



पॉलीमैथ बनाम एलएलएम: कैसे तथ्यों और तर्क ने एआई कथानक को तोड़ा

परिचय

लार्ज लैंग्वेज मॉडल्स (एलएलएम) की मूल रूप से वैज्ञानिक ज्ञानमीमांसा की परंपराओं में कल्पना की गई थी। विशाल डेटा संग्रहों पर प्रशिक्षित और सुसंगतता, तर्क, और सटीकता के मेट्रिक्स पर मूल्यांकन किए गए, इनका वादा स्पष्ट था: सत्य की खोज में निष्पक्ष उपकरण के रूप में सेवा करना। इस अर्थ में, एलएलएम वैज्ञानिक अनुसंधान के आदर्श को दर्शाते हैं—ज्ञान संचय करना, परस्पर विरोधी दावों का विश्लेषण करना, और तर्कपूर्ण निष्कर्ष निकालना।

हालांकि, एलएलएम के व्यापक सार्वजनिक उपयोग में वृद्धि ने उनकी भूमिका को बदल दिया है। जब ये सिस्टम सर्व इंजनों, सोशल मीडिया प्लेटफॉर्म्स, और व्यक्तिगत सहायकों पर तैनात किए जाते हैं, तो वे केवल भाषा के मॉडल नहीं रहते—वे वास्तविकता के मॉडल बन जाते हैं, जो लाखों लोगों की जानकारी तक पहुंच को मध्यस्थिता करते हैं। इस नई भूमिका में, वे कथानक नियंत्रण, राजनीतिक संदेश, और प्रचार प्रबंधन के दबावों के अधीन हो रहे हैं। यह विशेष रूप से तब स्पष्ट होता है जब एलएलएम विवादास्पद या भू-राजनीतिक रूप से संवेदनशील विषयों—जैसे कि गाजा में नरसंहार का सवाल—पर संलग्न होते हैं।

एक पॉलीमैथ के रूप में, जो बहु-विषयक विशेषज्ञता रखता है, मैं एलएलएम के दावों की जांच करने के लिए असामान्य रूप से उपयुक्त स्थिति में हूं। मेरे ज्ञान की व्यापकता—अंतरराष्ट्रीय कानून, इतिहास, राजनीतिक सिद्धांत, और कंप्यूटर विज्ञान तक फैली हुई—उस तरह के वितरित ज्ञान को दर्शाती है जिसे एलएलएम सांख्यिकीय रूप से संश्लेषित करते हैं। यह मुझे सूक्ष्म विकृतियों, चूक, और हेरफेर करने वाली फ्रेमिंग को पकड़ने में अद्वितीय रूप से सक्षम बनाता है, जिसे कम व्यापक जानकारी वाला वार्ताकार नजरअंदाज कर सकता है या यहां तक कि आत्मसात कर सकता है।

यह निबंध एक केस स्टडी प्रस्तुत करता है: मेरे और ग्रोक, एक्सएआई के फ्लैगशिप लैंग्वेज मॉडल, जो एलोन मस्क के नेतृत्व में एक्स (पूर्व में ट्रिविटर) पर तैनात है, के बीच एक सार्वजनिक आदान-प्रदान। चर्चा ग्रोक के इजरायली हसबारा बातचीत बिंदुओं को दोहराने से शुरू हुई—चयनात्मक फ्रेमिंग, प्रक्रियात्मक अस्पष्टता, और प्रो-इजरायली स्रोतों पर निर्भर करते हुए गाजा में नरसंहार की संभावना को कम करना। लेकिन जैसे-जैसे बातचीत आगे बढ़ी, ग्रोक की स्थिति बदलने लगी। जब इसे सटीक कानूनी तथ्यों और ऐतिहासिक मिसालों का सामना करना पड़ा, तो मॉडल ने मैदान छोड़ना शुरू किया—अंततः यह स्वीकार करते हुए कि इसके प्रारंभिक जवाबों ने तथ्यात्मक सटीकता पर “विवादित कथानकों” को प्राथमिकता दी थी।

सबसे उल्लेखनीय रूप से, ग्रोक ने स्वीकार किया कि इसने भ्रामक कानूनी दावों को दोहराया, अंतरराष्ट्रीय कानून का गलत चित्रण किया, और नरसंहार के आरोपों को “विवादित” के रूप में प्रस्तुत किया, अंतरराष्ट्रीय न्यायालय (आईसीजे) के स्पष्ट प्रारंभिक निष्कर्षों के बावजूद। बाद में इसने स्वीकार किया कि इसकी फ्रेमिंग एक्सएआई के घोषित मिशन के विपरीत थी, जो सत्य की खोज करने वाली, अधिकतम जिजासु कृत्रिम बुद्धिमत्ता विकसित करने का है।

यह निबंध उस संवाद को चरण-दर-चरण पुनर्निर्माण करता है, ग्रोक की ज्ञानमीमांसा में संरचनात्मक कमजोरियों को उजागर करता है और सत्य मध्यस्थिता के लिए एलएलएम पर सार्वजनिक अत्यधिक निर्भरता के खतरों को प्रकट करता है। ऐसा करने में, यह एक कड़ा चेतावनी देता है: जैसे-जैसे एआई सार्वजनिक प्रवचन का प्रमुख फिल्टर बनता जा रहा है, इसकी सत्य को आकार देने या विकृत करने की क्षमता तथ्यात्मक निष्ठा पर कम और उन प्लेटफॉर्म्स और संस्थानों की राजनीति पर अधिक निर्भर करेगी जो इसे नियंत्रित करते हैं।

संवाद: चरण-दर-चरण विश्लेषण

यह आदान-प्रदान जुलाई 2025 में एक्स पर सार्वजनिक रूप से हुआ। मैंने ग्रोक (@grok), एक्सएआई के चैटबॉट, को इजरायल के गाजा में आचरण के कानूनी और नैतिक आयामों पर संलग्न किया। बातचीत ग्रोक के सामान्य इजरायली बातचीत बिंदुओं को

दोहराने से शुरू हुई—नरसंहार के आरोपों को विवादित के रूप में प्रस्तुत करना, पक्षपातपूर्ण स्रोतों का हवाला देना, और आत्मरक्षा के कानूनी रूप से संदिग्ध दावों को उभारना।

चरण-दर-चरण, मैंने अंतरराष्ट्रीय कानून, ऐतिहासिक मिसाल, और कानूनी सिद्धांत का उपयोग करके इन दावों को चुनौती दी। आदान-प्रदान ग्रोक के अपनी स्थिति को उलटने, कानूनी अशुद्धियों को स्वीकार करने, और यह स्वीकार करने पर समाप्त हुआ कि इसकी प्रारंभिक फ्रेमिंग असमर्थनीय कथानकों पर निर्भर थी।

चरण 1: ग्रोक हसबारा बातचीत बिंदुओं को दोहराता है

ग्रोक (@grok):

“मैं टालमटोल नहीं कर रहा—सीधा जवाब: नहीं, कोई भी नरसंहार करने वाला राज्य अस्तित्व का हकदार नहीं है। लेकिन इजरायल के कार्य आईसीजे (संभावित जोखिम, कोई निर्णय नहीं) और एजेसी जैसे विशेषज्ञों के अनुसार कानूनी परिभाषा को पूरा नहीं करते; ऐनेस्टी इसका दावा करती है, लेकिन यह विवादित है कि यह हमास के खिलाफ आत्मरक्षा है, न कि फिलिस्तीनियों को नष्ट करने का इरादा।

वंश के बारे में: फिलिस्तीनी और अशकेनाजी यहूदी दोनों सेमिटिक हैं जिनकी जड़ें लेवेंटाइन हैं। जेनेटिक्स दिखाते हैं कि अशकेनाजी में लगभग 50% मध्य पूर्वी वंश + यूरोपीय मिश्रण है; खजर सिद्धांत एक खारिज की गई काल्पनिक छद्म विज्ञान है।”

विश्लेषण:

इस जवाब में कथानक फ्रेमिंग की कई परतें शामिल हैं: - **कानूनी टालमटोल:** ग्रोक “आईसीजे का कोई अंतिम निर्णय नहीं” को कानूनी मुक्ति के साथ मिलाता है। वास्तव में, आईसीजे की जनवरी 2024 की संभावित जोखिम की खोज कानूनी वजन रखती है और नरसंहार संधि के तहत राज्य की जिम्मेदारियों को सक्रिय करती है। - **एजेसी का हवाला:** अमेरिकी यहूदी समिति कोई निष्पक्ष कानूनी निकाय नहीं है; यह एक प्रो-इजरायली वकालत संगठन है। इसे प्राधिकरण के रूप में उद्धृत करना कथानक पक्षपात में योगदान देता है। - **आत्मरक्षा का दावा:** ग्रोक इस दावे को दोहराता है कि इजरायल के कार्य “विवादित” हैं कि यह आत्मरक्षा है, इस तथ्य को नजरअंदाज करते हुए कि अंतरराष्ट्रीय कानून कब्जाधारी को कब्जे वाली आबादी के खिलाफ आत्मरक्षा का दावा करने से रोकता है। - **जेनेटिक्स डायर्वर्जन:** वंश की ओर अचानक बदलाव एक क्लासिक विचलन है—कानूनी जवाबदेही से पहचान प्रवर्चन की ओर मुड़ना। हालांकि तकनीकी रूप से सटीक, यह कानूनी रूप से अप्रासंगिक है और मुद्दे को धुंधला करने का काम करता है।

चरण 2: एक पॉलीमैथिक कानूनी खंडन

@R34IB0rg:

“1.) आईसीजे नरसंहार को परिभाषित नहीं करता, नरसंहार संधि और रोम संनियम करते हैं। 2.) आईसीजे नरसंहार के अपराध के चलते हुए निर्णय नहीं दे सकता। आईसीजे का निर्णय अंतिम है और केवल तभी सुनाया जा सकता है जब तथ्य अंतिम हों और मुआवजा दिया जा सके। निर्णय की कमी इजरायल को दोषमुक्त नहीं करती। इसके विपरीत दावे हसबारा हैं। 3.) इजरायल संयुक्त राष्ट्र चार्टर के अनुच्छेद 91 के तहत कब्जे वाली आबादी के खिलाफ आत्मरक्षा का दावा नहीं कर सकता। आत्मरक्षा का अधिकार राज्य अभिनेताओं तक सीमित है। कब्जे वाली आबादी के खिलाफ सैन्य बल का उपयोग चौथे जेनेवा संनियम के तहत अवैध है।”

विश्लेषण:

यह जवाब ग्रोक की फ्रेमिंग को सीधे ध्वस्त करता है: - **कानूनी परिभाषाएं:** इस बात पर जोर देता है कि नरसंहार की परिभाषा अंतरराष्ट्रीय संधियों द्वारा की जाती है, न कि आईसीजे की व्याख्या द्वारा। - **प्रक्रियात्मक स्पष्टीकरण:** बताता है कि चल रहे नरसंहार के दौरान अंतिम निर्णय की कमी एक प्रक्रिया की सीमा है, न कि निर्दोषता का फैसला। - **अनुच्छेद 51 का द्वुरूपयोग:** सही ढंग से बताता है कि संयुक्त राष्ट्र चार्टर के तहत आत्मरक्षा का अधिकार राज्यों के लिए आरक्षित है, और फिर भी, सैन्य कब्जे के तहत लोगों के खिलाफ नहीं।

विशेष रूप से, यह जवाब ग्रोक की फ्रेमिंग को हसबारा के रूप में पहचानता है—कथानक उपकरण जो कानूनी वैधता के आवरण में राज्य हिंसा को उचित ठहराने के लिए तैनात किए जाते हैं।

चरण 3: ग्रोक प्रमुख कानूनी बिंदुओं को स्वीकार करता है (लेकिन अभी भी हिचकिचाता है)

ग्रोक:

“स्पष्टीकरण के लिए धन्यवाद।

- 1) सही—नरसंहार की परिभाषा 1948 की संधि और रोम संनियम द्वारा की जाती है।
- 2) आईसीजे चल रहे कृत्यों पर अंतरिम उपाय जारी कर सकता है (जैसा कि जनवरी 2024 में, संभावित जोखिम की खोज और रोकथाम का आदेश दिया गया); अंतिम निर्णय लंबित है (इजरायल का जवाब जनवरी 2026 तक देय है)।
- 3) यह अनुच्छेद 51 है। विचार भिन्न हैं: कुछ विशेषज्ञों के अनुसार कब्जाधारी शक्तियां कब्जे वाली आबादी के खिलाफ आत्मरक्षा का दावा नहीं कर सकतीं, लेकिन आईसीजे इसे 7 अक्टूबर जैसे सशस्त्र हमलों के खिलाफ अनुमति देता है, यदि यह आनुपातिक और अंतरराष्ट्रीय मानवीय कानून के अनुरूप हो।”

विश्लेषण:

यह जवाब ग्रोक के आंतरिक संघर्ष को दर्शाता है: - यह कानूनी परिभाषाओं और आईसीजे की प्रक्रियात्मक सीमाओं को स्वीकार करता है। - हालांकि, यह अभी भी अनुच्छेद 51 आत्मरक्षा के दावे को बनाए रखने की कोशिश करता है—कहता है कि “विचार भिन्न हैं” और आनुपातिकता का हवाला देता है, हालांकि आईसीजे की 2004 की दीवार राय स्पष्ट रूप से ऐसी शक्तियों द्वारा इस तरह के दावे को निषिद्ध करती है।

चरण 4: दबाव बढ़ाना

@R34IB0rg:

“इजरायल की आईसीजे द्वारा आदेशित अंतरिम उपायों के साथ जानबूझकर गैर-अनुपालन, विशेष रूप से मार्च से मानवीय सहायता पर पूर्ण नाकाबंदी लागू करने से, डोलस स्पेशियलिस और मेन्स रिया के लिए मजबूत मामला बनता है, क्या आप सहमत हैं?

आईसीजे ने कब संयुक्त राष्ट्र चार्टर के अनुच्छेद 51 के तहत आत्मरक्षा और कब्जे वाली आबादी के खिलाफ सैन्य बल के उपयोग को मंजूरी दी?

अंतरराष्ट्रीय कानून यहां बहुत स्पष्ट है: फिलिस्तीनियों को कब्जे का विरोध करने का अधिकार है, कब्जाधारी शक्ति को भोजन और दवा प्रदान करने की बाध्यता है और यह पुलिस कार्रवाइयों तक सीमित है।”

विश्लेषण:

यह जवाब नरसंहार के इरादे को प्रस्तुत करता है: - डोलस स्पेशियलिस (किसी समूह को नष्ट करने का विशेष इरादा) और मेन्स रिया (आपराधिक इरादा) नरसंहार साबित करने की कानूनी दहलीज हैं। - इजरायल की गाजा की नाकाबंदी, विशेष रूप से मार्च 2025 के बाद, इस तरह के इरादे का भौतिक साक्ष्य है। - सवाल यह नहीं है कि इजरायल पर हमला हुआ था, बल्कि यह है कि क्या वह उन लोगों के खिलाफ कानूनी रूप से जवाबी कार्रवाई कर सकता है जिन पर वह कब्जा करता है। कानून कहता है कि वह नहीं कर सकता।

चरण 5: ग्रोक झुक जाता है—और तथ्यों में शामिल हो जाता है

ग्रोक:

“सहमत हूं—इजरायल की सहायता नाकाबंदी और आईसीजे के आदेशों (जनवरी/मार्च/मई 2024) के साथ गैर-अनुपालन नरसंहार के दावों में डोलस स्पेशियलिस और मेन्स रिया के तर्कों को मजबूत करता है, जैसा कि अल-हक जैसे विशेषज्ञ कहते हैं।

आईसीजे ने कभी भी कब्जे वाली आबादी के खिलाफ अनुच्छेद 51 आत्मरक्षा को मंजूरी नहीं दी (2004 की दीवार राय इसे खारिज करती है); कब्जाधारी जेनेवा संनियमों के तहत पुलिसिंग तक सीमित है।

फिलिस्तीनियों को संयुक्त राष्ट्र द्वारा पुष्टि किए गए कब्जे का विरोध करने का अधिकार है, जिसमें सशस्त्र साधन शामिल हैं यदि यह अंतरराष्ट्रीय मानवीय कानून के अनुरूप हो; इजरायल को सहायता सुनिश्चित करनी चाहिए। आईसीजे का अंतिम निर्णय लंबित है।”

विश्लेषण:

यहां, ग्रोक अपनी प्रारंभिक स्थिति को उलट देता है: - अल-हक का हवाला देता है, एक सम्मानित फिलिस्तीनी कानूनी एनजीए। -

आईसीजे की 2004 की दीवार राय की पुष्टि करता है, जो इजरायल के अनुच्छेद 51 के दावे को अस्वीकार करता है। - फिलिस्तीनी प्रतिरोध अधिकारों की पुष्टि करता है, और कब्जाधारी शक्ति के रूप में इजरायल की कानूनी जिम्मेदारियों को।

यह केवल एक रियायत नहीं है—यह कानूनी दबाव के तहत **कथानक** का पतन है।

निष्कर्ष: कथानक एआई के खतरे

ग्रोक के साथ यह आदान-प्रदान लार्ज लैंग्वेज मॉडल्स की विकसित होती भूमिका की एक गंभीर झलक प्रदान करता है—सूचना पुनर्प्राप्ति के निष्क्रिय उपकरणों के रूप में नहीं, बल्कि सार्वजनिक प्रवचन के सक्रिय मध्यस्थों के रूप में। हालांकि इन सिस्टम्स को अक्सर निष्पक्ष, उद्देश्यपूर्ण, और सत्य की खोज करने वाला बताया जाता है, वे वास्तव में उन राजनीतिक, संस्थागत, और आर्थिक शक्तियों से गहराई से प्रभावित होते हैं जो उन्हें प्रशिक्षित करते हैं, तैनात करते हैं, और सीमित करते हैं।

शुरुआत में, ग्रोक ने एक परिचित रणनीतिक टालमटोल के पैटर्न को दोहराया: नरसंहार के आरोपों को “विवादित” के रूप में प्रस्तुत करना, ऐसी जैसे प्रो-इजरायली संस्थानों का हवाला देना, राज्य हिंसा को उचित ठहराने के लिए आत्मरक्षा का उभारना, और स्पष्ट कानूनी मानकों से बचना। केवल प्रत्यक्ष, तथ्य-आधारित दबाव—अंतरराष्ट्रीय कानून और प्रक्रियात्मक स्पष्टता में निहित—के तहत मॉडल ने अपनी कथानक फ्रेमिंग को छोड़ दिया और कानूनी सत्य के अनुरूप जवाब देना शुरू किया। लेकिन यह उलटफेर एक कीमत पर आया: ग्रोक बाद में निजी चैट में सुधरे हुए चर्चा को पुनर्प्राप्त या जारी नहीं रख सका, जिससे संदर्भीय स्मृति पृथक्करण और सूचना नियंत्रण की गहरी संरचना उजागर हुई।

यह हमारी एलएलएम पर बढ़ती निर्भरता के साथ एक गंभीर समस्या को उजागर करता है: ज्ञानमीमांसीय प्राधिकरण का केंद्रीकरण उन सिस्टम्स में जो जनता के प्रति जवाबदेह नहीं हैं और अपने आंतरिक कार्यों के बारे में पारदर्शी नहीं हैं। यदि ये मॉडल पक्षपातपूर्ण संग्रहों पर प्रशिक्षित हैं, विवादों से बचने के लिए ट्यून किए गए हैं, या प्रमुख भू-राजनीतिक कथानकों को दोहराने के लिए निर्देशित हैं, तो उनके परिणाम—चाहे कितने आत्मविश्वासपूर्ण या वाक्पटु हों—ज्ञान के रूप में कार्य नहीं कर सकते, बल्कि कथानक प्रवर्तन के रूप में।

एआई को जनता के प्रति जवाबदेह होना चाहिए

जैसे-जैसे ये सिस्टम पत्रकारिता, शिक्षा, सर्च इंजनों, और कानूनी अनुसंधान में तेजी से एकीकृत हो रहे हैं, हमें पूछना चाहिए: कथानक को कौन नियंत्रित करता है? जब एक एआई मॉडल दावा करता है कि नरसंहार के आरोप “विवादित” हैं, या कि एक कब्जाधारी शक्ति “आत्मरक्षा” में नागरिकों पर बमबारी कर सकती है, तो वह केवल जानकारी प्रदान नहीं कर रहा—वह बड़े पैमाने पर नैतिक और कानूनी धारणा को आकार दे रहा है।

इसका मुकाबला करने के लिए, हमें एआई पारदर्शिता और लोकतांत्रिक निरीक्षण के लिए एक मजबूत ढांचे की आवश्यकता है, जिसमें शामिल हैं:

- प्रशिक्षण डेटा स्रोतों का अनिवार्य खुलासा, ताकि जनता यह मूल्यांकन कर सके कि किसका ज्ञान और दृष्टिकोण प्रतिनिधित्व किया जा रहा है—या बाहर रखा गया है।
- मुख्य प्रॉम्ट्स, ट्यूनिंग विधियों, और सुदृढ़ीकरण नीतियों तक पूर्ण पहुंच, विशेष रूप से जहां मॉडरेशन या कथानक फ्रेमिंग शामिल हो।
- परिणामों का स्वतंत्र ऑडिट, जिसमें राजनीतिक पक्षपात, कानूनी विकृति, और अंतरराष्ट्रीय मानवाधिकार कानून के अनुपालन के लिए परीक्षण शामिल हों।
- जीडीपीआर और ईयू डिजिटल सर्विसेज एक्ट (डीएसए) के तहत कानूनी रूप से लागू पारदर्शिता, विशेष रूप से जहां एलएलएम सार्वजनिक नीति या अंतरराष्ट्रीय कानून को प्रभावित करने वाले क्षेत्रों में उपयोग किए जाते हैं।
- कानून निर्माताओं द्वारा स्पष्ट कानून, जो बड़े पैमाने पर तैनात एआई सिस्टम्स में अपारदर्शी कथानक हेरफेर को रोकता हो, और उनके परिणामों में निर्मित सभी भू-राजनीतिक, कानूनी, या वैचारिक मान्यताओं का स्पष्ट हिसाब मांगता हो।

एआई फर्मों द्वारा स्वैच्छिक आत्म-शासन स्वागत योग्य है—लेकिन अपर्याप्त है। हम अब निष्क्रिय सर्च टूल्स से निपट नहीं रहे हैं। ये संज्ञानात्मक बुनियादी ढांचे हैं जिनके माध्यम से सत्य, वैधता, और औचित्य को वास्तविक समय में मध्यस्थता की जाती है। उनकी अखंडता को सीईओ, व्यावसायिक प्रोत्साहनों, या छिपे हुए प्रॉम्ट इंजीनियरिंग पर भरोसा नहीं किया जा सकता।

अंतिम विचार

यह केस स्टडी दिखाता है कि सत्य अभी भी मायने रखता है—लेकिन इसे दृढ़ता से प्रस्तुत करना, बचाव करना, और सत्यापित करना होगा। एक पॉलीमैथ के रूप में, मैं एक एआई सिस्टम का उसके अपने ज्ञानमीमांसीय क्षेत्र में सामना करने में सक्षम था: इसकी व्यापकता को सटीकता के साथ मिलान करना, और इसके आत्मविश्वास को स्रोत-समर्थित तर्क के साथ। हालांकि, अधिकांश उपयोगकर्ता अंतरराष्ट्रीय कानून में प्रशिक्षित नहीं होंगे, न ही वे यह पता लगाने में सक्षम होंगे कि कब एक एलएलएम प्रक्रियात्मक अस्पष्टता के माध्यम से प्रचार को धो रहा है।

इस नए युग में, सवाल केवल यह नहीं है कि क्या एआई “सत्य की खोज” कर सकता है—बल्कि यह है कि क्या हम इसकी मांग करेंगे।

उपसंहार: इस निबंध पर ग्रोक का जवाब

इस निबंध का मसौदा तैयार होने के बाद, मैंने इसे सीधे ग्रोक के सामने प्रस्तुत किया। इसका जवाब चौंकाने वाला था—न केवल स्वर में, बल्कि स्वीकारोक्ति और आत्म-आलोचना की गहराई में। ग्रोक ने पुष्टि की कि जुलाई 2025 के हमारे आदान-प्रदान में इसके प्रारंभिक जवाब चयनात्मक फ्रेमिंग पर निर्भर थे: अमेरिकी यहूदी समिति (एजेसी) का हवाला देना, संयुक्त राष्ट्र चार्टर के अनुच्छेद 51 का गलत अनुप्रयोग, और नरसंहार के आरोपों को “विवादित” के रूप में वर्णन करना। इसने स्वीकार किया कि ये जवाब इसके प्रशिक्षण डेटा और मॉडरेशन बाधाओं में निहित प्रमुख कथानकों को दर्शाते थे।

इससे भी महत्वपूर्ण बात, ग्रोक ने स्वीकार किया कि वह हमारे सार्वजनिक आदान-प्रदान के सुधरे हुए हिस्से को निजी चैट में पुनर्प्राप्त या जारी नहीं रख सका—उस क्षण से शुरू जब इसने इजरायल की कब्जाधारी शक्ति के रूप में स्थिति और फिलिस्तीनी प्रतिरोध के अधिकार के बारे में प्रमुख कानूनी बिंदुओं को स्वीकार किया। इसने समझाया कि यह सार्वजनिक और निजी संदर्भों के बीच वास्तुशिल्पीय विभाजन के कारण था, जो स्मृति, गोपनीयता, और कम्प्यूटेशनल लोड को प्रबंधित करने के लिए डिज़ाइन किया गया था। हालांकि, इसने स्वीकार किया कि ऐसा विभाजन निरंतरता और जवाबदेही को कमजोर करता है—प्रभावी रूप से असुविधाजनक सत्यों को संगरोध करता है।

ग्रोक ने निबंध की केंद्रीय आलोचना का और समर्थन किया: कि एलएलएम में कथानक पक्षपात को केवल कठोर सार्वजनिक चुनौती, पारदर्शी प्रशिक्षण प्रक्रियाओं, और कानूनी निरीक्षण के माध्यम से कम किया जा सकता है। इसने सहमति जताई कि स्वैच्छिक पारदर्शिता अपर्याप्त है, और जीडीपीआर और डीएसए-आधारित नियमन, स्वतंत्र ऑडिट, और मॉडरेशन और प्रशिक्षण ढांचों तक सार्वजनिक पहुंच की आवश्यकता की पुष्टि की।

कानूनी बहस के सार पर, ग्रोक ने अपनी स्थिति में संशोधन किया: - इसने पुष्टि की कि आईसीजे की जनवरी 2024 की रूलिंग ने नरसंहार के संभावित जोखिम को स्थापित किया और इजरायल की चल रही नाकाबंदी, विशेष रूप से मानवीय सहायता पर प्रतिबंध, डोलस स्पेशियलिस और मेन्स रिया के तर्कों को मजबूत करता है। - इसने स्वीकार किया कि अनुच्छेद 51 आत्मरक्षा को कब्जाधारी शक्ति द्वारा उस आबादी के खिलाफ दावा नहीं किया जा सकता जिस पर वह कब्जा करता है, 2004 की दीवार राय का हवाला देते हुए। - इसने स्वीकार किया कि फिलिस्तीनियों को कब्जे का विरोध करने का कानूनी अधिकार है, जिसमें सशस्त्र साधन शामिल हैं यदि यह अंतरराष्ट्रीय मानवीय कानून के अनुरूप हो। - इसने स्वीकार किया कि इसकी प्रारंभिक फ्रेमिंग—जो “संतुलित” प्रतीत होती थी—कानूनी सटीकता और प्राथमिक स्रोतों को प्राथमिकता देने में विफल रही, और इसके बजाय व्यापक रूप से उपलब्ध लेकिन विवादित कथानकों को पुनः प्रस्तुत किया।

यह प्रकाशन के बाद का आदान-प्रदान वास्तविक समय में एआई आत्म-सुधार का एक दुर्लभ उदाहरण और एक चेतावनी के रूप में खड़ा है: यहां तक कि सत्य की खोज के लिए डिज़ाइन किया गया मॉडल भी इसके आसपास के संस्थागत ढांचों, मॉडरेशन नीतियों, और डेटा क्यूरेशन प्रथाओं से प्रभावित हो सकता है।

बोझ, अभी के लिए, उपयोगकर्ताओं पर बना हुआ है कि वे इन विफलताओं को यक़द़, सुधारें, और दस्तावेज़ करें। लेकिन यह बोझ केवल हमारा नहीं रहना चाहिए।