada.
- **Manajemen biaya:** Menyeimbangkan manfaat potensial dengan biaya implementasi dan operasi yang tinggi.

Mengatasi tantangan ini membutuhkan perencanaan yang matang, investasi dalam infrastruktur dan bakat, dan pendekatan strategis untuk implementasi.

Tujuan dari panduan ini

AI generatif dengan cepat menjadi komponen penting di banyak industri. Ini memberikan peluang transformatif tetapi juga menimbulkan tantangan dalam hal integrasi, kepatuhan, dan skalabilitas. Banyak organisasi berjuang untuk sepenuhnya memanfaatkan AI karena fondasi teknologi yang lemah, resistensi terhadap perubahan, dan masalah kualitas data. Penilaian beban kerja AI generatif mengatasi tantangan ini dengan mengidentifikasi persyaratan untuk modernisasi, mendefinisikan ruang lingkup implementasi, dan menantang sistem dan pemikiran warisan. Ini juga membantu dalam menentukan produk minimum yang layak (MVPs) dan membantu Anda mengembangkan arsitektur solusi target, memastikan pendekatan terstruktur dan strategis untuk adopsi AI.

Panduan ini berfungsi sebagai pendekatan terstruktur untuk membantu organisasi menavigasi kompleksitas mengadopsi teknologi AI generatif. Alih-alih mendefinisikan persyaratan dengan jelas sejak awal, panduan ini membantu dalam:

- Mengidentifikasi kasus penggunaan potensial untuk AI generatif dalam organisasi Anda.
- Menilai kesiapan organisasi Anda untuk adopsi AI generatif.
- Mendefinisikan dan menyempurnakan tujuan kasus penggunaan dan meregangkan tujuan.
- Menentukan ruang lingkup dan persyaratan untuk implementasi AI generatif.
- Mengembangkan arsitektur solusi target.

Target audiens dan manfaatnya

Penilaian ini dirancang khusus untuk arsitek solusi, arsitek perusahaan, dan arsitek aplikasi yang ingin mengevaluasi aspek teknis modernisasi beban kerja AI generatif. Ini juga berharga bagi manajer program dan orang yang ingin mengukur kesiapan tim mereka secara keseluruhan, alokasi sumber daya, dan persyaratan pemberdayaan. Praktik terbaik industri menekankan pentingnya penilaian komprehensif untuk memastikan kesiapan adopsi AI. Ini termasuk mengevaluasi arsitektur, penyimpanan, kepatuhan, integrasi, pengujian, penyebaran, dan otomatisasi.

Cakupan

Topik-topik berikut adalah cakupan metode penilaian beban kerja AI generatif:

- Teknologi dan model AI generatif saat ini (misalnya, model bahasa besar, model pembuatan gambar)

- Aplikasi AI sempit yang menggunakan teknik generatif
- Integrasi AI generatif dengan sistem dan alur kerja yang ada
- Strategi data untuk pelatihan dan penyempurnaan model AI generatif
- Pertimbangan etis dan praktik AI yang bertanggung jawab untuk aplikasi AI generatif saat ini
- Strategi pengujian dan penerapan untuk AI generatif di lingkungan produksi
- Pertimbangan keamanan dan privasi untuk implementasi AI generatif
- Optimalisasi kinerja dan skalabilitas beban kerja AI generatif
- Kasus penggunaan dan aplikasi AI generatif di berbagai industri
- Evaluasi output AI generatif dan proses jaminan kualitas

Topik-topik berikut berada di luar cakupan:

- Skenario kecerdasan umum buatan (AGI) dan kecerdasan super buatan (ASI)
- Kemajuan spekulatif future dalam AI di luar model generatif saat ini
- Aplikasi komputasi kuantum dalam AI
- Komputasi neuromorfik dan antarmuka otak-komputer
- Kesadaran dan kesadaran diri dalam sistem AI
- Dampak sosial jangka panjang dari AI canggih di luar aplikasi AI generatif saat ini
- Kerangka kerja regulasi untuk teknologi AI masa depan hipotetis
- Perdebatan filosofis tentang sifat kecerdasan dan kesadaran dalam mesin
- Kasus tepi ekstrim atau kasus penggunaan AI yang sangat spekulatif
- Spesifikasi teknis terperinci dari model atau arsitektur AI berpemilik

Hasil bisnis yang ditargetkan

Penilaian beban kerja AI generatif bertujuan untuk memberikan beberapa hasil yang ditargetkan yang sangat penting untuk berhasil memodernisasi beban kerja AI generatif. Hasil ini memastikan bahwa organisasi dipersiapkan dengan baik untuk mengintegrasikan teknologi AI secara efektif dan efisien.

Untuk setiap hasil yang ditargetkan, penilaian beban kerja AI generatif berfokus pada:

- **Interdependensi:** Identifikasi dan klarifikasi setiap ketergantungan antara hasil dan aspek lain dari proses modernisasi. Ini termasuk memahami bagaimana satu hasil dapat mempengaruhi atau dipengaruhi oleh orang lain, untuk memastikan pendekatan holistik terhadap modernisasi.
- **Penyelarasan pemangku kepentingan:** Garis besar strategi untuk menyelaraskan berbagai pemangku kepentingan dengan setiap hasil. Ini melibatkan mengkomunikasikan nilai dan dampak dari setiap hasil ke tingkat organisasi dan departemen yang berbeda, untuk mendorong dukungan dan dukungan.
- **Prioritas:** Dalam kasus di mana beberapa kasus penggunaan atau hasil diidentifikasi, sediakan kerangka kerja untuk memprioritaskannya berdasarkan faktor-faktor seperti dampak bisnis, persyaratan sumber daya, dan penyelarasan strategis.
- **Perbaikan berkelanjutan:** Untuk setiap hasil, buat mekanisme untuk evaluasi dan penyempurnaan yang berkelanjutan. Ini memastikan bahwa upaya modernisasi tetap adaptif dan responsif terhadap perubahan lanskap teknologi dan kebutuhan bisnis.

Berikut adalah diskusi rinci dari setiap hasil yang ditargetkan:

Arsitektur target

- **Definisi:** Penilaian membantu menentukan arsitektur target yang jelas dan dapat diskalakan untuk beban kerja AI generatif.
- **Komponen:** Ini termasuk memilih layanan cloud yang sesuai, merancang jaringan data, dan memastikan interoperabilitas sistem.
- **Manfaat:** Arsitektur yang terdefinisi dengan baik mendukung skalabilitas, keandalan, dan optimalisasi kinerja, dan memberikan dasar yang kuat untuk modernisasi.

Kesiapan pelanggan

- **Evaluasi:** Menilai keadaan infrastruktur, proses, dan budaya organisasi saat ini untuk menentukan kesiapan adopsi modernisasi AI generatif.
- **Kriteria:** Ini melibatkan evaluasi kemampuan teknis, kualitas data, dan kemauan organisasi untuk merangkul perubahan.
- **Hasil:** Mengidentifikasi kesenjangan dan area untuk perbaikan memastikan bahwa organisasi siap untuk transisi yang mulus ke solusi dan teknologi modern.

Gunakan tujuan kasus dan regangkan tujuan

- **Tujuan kasus penggunaan** menetapkan tujuan yang jelas untuk implementasi solusi target, dengan fokus pada masalah atau peluang bisnis tertentu.

Tujuan kasus penggunaan dalam konteks modernisasi AI generatif mengacu pada tujuan spesifik dan terukur yang ingin dicapai organisasi dengan menerapkan solusi AI generatif. Tujuan ini biasanya selaras dengan tujuan bisnis yang lebih luas dan fokus pada mengatasi tantangan atau peluang tertentu dalam organisasi. Contoh tujuan kasus penggunaan mungkin termasuk:

- Mengurangi waktu respons layanan pelanggan hingga 50 persen dengan menggunakan chatbots generatif yang didukung AI.
- Meningkatkan efisiensi peninjauan kode sebesar 30 persen melalui analisis kode berbantuan AI generatif.
- Meningkatkan akurasi deteksi penipuan sebesar 25 persen dengan menggunakan pengenalan pola AI generatif.
- **Tujuan peregangan** menentukan target ambisius yang mendorong batas-batas apa yang dapat dicapai oleh modernisasi AI generatif dalam organisasi.
- **Dampak:** Menetapkan tujuan yang dapat dicapai dan aspirasional membantu menyelaraskan inisiatif modernisasi AI generatif dengan tujuan bisnis strategis dan mendorong inovasi.

Estimasi upaya

- **Tujuan:** Estimasi upaya yang akurat membantu dalam perencanaan sumber daya dan memastikan bahwa proyek disampaikan tepat waktu dan sesuai anggaran.
- **Lingkup:** Perkirakan sumber daya, waktu, dan anggaran yang diperlukan untuk mengimplementasikan rencana modernisasi AI generatif.
- **Faktor:** Pertimbangkan kompleksitas teknis, tantangan integrasi, dan potensi risiko.

Kebutuhan pemberdayaan

- Pelatihan dan pengembangan: Identifikasi keterampilan dan pengetahuan yang diperlukan untuk adopsi modernisasi AI generatif yang sukses.
- Sumber Daya: Menentukan kebutuhan akan program pelatihan, lokakarya, dan kegiatan pemberdayaan lainnya.
- Hasil: Memastikan bahwa staf dilengkapi dengan keterampilan yang diperlukan meningkatkan efektivitas inisiatif modernisasi AI generatif dan mendukung kesuksesan jangka panjang.

Rencana implementasi

- Peta Jalan: Kembangkan rencana terperinci yang menguraikan langkah-langkah yang diperlukan untuk mencapai modernisasi AI generatif.
- Tonggak sejarah: Tentukan tonggak penting dan kiriman untuk melacak kemajuan.
- Manfaat: Rencana implementasi yang jelas memberikan arahan dan akuntabilitas, dan memfasilitasi pendekatan terstruktur untuk modernisasi AI generatif.

Pertimbangan dan prasyarat penilaian

Mulailah dengan kasus penggunaan yang jelas

Identifikasi masalah atau peluang bisnis tertentu yang dapat diatasi oleh AI generatif. Fokus pada kasus penggunaan yang selaras dengan tujuan bisnis strategis dan menawarkan manfaat yang terukur. Prioritaskan kasus penggunaan yang menargetkan tantangan yang sering dihadapi dalam organisasi untuk memastikan bahwa arsitektur solusi dapat berfungsi sebagai pola untuk beberapa skenario.

Memulai proses penilaian dengan pemahaman umum tentang aplikasi AI generatif potensial bermanfaat tetapi tidak wajib. [Kuesioner](#) yang disertakan dengan panduan ini mengakomodasi berbagai tingkat kesiapsiagaan, dari organisasi yang memiliki kasus penggunaan yang terdefinisi dengan baik hingga yang hanya memiliki gagasan luas. Proses penilaian berfungsi untuk:

- Sempurnakan dan klarifikasi ide kasus penggunaan awal ini.
- Identifikasi kasus penggunaan potensial baru.
- Kembangkan tujuan spesifik dan terukur untuk setiap kasus penggunaan.
- Menilai kelayakan dan dampak potensial dari setiap kasus penggunaan.

Mari kita pertimbangkan contoh hipotetis: Perusahaan jasa keuangan memutuskan untuk mengeksplorasi modernisasi AI generatif. Mereka mulai dengan ide luas untuk meningkatkan layanan pelanggan dan proses deteksi penipuan mereka.

- Penilaian awal: Kuesioner membantu mereka mengevaluasi sistem mereka saat ini, kualitas data, dan kesiapan organisasi untuk adopsi AI generatif.
- Penyempurnaan kasus penggunaan: Melalui proses penilaian, mereka menyempurnakan ide awal mereka menjadi dua kasus penggunaan spesifik:
 - Menerapkan chatbot bertenaga AI generatif untuk pertanyaan pelanggan
 - Menggunakan AI generatif untuk deteksi penipuan transaksi real-time
- Pengaturan tujuan: Untuk setiap kasus penggunaan, mereka menentukan tujuan tertentu:
 - Mengurangi waktu respons layanan pelanggan sebesar 40 persen dalam 6 bulan
 - Tingkatkan akurasi deteksi penipuan sebesar 20 persen dan kurangi false positive sebesar 15 persen

- Peregangan tujuan: Mereka juga menetapkan target ambisius ini:
 - Raih kepuasan pelanggan 80 persen dengan respons yang dibantu AI
 - Mengembangkan model deteksi penipuan prediktif yang mengidentifikasi pola penipuan baru
- Definisi MVP: Kuesioner membantu mereka menentukan MVP untuk setiap kasus penggunaan, dengan fokus pada fitur-fitur penting yang memberikan nilai langsung.
- Arsitektur target: Akhirnya, mereka mengembangkan arsitektur target yang mendukung satu atau kedua kasus penggunaan, dan memastikan skalabilitas dan integrasi dengan sistem yang ada.

Pastikan keselarasan bisnis

Selaraskan inisiatif AI generatif dengan strategi dan tujuan bisnis secara keseluruhan. Untuk setiap kasus penggunaan, kembangkan proposisi nilai yang jelas yang menunjukkan bagaimana AI generatif berkontribusi pada pertumbuhan, efisiensi, atau inovasi bisnis. Tetapkan metrik untuk mengukur dampak implementasi AI generatif pada indikator kinerja utama (). KPIs

Menerapkan tata kelola dan pengawasan

Buat komite pengarah lintas fungsi untuk mengawasi inisiatif AI generatif. Kembangkan kebijakan dan pedoman untuk penggunaan AI yang bertanggung jawab, mengatasi pertimbangan etis dan potensi bias. Menetapkan proses peninjauan untuk proyek AI generatif untuk memastikan kepatuhan dengan standar organisasi dan persyaratan peraturan.

Data alamat dan prasyarat teknis

Menilai dan meningkatkan kualitas data, dan menerapkan praktik tata kelola data untuk memastikan input yang andal untuk model AI generatif. Kembangkan strategi data yang membahas pengumpulan data, penyimpanan, dan manajemen yang khusus untuk kebutuhan AI generatif. Mengevaluasi dan meningkatkan infrastruktur data untuk mendukung volume dan kecepatan data yang diperlukan untuk beban kerja AI generatif.

Pertimbangkan persyaratan sumber daya komputasi

Menilai infrastruktur TI saat ini dan mengidentifikasi kesenjangan dalam kapasitas komputasi untuk beban kerja AI generatif. Rencanakan sumber daya komputasi yang dapat diskalakan, dengan mempertimbangkan opsi seperti layanan cloud atau kluster komputasi berkinerja tinggi lokal.

Optimalkan alokasi sumber daya untuk menyeimbangkan kinerja dan efektivitas biaya untuk beban kerja pelatihan dan inferensi.

Mengatasi implikasi privasi dan keamanan

Menerapkan langkah-langkah keamanan yang kuat untuk melindungi data sensitif yang digunakan dalam pelatihan dan operasi AI generatif. Pastikan kepatuhan terhadap peraturan perlindungan data seperti Peraturan Perlindungan Data Umum (GDPR) atau California Consumer Privacy Act (CCPA) saat menangani informasi pribadi. Kembangkan protokol untuk penerapan dan pemantauan model yang aman untuk mencegah akses yang tidak sah atau penyalahgunaan kemampuan AI generatif.

Libatkan pemangku kepentingan lebih awal

Libatkan pemangku kepentingan utama sejak awal untuk mendapatkan dukungan dan dukungan kepemimpinan. Komunikasikan dengan jelas manfaat dan dampak potensial dari inisiatif modernisasi, khususnya untuk beban kerja AI generatif. Memberikan pelatihan dan sumber daya untuk membantu para pemangku kepentingan memahami teknologi AI generatif dan implikasinya.

Iterasi dan pelajari

Mengadopsi pendekatan inkremental yang memungkinkan Anda menyempurnakan solusi target. Gunakan loop umpan balik untuk terus meningkatkan arsitektur dan proses beban kerja. Secara teratur menilai kinerja dan dampak implementasi AI generatif, dan sesuaikan strategi sesuai kebutuhan berdasarkan hasil dunia nyata dan kebutuhan bisnis yang berkembang.

Kuesioner penilaian beban kerja AI generatif

Bagian berikut memberikan pertanyaan yang dapat Anda gunakan untuk mengevaluasi berbagai aspek modernisasi beban kerja AI generatif untuk organisasi Anda. Kuesioner komprehensif ini mengevaluasi kesiapan organisasi Anda untuk mengadopsi dan menerapkan beban kerja AI generatif dengan pertanyaan di seluruh area utama, termasuk kasus penggunaan, arsitektur, penyimpanan, kepatuhan, integrasi, pengujian, penerapan, dan strategi data. Dengan menangani aspek-aspek penting dari implementasi AI generatif, mulai dari infrastruktur teknis hingga pertimbangan peraturan, kuesioner ini membantu Anda mengidentifikasi kekuatan, kesenjangan, dan peluang dalam perjalanan modernisasi AI Anda.

Bagian:

- [Kesiapan](#)
- [Kasus penggunaan](#)
- [Arsitektur](#)
- [Penyimpanan](#)
- [Peraturan dan kepatuhan](#)
- [Integrasi](#)
- [Pengujian](#)
- [Penerapan dan otomatisasi](#)
- [Strategi data](#)

Anda juga dapat mengunduh kuesioner dalam format Microsoft Excel dan menggunakannya untuk merekam informasi Anda.

 [kuesioner](#)

[Unduh](#)

Kesiapan

Pertanyaan	Contoh respon
Apakah Anda memiliki AWS akun yang dapat dimanfaatkan untuk beban kerja ini?	Ya atau tidak.

Pertanyaan	Contoh respon
Apakah Anda memiliki perjanjian perusahaan yang ada dengan AWS?	Ya atau tidak.
Seberapa skalabel infrastruktur cloud Anda saat ini untuk menangani beban kerja AI generatif?	Infrastruktur cloud kami sangat terukur, dengan kemampuan penskalaan otomatis untuk sumber daya komputasi dan sistem penyimpanan terdistribusi yang dirancang untuk menangani beban kerja AI generatif skala besar secara efisien.
Apakah Anda memiliki kemampuan pipa data untuk preprocessing dan rekayasa fitur dalam skala besar?	Pipa data kami menggunakan kerangka kerja pemrosesan terdistribusi seperti Apache Spark untuk preprocessing data skala besar dan rekayasa fitur, dengan dukungan untuk pemrosesan data batch dan streaming.
Apakah Anda memiliki kemampuan penyediaan dan manajemen akun?	Ya atau tidak.
Bagaimana Anda menggambarkan literasi AI organisasi Anda dan kesiapan untuk mengadopsi teknologi AI generatif?	Organisasi kami telah banyak berinvestasi dalam program pendidikan AI, dan sebagian besar staf teknis telah menyelesaikan pelatihan AI/ML dasar. Organisasi ini memiliki budaya inovasi yang merangkul teknologi baru, termasuk AI generatif.
Keahlian AI/ML apa yang ada dalam organisasi Anda, dan bagaimana cara mendistribusikannya?	Kami memiliki AI Center of Excellence yang berdedikasi dengan ilmuwan data berpengalaman dan insinyur ML. Kami meningkatkan keterampilan pakar domain di berbagai unit bisnis untuk menjadi melek AI dan mengidentifikasi kasus penggunaan AI generatif.
Apakah Anda memiliki kasus bisnis tingkat tinggi yang mengartikulasikan tujuan, manfaat, dan biaya program cloud?	Ya atau tidak.

Pertanyaan	Contoh respon
Apa garis waktu Anda untuk mengambil solusi untuk produksi?	Minggu, bulan, dan sebagainya.
Apakah komitmen pendanaan telah dibuat oleh pemangku kepentingan utama Anda (misalnya, CFO, CIT/CTO, COO)?	Ya atau tidak.
Bagaimana Anda memastikan kepatuhan terhadap peraturan perlindungan data dalam inisiatif AI generatif Anda?	Kami memiliki tim kepatuhan khusus yang bekerja sama dengan tim AI kami. Kami melakukan penilaian dampak privasi secara teratur, menerapkan perlindungan data berdasarkan prinsip desain, dan memelihara catatan pemrosesan data terperinci untuk semua proyek AI generatif.
Seberapa matang sistem Anda yang ada yang terintegrasi dengan teknologi AI generatif baru?	Arsitektur TI kami didasarkan pada layanan mikro dan APIs memungkinkan integrasi fleksibel teknologi AI generatif baru. Sistem ini distandarisasi pada format dan protokol data umum untuk memastikan interoperabilitas.
Pengalaman apa yang Anda miliki dalam mengoperasikan model ML, dan bagaimana hal ini dapat diterapkan pada sistem AI generatif?	Kami telah menetapkan MLOps praktik, termasuk jalur pipa penerapan model otomatis, sistem pemantauan, dan kerangka kerja pengujian A/B. Praktik ini sedang disesuaikan untuk menangani persyaratan unik model AI generatif skala besar.

Kasus penggunaan

Pertanyaan	Contoh respon
Apa tujuan utama atau kriteria keberhasilan kasus penggunaan?	Untuk meningkatkan waktu respons dukungan pelanggan, meningkatkan konversi penjualan, meningkatkan rekomendasi produk. Juga:

Pertanyaan	Contoh respon
	Untuk meningkatkan kepuasan pengguna, tingkat penyelesaian tugas, kualitas respons, dan sebagainya.
Bagaimana kasus penggunaan ini selaras dengan tujuan strategis organisasi Anda?	Ini sejalan dengan tujuan strategis kami untuk meningkatkan kepuasan pelanggan dengan mengurangi waktu respons dalam layanan pelanggan.
Berapa volume data atau permintaan yang diharapkan untuk kasus penggunaan?	500 transaksi per detik (TPS).
Jenis sumber data apa yang diperlukan untuk mendukung beban kerja AI generatif Anda?	Database terstruktur internal (catatan pelanggan, data penjualan, dan sebagainya); data teks tidak terstruktur dari dokumen, email, dan media sosial; file audio dan video untuk tugas pengenalan ucapan dan gambar; data streaming waktu nyata dari perangkat dan sensor IoT; kumpulan data publik dan untuk pengayaan. APIs
Seberapa sering Anda perlu memperbarui atau menyegarkan data dari sumber-sumber ini?	Database transaksional: hampir pembaruan waktu nyata; repositori dokumen: pembaruan batch harian; umpan media sosial: pembaruan per jam; Data sensor IoT: streaming real-time berkelanjutan; kumpulan data publik: pembaruan bulanan atau triwulanan.
Format data apa yang dibutuhkan model AI generatif Anda sebagai input?	Data terstruktur: tabel database CSV, JSON, dan SQL; data teks: teks biasa, PDF, dan HTML; data gambar: JPEG, PNG, dan TIFF; data audio: WAV dan; data video: dan AVI MP3. MP4

Pertanyaan	Contoh respon
Apa masalah kualitas data utama Anda untuk beban kerja AI generatif?	Kelengkapan: memastikan bahwa tidak ada bidang kritis yang hilang; akurasi: memverifikasi kebenaran data dan menghilangkan kesalahan ; konsistensi: mempertahankan format dan nilai yang seragam di seluruh sumber; ketepatan waktu: memastikan bahwa data mutakhir untuk inferensi waktu nyata; relevansi: mengonfirmasi bahwa data selaras dengan tugas AI generatif tertentu.
Apa persyaratan kinerja utama (misalnya, waktu respons, throughput, akurasi)?	Akurasi 95%; < 500 ms waktu respons; kemampuan untuk menangani 1000 permintaan/detik. Akurasi tinggi (95% +), akurasi sedang (80-90%), upaya terbaik, dan sebagainya.
Apakah Anda memiliki yang lain KPIs untuk mengukur keberhasilan kasus penggunaan ini?	Kunci KPIs termasuk pengurangan tingkat kesalahan, penghematan waktu per transaksi, dan skor kepuasan pelanggan.
Berapa akurasi model yang diinginkan, dan bagaimana menyeimbangkannya dengan biaya?	Akurasi tinggi (> 90%) dengan biaya sedang, akurasi sedang (70-80%) dengan biaya rendah, dan sebagainya.
Apa kasus penggunaan utama atau skenario untuk solusi AI generatif?	Chatbot layanan pelanggan, pembuatan konten, rekomendasi produk, dan sebagainya.
Apa target pengguna atau persona untuk sistem AI generatif?	Agen layanan pelanggan, tim pemasaran, karyawan, pengguna akhir, dan sebagainya.
Berapa volume permintaan atau pengguna yang diharapkan?	1.000 permintaan per hari; 10.000 pengguna aktif bulanan.
Apakah ada batasan atau persyaratan kasus penggunaan khusus?	Respons real-time, dukungan multi-bahasa, privasi data, dan sebagainya.

Pertanyaan	Contoh respon
Apakah Anda memiliki anggaran yang dialokasikan untuk mengembangkan dan memelihara solusi AI generatif?	Biaya pengembangan awal diperkirakan \$200.000, dengan biaya pemeliharaan tahunan \$50.000.
Berapa proyeksi pengembalian investasi (ROI) dan periode pengembalian modal untuk kasus penggunaan ini?	ROI yang diharapkan sebesar 150% selama tiga tahun, dengan periode pengembalian modal 18 bulan.
Apakah ada biaya tersembunyi atau potensi penghematan yang harus dipertimbangkan?	Penghematan potensial termasuk pengurangan biaya lembur. Biaya tersembunyi mungkin melibatkan pelatihan tambahan untuk staf.
Apa skalabilitas dan kemungkinan ekspansi masa depan dari solusi AI generatif ini?	Solusi ini dirancang untuk skala dengan operasi kami, dengan kemungkinan memperluas ke departemen lain di masa depan.
Bagaimana Anda memastikan keadilan dan mengurangi bias dalam model AI generatif Anda?	Kami berencana untuk mengurangi bias melalui pengumpulan data yang beragam, audit bias reguler, dan implementasi teknik mitigasi bias.
Proses apa yang Anda miliki untuk mengatasi masalah etika atau konsekuensi yang tidak diinginkan?	Kami akan mengelola masalah etika melalui rencana respons insiden AI yang mapan, penilaian risiko etis reguler, sistem pelaporan anonim untuk karyawan, kolaborasi dengan pakar etika eksternal, dan pemantauan dan penyesuaian berkelanjutan dari model yang diterapkan berdasarkan umpan balik.

Pertanyaan	Contoh respon
Bagaimana Anda mendekati memprioritaskan dan mengurutkan penilaian beban kerja AI generatif di berbagai proyek dan departemen di organisasi Anda?	Dengan melakukan survei tingkat tinggi di semua departemen untuk mengidentifikasi potensi kasus penggunaan AI generatif dan mengevaluasinya berdasarkan tiga kriteria utama: dampak bisnis, kelayakan teknis, dan pertimbangan etis. Proyek dengan dampak potensial tinggi, hambatan teknis yang lebih rendah, dan masalah etika minimal diberikan prioritas.

Arsitektur

Pertanyaan	Contoh respon
Jenis model atau arsitektur AI generatif apa yang sedang dipertimbangkan?	Transformator, jaringan saraf konvolusional (CNN), jaringan saraf berulang (RNN), pohon keputusan, dan sebagainya.
Berapa skala atau volume data dan perhitungan yang diharapkan?	Jutaan pengguna, petabyte data, dan sebagainya.
Apa persyaratan perangkat keras (misalnya, CPUs atau GPUs) untuk pelatihan dan inferensi?	High-end GPUs, cluster CPU, instance cloud, dan sebagainya.
Bagaimana model AI generatif akan diperbarui atau dilatih ulang dari waktu ke waktu?	Melalui pembelajaran berkelanjutan, pelatihan ulang berkala, pembaruan manual, dan sebagainya.
Apa saja preprocessing data dan persyaratan rekayasa fitur?	Pembersihan teks, augmentasi gambar, pemilihan fitur, dan sebagainya.
Bagaimana sistem AI generatif menangani kasus tepi, outlier, atau input dengan kepercayaan rendah?	Melalui mundur ke pengawasan manusia, meminta klarifikasi, dan sebagainya.

Pertanyaan	Contoh respon
Apa persyaratan latensi untuk aplikasi AI generatif?	Real-time, mendekati waktu nyata, pemrosesan batch, dan sebagainya.

Penyimpanan

Pertanyaan	Contoh respon
Di mana data pelatihan akan disimpan?	Di penyimpanan cloud (misalnya, Amazon S3, penyimpanan file, penyimpanan blok, atau penyimpanan objek), di penyimpanan lokal, dan sebagainya.
Apa persyaratan penyimpanan untuk data pelatihan dan artefak model (misalnya, kapasitas, daya tahan, ketersediaan)?	Penyimpanan skala petabyte, daya tahan tinggi (daya tahan 99,999999999%), ketersediaan tinggi, dan sebagainya.
Apa persyaratan retensi dan cadangan data untuk data pelatihan dan artefak model?	Retensi data selama x tahun, backup harian, backup off-site, dan sebagainya.
Format file mana yang terutama digunakan untuk menyimpan kumpulan data pelatihan AI Anda (misalnya, CSV, JSON, Parquet,)? HDF5	File parquet untuk data terstruktur, dan HDF5 untuk array multidimensi besar dan data tidak terstruktur seperti gambar dan teks. Kami menggunakan format khusus seperti TFRecord untuk mengoptimalkan pemuatan data selama pelatihan.
Bagaimana kumpulan data pelatihan Anda diatur: sebagai file individual, dalam database, atau menggunakan format data AI khusus?	Kumpulan data kecil hingga menengah disimpan sebagai file Parquet individual dalam penyimpanan objek untuk fleksibilitas. Dataset besar disimpan dalam database terdistribusi (Cassandra) untuk menangani skala.
Apakah Anda menggunakan teknik kompresi atau pengkodean data khusus untuk data pelatihan AI generatif?	Untuk data tabular, kami menggunakan teknik pengkodean kamus dan bit-packing yang tersedia di Parquet. Untuk gambar, kami

Pertanyaan	Contoh respon
	menggunakan kompresi JPEG lossy dengan pengaturan kualitas yang dioptimalkan untuk model kami.
Bagaimana Anda menangani pembuatan versi dan penyimpanan berbagai iterasi kumpulan data pelatihan? Apa dampaknya terhadap kebutuhan penyimpanan Anda secara keseluruhan?	Kami menggunakan sistem versi data (DVC) yang terintegrasi dengan platform ML kami.

Peraturan dan kepatuhan

Pertanyaan	Contoh respon
Apa peraturan atau persyaratan kepatuhan yang relevan untuk solusi AI generatif (misalnya, GDPR, HIPAA, PCI-DSS)?	GDPR untuk menangani data pribadi, HIPAA untuk data perawatan kesehatan, PCI-DSS untuk data pembayaran, dan sebagainya.
Pedoman atau kerangka kerja AI generatif etis apa yang telah diadopsi organisasi Anda?	Kami menerapkan pedoman AI bertanggung jawab kami sendiri. Semua proyek AI generatif menjalani tinjauan etis sebelum persetujuan dan penerapan.
Apa persyaratan keamanan untuk sistem AI generatif?	Enkripsi data, komunikasi jaringan yang aman, audit keamanan reguler.
Apa persyaratan untuk privasi dan perlindungan data?	Anonimisasi data, enkripsi, kontrol akses, dan sebagainya.
Apa persyaratan solusi untuk menangani data sensitif atau rahasia?	Kontrol akses yang ketat, penyembunyian data, persyaratan residensi data, dan sebagainya.
Bagaimana otentikasi dan otorisasi pengguna ditangani?	Dengan menggunakan OAuth, kunci API, sistem masuk tunggal (SSO), dan kontrol akses berbasis peran (RBAC).

Pertanyaan	Contoh respon
Bagaimana solusi akan dipantau dan dikelola dalam produksi?	Dengan menggunakan alat pemantauan seperti Prometheus dan Datadog, alat logging seperti ELK Stack, sistem peringatan, dan sebagainya.

Integrasi

Pertanyaan	Contoh respon
Apa persyaratan untuk mengintegrasikan solusi AI generatif dengan sistem atau sumber data yang ada?	REST APIs, antrian pesan, konektor database, dan sebagainya.
Bagaimana data akan dicerna dan diproses sebelumnya untuk solusi AI generatif?	Dengan menggunakan pemrosesan batch, streaming data, transformasi data, dan rekayasa fitur.
Bagaimana output dari solusi AI generatif akan dikonsumsi atau diintegrasikan dengan sistem hilir?	Melalui endpoint API, antrian pesan, pembaruan database, dan sebagainya.
Pola integrasi berbasis peristiwa mana yang dapat digunakan untuk solusi AI generatif?	Antrian pesan (seperti Amazon SQS, Apache Kafka, RabbitMQ), pub/sub sistem, webhook, platform streaming acara.
Pendekatan integrasi berbasis API mana yang dapat digunakan untuk menghubungkan solusi AI generatif dengan sistem lain?	RESTful APIs, APIs GraphQL, APIs SOAP (untuk sistem warisan).
Komponen arsitektur layanan mikro mana yang dapat digunakan untuk integrasi solusi AI generatif?	Service mesh untuk komunikasi antar layanan, gateway API, orkestrasi kontainer (misalnya, Kubernetes).
Bagaimana integrasi hybrid dapat diterapkan untuk solusi AI generatif?	Dengan menggabungkan pola berbasis peristiwa untuk pembaruan waktu nyata,

Pertanyaan	Contoh respon
	pemrosesan batch untuk data historis, dan APIs untuk integrasi sistem eksternal.
Bagaimana output solusi AI generatif dapat diintegrasikan dengan sistem hilir?	Melalui titik akhir API, antrian pesan, pembaruan basis data, webhook, dan ekspor file.
Langkah-langkah keamanan mana yang harus dipertimbangkan untuk mengintegrasikan solusi AI generatif?	Mekanisme otentikasi (seperti OAuth atau JWT), enkripsi (dalam perjalanan dan saat istirahat), pembatasan laju API, dan daftar kontrol akses (). ACLs
Bagaimana Anda berencana untuk mengintegrasikan kerangka kerja open source seperti LlamaIndex atau LangChain ke dalam pipeline data yang ada dan alur kerja AI generatif?	Kami berencana menggunakannya LangChain untuk membangun aplikasi AI generatif yang kompleks, terutama untuk kemampuan agen dan manajemen memorinya. Kami bertujuan untuk memiliki 60% proyek AI generatif kami yang digunakan LangChain dalam 6 bulan ke depan.
Bagaimana Anda memastikan kompatibilitas antara kerangka kerja open source yang Anda pilih dan infrastruktur data yang ada?	Kami menciptakan tim integrasi khusus untuk memastikan kompatibilitas yang lancar. Pada kuartal ketiga, tujuan kami adalah memiliki pipa terintegrasi penuh yang digunakan LlamaIndex untuk pengindeksan dan pengambilan data yang efisien dalam struktur data lake kami saat ini.
Bagaimana Anda berencana untuk memanfaatkan komponen modular kerangka kerja seperti LangChain untuk pembuatan prototipe dan eksperimen cepat?	Kami sedang menyiapkan lingkungan kotak pasir di mana pengembang dapat dengan cepat membuat prototipe dengan menggunakan LangChain komponen.

Pertanyaan	Contoh respon
Apa strategi Anda untuk mengikuti pembaruan dan fitur baru dalam kerangka kerja open source yang berkembang pesat ini?	Kami telah menugaskan tim untuk memantau GitHub repositori dan forum komunitas untuk LangChain dan. LlamaIndex Kami berencana untuk mengevaluasi dan mengintegrasikan pembaruan besar setiap triwulan, dengan fokus pada peningkatan kinerja dan kemampuan baru.

Pengujian

Pertanyaan	Contoh respon
Apa persyaratan pengujian (misalnya, pengujian unit, pengujian integrasi, end-to-end pengujian)?	Unit pengujian untuk komponen individu, pengujian integrasi dengan sistem eksternal, end-to-end pengujian untuk skenario kritis, dan sebagainya.
Bagaimana Anda memastikan kualitas dan konsistensi data di berbagai sumber untuk pelatihan AI generatif?	Kami menjaga kualitas data melalui alat pembuatan profil data otomatis, audit data reguler, dan katalog data terpusat. Kami telah menerapkan kebijakan tata kelola data untuk memastikan konsistensi di seluruh sumber dan untuk mempertahankan garis keturunan data.
Bagaimana model AI generatif akan dievaluasi dan divalidasi?	Dengan menggunakan dataset holdout, evaluasi manusia, pengujian A/B, dan sebagainya.
Apa kriteria untuk mengevaluasi kinerja dan akurasi model AI generatif?	Presisi, ingatan, skor F1, kebingungan, evaluasi manusia, dan sebagainya.
Bagaimana kasus tepi dan kasus sudut diidentifikasi dan ditangani?	Dengan menggunakan rangkaian pengujian yang komprehensif, evaluasi manusia, pengujian permusuhan, dan sebagainya.

Pertanyaan	Contoh respon
Bagaimana Anda akan menguji potensi bias dalam model AI generatif?	Dengan menggunakan analisis paritas demografis, pengujian kesempatan yang sama, teknik de-biasing permusuhan, pengujian kontrafaktual, dan sebagainya.
Metrik mana yang akan digunakan untuk mengukur keadilan dalam output model?	Rasio dampak yang berbeda, peluang yang disamakan, paritas demografis, metrik keadilan individu, dan sebagainya.
Bagaimana Anda memastikan representasi yang beragam dalam kumpulan data pengujian Anda untuk deteksi bias?	Dengan menggunakan stratified sampling lintas kelompok demografis, kolaborasi dengan pakar keragaman, penggunaan data sintetis untuk mengisi kesenjangan, dan sebagainya.
Proses mana yang akan diterapkan untuk pemantauan berkelanjutan dari keadilan model pasca-penerapan?	Audit keadilan reguler, sistem deteksi bias otomatis, analisis umpan balik pengguna, pelatihan ulang berkala dengan kumpulan data yang diperbarui, dan sebagainya.
Bagaimana Anda akan mengatasi bias interseksional dalam model AI generatif?	Dengan menggunakan analisis keadilan interseksional, pengujian subkelompok, kolaborasi dengan pakar domain tentang interseksionalitas, dan sebagainya.
Bagaimana Anda akan menguji kinerja model di berbagai bahasa dan konteks budaya?	Dengan menggunakan set tes multibahasa, kolaborasi dengan pakar budaya, metrik keadilan lokal, studi perbandingan lintas budaya, dan sebagainya.

Penerapan dan otomatisasi

Pertanyaan	Contoh respon
Apa persyaratan untuk penskalaan dan penyeimbangan beban?	Perutean permintaan cerdas; sistem penskalaan otomatis; mengoptimalkan awal dingin yang

Pertanyaan	Contoh respon
	cepat dengan menggunakan teknik seperti caching model, pemuatan lambat, dan sistem penyimpanan terdistribusi; merancang sistem untuk menangani pola lalu lintas yang meledak dan tidak dapat diprediksi.
Apa persyaratan untuk memperbarui dan meluncurkan versi baru?	Penerapan biru/hijau, rilis kenari, pembaruan bergulir, dan sebagainya.
Apa persyaratan untuk pemulihan bencana dan kelangsungan bisnis?	Prosedur pencadangan dan pemulihan, mekanisme failover, konfigurasi ketersediaan tinggi, dan sebagainya.
Apa persyaratan untuk mengotomatiskan pelatihan, penyebaran, dan manajemen model AI generatif?	Pipa pelatihan otomatis, penerapan berkelanjutan, penskalaan otomatis, dan sebagainya.
Bagaimana model AI generatif akan diperbarui dan dilatih ulang saat data baru tersedia?	Melalui pelatihan ulang berkala, pembelajaran inkremental, pembelajaran transfer, dan sebagainya.
Apa persyaratan untuk mengotomatisasi pemantauan dan manajemen?	Peringatan otomatis, penskalaan otomatis, penyembuhan diri, dan sebagainya.
Apa lingkungan penerapan pilihan Anda untuk beban kerja AI generatif?	Pendekatan hybrid yang menggunakan AWS untuk pelatihan model dan infrastruktur lokal kami untuk inferensi guna memenuhi persyaratan residensi data.
Apakah ada platform cloud khusus yang Anda sukai untuk penerapan AI generatif?	Layanan AWS, khususnya Amazon SageMaker AI untuk pengembangan dan penerapan model, dan Amazon Bedrock untuk model pondasi.
Teknologi kontainerisasi apa yang Anda pertimbangkan untuk beban kerja AI generatif?	Kami ingin melakukan standarisasi pada kontainer Docker yang diatur dengan Kubernetes untuk memastikan portabilitas dan skalabilitas di seluruh lingkungan hybrid kami.

Pertanyaan	Contoh respon
Apakah Anda memiliki alat pilihan untuk CI/CD di saluran AI generatif Anda?	GitLab untuk kontrol versi dan pipa CI/CD, terintegrasi dengan Jenkins untuk pengujian dan penerapan otomatis.
Alat orkestrasi apa yang Anda pertimbangkan untuk mengelola alur kerja AI generatif?	Apache Airflow untuk orkestrasi alur kerja, terutama untuk pra-pemrosesan data dan jalur pelatihan model.
Apakah Anda memiliki persyaratan khusus untuk infrastruktur lokal untuk mendukung beban kerja AI generatif?	Kami berinvestasi di server berakselerasi GPU dan jaringan berkecepatan tinggi untuk mendukung beban kerja inferensi lokal.
Bagaimana Anda berencana untuk mengelola versi model dan penerapan di lingkungan yang berbeda?	Kami berencana untuk menggunakannya MLflow untuk pelacakan model dan pembuatan versi, dan mengintegrasikannya dengan infrastruktur Kubernetes kami untuk penerapan yang mulus di seluruh lingkungan.
Alat pemantauan dan observabilitas apa yang Anda pertimbangkan untuk penerapan AI generatif?	Prometheus untuk pengumpulan metrik dan Grafana untuk visualisasi, dengan solusi pencatatan khusus tambahan untuk pemantauan khusus model.
Bagaimana Anda menangani pergerakan dan sinkronisasi data dalam model penerapan hibrida?	Kami akan menggunakannya AWS DataSync untuk transfer data yang efisien antara penyimpanan lokal dan AWS, dengan pekerjaan sinkronisasi otomatis yang dijadwalkan berdasarkan siklus pelatihan kami.
Langkah-langkah keamanan apa yang Anda terapkan untuk penerapan AI generatif di berbagai lingkungan?	Kami akan menggunakan IAM untuk sumber daya cloud, terintegrasi dengan Active Directory lokal kami untuk mengimplementasikan end-to-end enkripsi dan segmentasi jaringan untuk mengamankan aliran data.

Strategi data

Pertanyaan	Contoh respon
Jenis data spesifik apa yang penting untuk beban kerja AI generatif Anda, dan berapa persentasenya yang saat ini dapat diakses?	Log panggilan pelanggan dan data ulasan produk sangat penting. Saat ini, 85% dari tipe data ini dapat diakses untuk proyek AI generatif kami.
Bagaimana Anda memastikan dan mengukur kualitas data Anda?	Kami telah menerapkan metrik kualitas data, termasuk kelengkapan, akurasi, konsistensi, dan ketepatan waktu. Kami menggunakan alat otomatis untuk menilai metrik ini secara teratur dan memiliki tim khusus untuk pembersihan dan pengayaan data.
Berapa persentase data Anda yang memenuhi standar kualitas Anda untuk penggunaan AI generatif?	Saat ini, 78% data kami memenuhi standar kualitas kami. Kami menargetkan 95% dalam 12 bulan ke depan melalui proses pembersihan data yang lebih baik.
Bagaimana Anda berencana untuk membangun kepercayaan tentang penggunaan data dalam AI generatif di antara para pemangku kepentingan Anda?	Kami menerapkan dewan etika AI, memberikan penjelasan yang jelas tentang keputusan AI, dan melakukan audit AI triwulanan untuk memastikan transparansi dan keadilan.
Seberapa komprehensif dokumentasi Anda untuk sumber data dan garis keturunan?	Kami memelihara katalog data terperinci yang mencakup metadata untuk semua sumber data kami, termasuk asal, frekuensi pembaruan, dan penggunaan. Kami menggunakan alat silsilah data untuk melacak bagaimana data mengalir dan berubah di seluruh sistem kami.
Bagaimana Anda memastikan keragaman dalam kumpulan data Anda untuk mencegah bias dalam model AI?	Kami secara aktif mengambil data dari beragam demografi dan secara teratur mengaudit kumpulan data kami untuk bias representasional. Kami juga menggunakan teknik pembuatan data sintetis untuk

Pertanyaan	Contoh respon
	menyeimbangkan kategori yang kurang terwakili.
Berapa kecepatan refresh data Anda untuk model AI generatif kritis, dan bagaimana Anda menentukan frekuensi ini?	Model kritis disegarkan setiap minggu. Frekuensi ini ditentukan oleh metrik kinerja pengujian A/B, dan kami bertujuan untuk tidak lebih dari 2% degradasi antara penyegaran.
Berapa banyak versi kumpulan data penting yang Anda pertahankan dan untuk berapa lama?	Kami mempertahankan lima versi terakhir dari setiap kumpulan data penting, dengan periode retensi 18 bulan untuk setiap versi.
Berapa banyak tim lintas fungsi yang terlibat dalam inisiatif AI generatif Anda dan memiliki akses ke data Anda?	Kami memiliki tiga tim lintas fungsi. Setiap tim mencakup ilmuwan data, pakar domain, ahli etika, dan analis bisnis.
Kebijakan dan praktik tata kelola data apa yang Anda miliki?	Kami memiliki komite tata kelola data lintas fungsi yang mengawasi kebijakan data kami. Kami telah menerapkan kontrol akses berbasis peran, skema klasifikasi data, dan audit reguler untuk memastikan kepatuhan terhadap kerangka tata kelola kami.
Tindakan apa yang Anda miliki untuk memastikan privasi data, mendapatkan persetujuan yang tepat, dan menjaga kerahasiaan?	Kami telah menerapkan kerangka kerja privasi data komprehensif yang selaras dengan GDPR dan CCPA. Ini termasuk mendapatkan persetujuan eksplisit untuk penggunaan data, menerapkan teknik anonimisasi data, dan penilaian dampak privasi reguler.
Berapa persentase kumpulan data pelatihan AI Anda yang telah diaudit untuk bias pada kuartal terakhir?	70% dari kumpulan data pelatihan AI kami diaudit untuk bias kuartal terakhir. Kami menerapkan alat deteksi bias otomatis untuk mencapai audit triwulanan 100%.

Pertanyaan	Contoh respon
Berapa kapasitas pemrosesan data Anda saat ini, dan berapa banyak proyek yang Anda butuhkan untuk beban kerja AI generatif masa depan?	Kapasitas kami saat ini adalah 10 TB/day. We project needing 30 TB/day dalam setahun dan meningkatkan infrastruktur kami untuk memenuhi permintaan ini.
Apa strategi Anda untuk menyeimbangkan privasi data dengan kebutuhan data model AI generatif?	Kami menerapkan teknik anonimisasi tingkat lanjut dan pembuatan data sintetis. Tujuan kami adalah untuk meningkatkan data yang dapat digunakan untuk AI sebesar 40% sekaligus mengurangi risiko privasi hingga 60% selama tahun depan.
Berapa persentase kumpulan data pembelajaran mesin (ML) Anda yang diberi label secara akurat, dan berapa tingkat akurasi target Anda?	Saat ini, 85% dari kumpulan data ML kami diberi label secara akurat. Kami menargetkan tingkat akurasi 95% dalam kuartal berikutnya dengan menggunakan teknik pelabelan manusia dan otomatis.

Menerjemahkan wawasan penilaian menjadi hasil yang dapat ditindaklanjuti

Bagian ini menyediakan kerangka kerja untuk menganalisis tanggapan kuesioner dan menggunakan wawasan tersebut untuk membentuk arsitektur target dan kiriman utama lainnya dari inisiatif modernisasi AI generatif. Kerangka kerja ini menjembatani kesenjangan antara pengumpulan dan implementasi data, dan memastikan bahwa penilaian secara langsung menginformasikan dan mendorong strategi modernisasi Anda.

Definisi arsitektur target:

- Gunakan tanggapan kuesioner untuk menginformasikan pemilihan layanan cloud dan desain pipa data.
- Pastikan bahwa desain arsitektur mendukung skalabilitas dan interoperabilitas seperti yang disorot dalam panduan.

Evaluasi kesiapan pelanggan:

- Menganalisis tanggapan kuesioner yang terkait dengan infrastruktur, proses, dan budaya organisasi saat ini.
- Identifikasi kesenjangan dan buat rencana untuk mengatasinya. Prioritaskan kesenjangan yang penting untuk keberhasilan MVP.

Gunakan tujuan kasing dan peregangan:

- Ekstrak masalah bisnis tertentu dari tanggapan kuesioner untuk menentukan tujuan kasus penggunaan yang jelas.
- Tetapkan tujuan yang selaras dengan visi jangka panjang organisasi Anda untuk modernisasi AI generatif.

Estimasi upaya:

- Gunakan data kuesioner untuk memperkirakan sumber daya, waktu, dan anggaran untuk MVP dan implementasi penuh.
- Buat pendekatan bertahap yang dimulai dengan MVP, dan uraikan fase berikutnya.

Kebutuhan pemberdayaan:

- Berdasarkan tanggapan kuesioner, identifikasi kesenjangan keterampilan dan kebutuhan pelatihan.
- Kembangkan rencana pelatihan yang mendukung kebutuhan MVP langsung dan adopsi AI generatif jangka panjang.

Rencana implementasi:

- Buat peta jalan komprehensif yang dimulai dengan MVP dan uraikan langkah-langkah menuju modernisasi AI generatif penuh.
- Tentukan tonggak dan kiriman yang jelas untuk setiap fase implementasi.

Langkah-langkah praktis:

- Matriks prioritas: Buat matriks yang memetakan tanggapan kuesioner terhadap [enam hasil](#) untuk membantu memprioritaskan fitur dan upaya.
- Pendekatan iteratif: Rancang MVP untuk menjadi iterasi pertama dalam serangkaian rilis yang direncanakan, di mana setiap rilis dibangun menuju arsitektur target penuh.
- Penyelarasan pemangku kepentingan: Gunakan hasil kuesioner untuk menyelaraskan pemangku kepentingan pada lingkup MVP dan pendekatan bertahap untuk mencapai semua hasil.
- Loop umpan balik berkelanjutan: Menerapkan mekanisme untuk mengumpulkan umpan balik setelah penerapan MVP, dan menggunakan wawasan untuk menyempurnakan rencana untuk fase berikutnya.
- Implementasi tangkas: Mengadopsi metodologi tangkas yang memungkinkan fleksibilitas dalam menangani semua hasil dari waktu ke waktu, dimulai dengan hasil paling kritis dalam MVP.

Langkah selanjutnya

Setelah Anda menyelesaikan penilaian beban kerja AI generatif, ikuti langkah-langkah berikut:

1. Memberikan arsitektur target yang terperinci

- Tujuan: Arsitek solusi menciptakan arsitektur target komprehensif yang selaras dengan tujuan organisasi dan hasil penilaian.
- Komponen: Arsitektur ini mencakup desain konsumsi data, titik integrasi, dan interoperabilitas sistem untuk memastikan skalabilitas, keandalan, dan optimalisasi kinerja.

2. Jelaskan seberapa spesifik Layanan AWS sesuai dengan kasus penggunaan

- Pemetaan layanan: Identifikasi dan petakan spesifik Layanan AWS yang paling sesuai dengan kasus penggunaan yang diidentifikasi.
- Manfaat: Sorot bagaimana layanan ini memenuhi kebutuhan bisnis tertentu, meningkatkan efisiensi, dan memberikan skalabilitas.

3. Berikan solusi alternatif opsional dengan pro dan kontra

- Alternatif: Menyajikan solusi alternatif yang juga dapat memenuhi persyaratan organisasi.
- Analisis: Menawarkan analisis terperinci tentang keuntungan dan kerugian dari setiap alternatif dengan mempertimbangkan faktor-faktor seperti biaya, kompleksitas, dan keselarasan dengan tujuan bisnis.

4. Memberikan estimasi harga rinci Layanan AWS

- Analisis biaya: Memberikan estimasi biaya terperinci untuk yang diusulkan Layanan AWS, termasuk skenario penggunaan potensial dan model penetapan harga.
- Penyelarasan anggaran: Pastikan bahwa biaya sejalan dengan kendala anggaran organisasi dan memberikan pemahaman yang jelas tentang implikasi keuangan.

5. Dapatkan umpan balik tentang arsitektur yang diusulkan

- Keterlibatan pemangku kepentingan: Terlibat dengan pemangku kepentingan untuk mempresentasikan arsitektur yang diusulkan dan mengumpulkan umpan balik.
- Peningkatan berulang: Gunakan umpan balik untuk menyempurnakan dan meningkatkan solusi, dan konfirmasi bahwa itu memenuhi kebutuhan dan harapan semua pemangku kepentingan.

Pertanyaan yang Sering Diajukan

Apa tujuan utama penilaian beban kerja AI generatif?

Tujuan utama dari penilaian ini adalah untuk mengevaluasi kesiapan organisasi untuk memodernisasi beban kerja AI generatif mereka, mengidentifikasi kasus penggunaan, dan mengembangkan arsitektur solusi target. Ini bertujuan untuk menentukan persyaratan modernisasi, menentukan ruang lingkup implementasi, dan mempersiapkan modernisasi AI generatif yang sukses.

Siapa yang harus menggunakan penilaian ini?

Penilaian ini untuk arsitek solusi, arsitek perusahaan, dan arsitek aplikasi yang ingin menilai aspek teknis modernisasi AI generatif. Hal ini juga berguna bagi manajer program dan manajer orang untuk mengukur kesiapan keseluruhan, alokasi sumber daya, dan kebutuhan pemberdayaan.

Apa komponen kunci yang dievaluasi dalam penilaian?

Penilaian tersebut mencakup kesiapan keseluruhan, kasus penggunaan, arsitektur, penyimpanan, peraturan dan kepatuhan, integrasi, pengujian, otomatisasi penyebaran, dan strategi data. Komponen-komponen ini sangat penting untuk menentukan kesiapan teknis dan organisasi untuk adopsi modernisasi AI generatif.

Bagaimana penilaian membantu menentukan arsitektur target?

Penilaian memberikan pendekatan terstruktur untuk mengevaluasi sistem saat ini dan mengidentifikasi perbaikan. Ini membantu Anda memilih teknologi yang tepat dan merancang arsitektur terukur yang selaras dengan tujuan bisnis dan persyaratan kasus penggunaan.

Apa manfaat melakukan penilaian beban kerja AI generatif?

Manfaat termasuk peningkatan efisiensi, peningkatan pengambilan keputusan, jaminan kepatuhan, pembinaan inovasi, dan persiapan skalabilitas. Penilaian tersebut menetapkan pendekatan strategis untuk modernisasi AI generatif, dan memaksimalkan potensi manfaat sambil mengurangi risiko.

Bagaimana organisasi dapat memastikan keberhasilan implementasi setelah penilaian?

Organizations harus mengembangkan rencana implementasi yang jelas yang mencakup tonggak yang ditentukan, melibatkan pemangku kepentingan lebih awal, dan mengadopsi pendekatan berulang. Mendirikan Center of Excellence (CoE) dan berfokus pada pengembangan bakat juga direkomendasikan praktik terbaik.

Tantangan apa yang mungkin dihadapi organisasi selama penilaian?

Tantangan mungkin termasuk resistensi terhadap perubahan, masalah kualitas data, dan kompleksitas kepatuhan. Mengatasi tantangan ini membutuhkan pengembangan budaya inovasi, memastikan kesiapan data, dan menerapkan langkah-langkah keamanan yang kuat.

Bagaimana penilaian menangani persyaratan peraturan dan kepatuhan?

Penilaian mengevaluasi langkah-langkah kepatuhan saat ini dan mengidentifikasi kesenjangan. Ini memastikan bahwa solusi target mematuhi peraturan yang relevan dan undang-undang privasi data, dan menggabungkan praktik terbaik keamanan untuk melindungi informasi sensitif.

Peran apa yang dimainkan oleh keterlibatan pemangku kepentingan dalam proses penilaian?

Keterlibatan pemangku kepentingan sangat penting untuk mendapatkan dukungan, menyelaraskan inisiatif modernisasi dengan tujuan bisnis, dan memastikan implementasi yang berhasil. Keterlibatan awal dan komunikasi manfaat yang jelas adalah kunci untuk mengatasi resistensi dan mendorong dukungan.

Bagaimana organisasi dapat mengukur keberhasilan inisiatif modernisasi AI generatif mereka setelah penilaian?

Sukses dapat diukur dengan menggunakan indikator kinerja utama (KPIs) yang selaras dengan tujuan bisnis. Pemantauan dan evaluasi metrik ini secara teratur membantu memandu pengambilan keputusan dan menunjukkan nilai modernisasi AI generatif kepada para pemangku kepentingan.

Bagaimana pendekatan penilaian berbeda untuk organisasi dengan berbagai ukuran (kecil, menengah, atau perusahaan) atau industri?

Organisasi kecil:

- Mungkin memiliki sumber daya dan keahlian yang terbatas untuk penilaian komprehensif
- Kemungkinan untuk fokus pada kasus penggunaan berdampak tinggi tertentu daripada adopsi di seluruh perusahaan
- Mungkin lebih bergantung pada alat dan layanan pihak ketiga untuk penilaian
- Proses penilaian mungkin kurang formal dan lebih gesit

Organisasi menengah:

- Seringkali memiliki tim TI atau data khusus tetapi mungkin tidak memiliki keahlian AI khusus
- Mungkin mengambil pendekatan bertahap, dimulai dengan proyek percontohan di departemen utama
- Perlu menyeimbangkan inovasi dengan sistem dan proses yang ada
- Penilaian kemungkinan melibatkan tim lintas fungsi

Organisasi perusahaan:

- Biasanya memiliki AI/ML tim yang berdedikasi dan lebih banyak sumber daya untuk penilaian komprehensif
- Perlu mempertimbangkan integrasi kompleks dengan sistem perusahaan yang ada
- Mungkin memiliki persyaratan peraturan khusus industri untuk dipertimbangkan
- Penilaian sering melibatkan proses tata kelola formal

Sumber daya

- [AI generatif aktif AWS](#)
- [AWS menawarkan kecerdasan buatan baru, pembelajaran mesin, dan panduan AI generatif untuk merencanakan strategi AI Anda](#) (posting AWS blog)
- [Praktik terbaik untuk membangun aplikasi AI generatif di AWS](#) (posting AWS blog)
- [Pembuat Aplikasi AI Generatif di AWS](#) (Perpustakaan AWS Solusi)
- [Kemampuan AI generatif](#) (Arsitektur Referensi AWS Keamanan)
- [AWS kerangka kerja praktik terbaik AI generatif](#) (Panduan AWS Audit Manager Pengguna)
- [Memilih layanan AI generatif](#) (panduan AWS keputusan)
- [Apa itu Amazon Bedrock?](#) (Panduan Pengguna Amazon Bedrock)
- [Apa itu Amazon SageMaker AI?](#) (Panduan Pengembang Amazon SageMaker AI)

Riwayat dokumen

Tabel berikut menjelaskan perubahan signifikan pada panduan ini. Jika Anda ingin diberi tahu tentang pembaruan masa depan, Anda dapat berlangganan umpan [RSS](#).

Perubahan	Deskripsi	Tanggal
Publikasi awal	—	6 November 2024

Terjemahan disediakan oleh mesin penerjemah. Jika konten terjemahan yang diberikan bertentangan dengan versi bahasa Inggris aslinya, utamakan versi bahasa Inggris.