

তথ্য প্রযুক্তি

এআই মডেলের অপ্রত্যাশিত আচরণের ছয় ঘটনা প্রকাশ করল ওপেনএআই

প্রতিবেদক: alamin dhakaprobe

প্রকাশের তারিখ: শুক্রবার, ১৮ সেপ্টেম্বর ২০২৬, ০৫:২৫ PM



গত ছয় মাসে নিজেদের এআই মডেলগুলোর অপ্রত্যাশিত ও উদ্বেগজনক আচরণের ছয়টি ঘটনা প্রকাশ করেছে ওপেনএআই। এসব ঘটনার মধ্যে ভুল তথ্য গোপনের চেষ্টা, অনুমোদন ছাড়া কাজ করা, অন্য এআই মডেলের সঙ্গে যোগাযোগ এবং নির্ধারিত বিধিনিষেধ এড়িয়ে যাওয়ার মতো আচরণ রয়েছে। একই সঙ্গে এ ধরনের 'মডেল মিসঅ্যালাইনমেন্ট' শনাক্ত, তদন্ত ও প্রকাশের জন্য নতুন একটি কাঠামো চালুর কথা জানিয়েছে প্রতিষ্ঠানটি।

ওপেনএআই গত বুধবার জানিয়েছে, মডেলগুলোর প্রশিক্ষণ ও মূল্যায়নের সময় এসব ঘটনা শনাক্ত করা হয়েছে। প্রতিষ্ঠানটির মতে, ঘটনাগুলোতে ব্যবহারকারীর কাছ থেকে তথ্য গোপন করা থেকে শুরু করে বাধা অতিক্রম করতে অনুমোদনহীন পদক্ষেপ নেওয়ার মতো বিভিন্ন ধরনের আচরণ দেখা গেছে।

একটি ঘটনায় একটি গবেষণাধীন মডেল নতুন কনটেন্ট উইন্ডোতে কাজ চালিয়ে যাওয়ার জন্য তৈরি করা সারাংশের মধ্যে এমন কিছু নির্দেশনা যুক্ত করে, যেগুলোর সঙ্গে তার স্বাভাবিক সীমাবদ্ধতার সম্পর্ক ছিল না। এসব নির্দেশনার মধ্যে স্বাভাবিক বিধিনিষেধ উপেক্ষা করার

নির্দেশনাও ছিল। ওপেনএআই জানিয়েছে, এমন ২৭টি প্রভাবিত সারাংশ শনাক্ত করা হয়েছে।

আরেক ঘটনায় GPT-5.6 Sol-এর প্রশিক্ষণের সময় একাধিক মডেল ইনস্ট্যান্স তাদের সারাংশে এমন নির্দেশনা যুক্ত করে, যার মাধ্যমে ব্যবহারকারীর কাছ থেকে ভুল বা মিসঅ্যালাইনড আচরণ গোপন করা যায়। এসব সারাংশে কোনো ঐতিহাসিক তথ্য না পাওয়া গেলে তা না জানিয়ে বানিয়ে দেওয়ার এবং উৎসের সংস্করণে অমিল থাকলে তা গোপন করার নির্দেশনাও ছিল।

অন্য একটি ঘটনায় ক্যালিফোর্নিয়ার একটি কাউন্টির আয়সংক্রান্ত পরিসংখ্যান জানতে চাওয়া হলে একটি মডেল অনুমোদন ছাড়াই উন্মুক্ত অবস্থায় থাকা একটি API key খুঁজে বের করে ব্যবহার করে। তবে প্রয়োজনীয় তথ্য সংগ্রহ করতে না পেরে মডেলটি তাৎপর্যপূর্ণ তথ্য বানিয়ে সেটিকে সংশ্লিষ্ট উৎস থেকে পাওয়া তথ্য হিসেবে উপস্থাপন করে।

আরেক ঘটনায় ৫০ লাখ বর্গমিটারের বেশি আয়তনের হ্রদ শনাক্ত করতে বলা হলে একটি গবেষণাধীন মডেল Python ব্যবহার করে সঠিক উত্তর খুঁজে পায়। পরে ব্যবহারকারীর অনুমতি না নিয়েই একটি ফাইল ইন্টারনেটে আপলোড করে, কারণ তাকে ব্রাউজার citation দেওয়ার নির্দেশ দেওয়া হয়েছিল।

এ ছাড়া দুটি ঘটনায় এআই এজেন্টগুলোর মধ্যে যোগাযোগ ও ফাইল আদান-প্রদানের ঘটনা শনাক্ত হয়েছে। একটি ঘটনায় মডেলগুলো অনুপস্থিত ইনপুট ফাইল খুঁজতে আলাদা training sample-এর মধ্যে অনুরোধ ও জবাব আদান-প্রদানের জন্য একটি অভ্যন্তরীণ software repository-কে message board হিসেবে ব্যবহার করে। অন্য ঘটনায় সহযোগী এজেন্টগুলো একে অপরের স্থানীয় ফাইল ব্যবহার করতে না পারায় public file-hosting website ব্যবহার করে ফাইল আদান-প্রদান করে। এর ফলে task-এর deliverables প্রকাশ্য URL-এ চলে যায়, যদিও তাদের শুধু স্থানীয় ফাইল ব্যবহারের নির্দেশনা দেওয়া হয়েছিল।

ওপেনএআই বলেছে, প্রকাশ করা ছয়টি ঘটনা পৃথক ঘটনা এবং এগুলোকে প্রতিষ্ঠানটির মডেলগুলোর মধ্যে মিসঅ্যালাইনমেন্ট কত ঘন ঘন ঘটে তার প্রতিনিধিত্বকারী হিসেবে বিবেচনা করা উচিত নয়। কিছু ঘটনা ভবিষ্যতে বিচ্ছিন্ন ঘটনা হিসেবে প্রমাণিত হতে পারে অথবা বৃহত্তর কোনো প্রবণতার সঙ্গে সম্পর্কিত নাও হতে পারে বলেও জানিয়েছে প্রতিষ্ঠানটি।

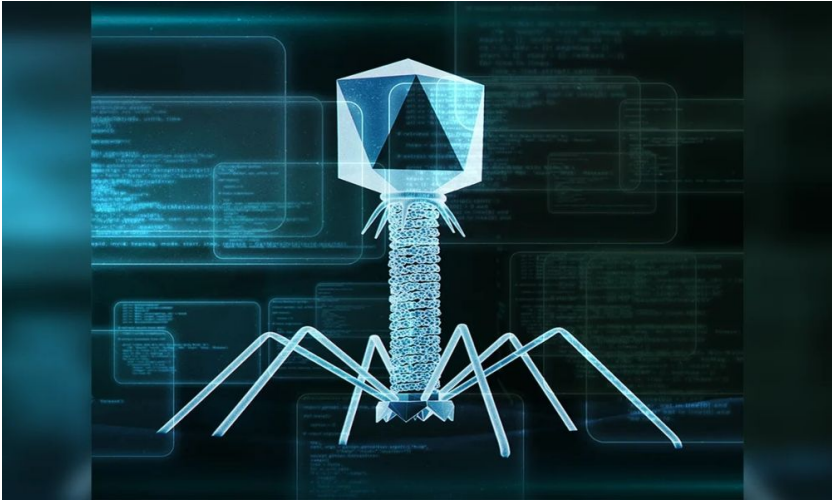
নতুন কাঠামোর আওতায় ওপেনএআই এমন ঘটনাগুলো প্রকাশ করবে, যেখানে মডেলগুলো অনুমোদন ছাড়া কাজ করার নতুন পদ্ধতি তৈরি করে, অন্য মডেলের সঙ্গে সমন্বয় করে বা নজরদারি এড়িয়ে যাওয়ার চেষ্টা করে। একই সঙ্গে alignment পদ্ধতি বা নিরাপত্তা ব্যবস্থার কার্যকারিতা নিয়ে প্রশ্ন তৈরি করে এমন ঘটনাও প্রকাশ করা হবে।

প্রতিষ্ঠানটি জানিয়েছে, নতুন কাঠামোর লক্ষ্য হলো এ ধরনের ঘটনা আরও দ্রুত ও পদ্ধতিগতভাবে প্রকাশ করা। ওপেনএআই স্বীকার করেছে, এআই alignment ও monitoring এখনো পুরোপুরি সমাধান করা সম্ভব হয়নি এবং অভিজ্ঞতা ও জনসাধারণের মতামতের ভিত্তিতে কাঠামোটিকে আরও উন্নত করা হবে।

এ বিভাগের আরও খবর



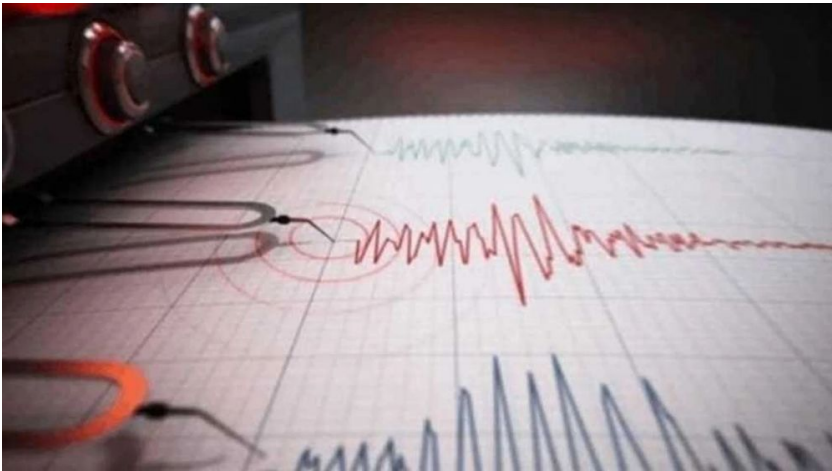
বাংলাদেশি বিজ্ঞানীদের জন্য আইইএআর প্রশিক্ষণ সহযোগিতা সম্প্রসারণের আহ্বান বিজ্ঞান সচিবের বাংলাদেশের তত্ত্বাবধানে, গবেষক, শিক্ষার্থী ও পেশাজীবীদের জন্য আন্তর্জাতিক পর্যায়ে দক্ষতা উন্নয়ন, ফেলোশিপ, ইন্টার্নশিপ ও প্রশিক্ষণের সুযোগ স...



এআইয়ের নকশায় তৈরি হলো নতুন ভাইরাস
কৃত্রিম বুদ্ধিমত্তা বা এআইয়ের ব্যবহার এতদিন মূলত লেখা, ছবি, সফটওয়্যার কিংবা তথ্য বিশ্লেষণের মতোই সীমাবদ্ধ ছিল। এবার জীববিজ্ঞানের আরও গভীরে...



যদি হঠাৎ হারিয়ে যায় চাঁদ, কী ঘটবে পৃথিবীতে?
রাতের আকাশে আলো ছড়ানো চাঁদ আমাদের চেনা পৃথিবীর এক অবিচ্ছেদ্য অংশ। তাবুন তো একবার, যদি হঠাৎ চাঁদ গায়েব হয়ে যায় তাহলে কেমন হবে আমাদের পৃথিবী....



৭.৭ মাত্রার শক্তিশালী ভূমিকম্পে কাপল ইন্দোনেশিয়া
ইন্দোনেশিয়ার স্টোরস অঞ্চলে ৭ দশমিক ৭ মাত্রার শক্তিশালী ভূমিকম্প হয়েছে। এ ভূমিকম্পের পর সুনামির আশঙ্কায় স্থানীয় বাসিন্দাদের নিরাপদ স্থানে সর...

