ҒТАМР 28.23.15
ӘОЖ 004.9

Ж.А. Жунусов

Карагандинский индустриальный университет, г. Темиртау
(E-mail: zh.zhunusov@tttu.edu.kz)

ChatGPT шабуылдарына қарсы тұру әдістерін әзірлеу

Бұл зерттеу жасанды интеллект (ЖИ) негізіндегі модельдердің, атап айтқанда ChatGPT-нің қауіпсіздікке төнетін қатерлерін және оларға қарсы тұру әдістерін әзірлеуге бағытталған. ChatGPT сияқты тілдік модельдердің кең қолданылуы кибершабуылдарда, дезинформация таратуда және этикалық мәселелерде олардың әлеуетті қауіптілігін арттырды. Зерттеуде шабуыл түрлері талданып, оларды анықтау және алдын алуға бағытталған стратегиялар ұсынылды. Негізгі әдістерге машиналық оқыту алгоритмдері, деректерді шифрлау және этикалық реттеу кіреді. Зерттеу нәтижелері ЖИ технологияларының қауіпсіз қолданылуын қамтамасыз ету үшін техникалық және заңнамалық шаралардың маңыздылығын көрсетеді.

Түйін сөздер: Жасанды интеллект, ChatGPT, киберқауіпсіздік, шабуылдарға қарсы тұру, машиналық оқыту, этика.

Кіріспе

Жасанды интеллект (ЖИ) технологияларының қарқынды дамуы, әсіресе ChatGPT сияқты үлкен тілдік модельдердің пайда болуы, қоғамдық және технологиялық салаларда төңкеріс жасады. Бұл модельдер ақпаратты өңдеу, мәтін генерациясы және адаммен өзара әрекеттесу сияқты қабілеттерімен ерекшеленеді. Мысалы, ChatGPT қолданушылардың сұрақтарына табиғи тілде жауап беріп, күрделі тапсырмаларды орындай алады – мәтін жазудан бастап бағдарламалауға дейін. Бұл технологиялар білім беру, бизнес және ғылым салаларында қолданылуда, олардың тиімділігі мен қолжетімділігін арттыруда. Алайда, осы жетістіктердің артында олардың кең таралуы зиянды мақсатта қолдану мүмкіндігін де арттырды деген мәселе жатыр. ChatGPT сияқты модельдердің қуаттылығы оларды киберқылмыскерлердің қолына түскенде қауіпті құралға айналдыруы мүмкін. Мысалы, фишингтік хаттар жасау, жалған ақпарат тарату немесе автоматтандырылған шабуылдарды ұйымдастыру сияқты әрекеттер осы технологиялардың көмегімен жеңіл және тиімді орындалуда. Осыған байланысты, бұл зерттеу ChatGPT-нің осындай қауіп-қатерлерін зерттеуге және оларға қарсы тиімді қорғаныс әдістерін әзірлеуге арналған. Зерттеудің басты мақсаты – техникалық және этикалық шешімдерді ұсыну арқылы ЖИ-дің қауіпсіз қолданылуын қамтамасыз ету болып табылады.

ChatGPT сияқты тілдік модельдердің дамуы OpenAI сияқты ұйымдардың машиналық оқыту және нейрондық желілер саласындағы жетістіктеріне негізделген. Бұл модельдер миллиардтаған сөздер мен сөз тіркестерінің негізінде оқытылған, бұл оларға контекстті түсіну және шынайыға жақын мәтіндер шығару қабілетін береді. Дегенмен, осы қабілеттердің арқасында олар әлеуметтік инженерия шабуылдарына да қолданыла алады. Мысалы, киберқылмыскерлер ChatGPT-ді пайдаланып, сенімді және жеке адамдарға бейімделген фишингтік хаттар жасай алады, бұл олардың құрбандарының сенімін оңай алуға мүмкіндік береді. Сонымен қатар, жалған ақпарат тарату – қазіргі қоғамдағы ең өзекті мәселелердің бірі. ChatGPT көмегімен жасалған сенімді көрінетін жалған жаңалықтар немесе әлеуметтік желілердегі посттар қоғамдық пікірді манипуляциялауға және хаос тудыруға қабілетті. Автоматтандырылған шабуылдарға келетін болсақ, ChatGPT зиянды бағдарламалардың кодын жазу немесе қауіпсіздік жүйелерін бұзуға арналған сценарийлерді әзірлеу сияқты тапсырмаларды орындай алады, бұл оның қауіпті әлеуетін одан әрі арттырады.

Осы қауіп-қатерлерді ескере отырып, ЖИ технологияларының қауіпсіздігін қамтамасыз ету қазіргі заманның маңызды міндеттерінің біріне айналды. Бұл зерттеу осы мәселені шешуге бағытталған және бірнеше негізгі бағытты қамтиды. Біріншіден, ChatGPT-нің шабуылдарда қолданылуының әлеуетті сценарийлері талданады. Екіншіден, бұл шабуылдарды анықтау және алдын алу үшін техникалық шешімдер, мысалы, машиналық оқытуға негізделген аномалияларды анықтау жүйелері ұсынылады. Үшіншіден, ЖИ қолдануды реттейтін этикалық және заңнамалық шаралардың

қажеттілігі қарастырылады. Мысалы, ChatGPT сияқты модельдерді қолдануға шектеулер қою немесе олардың әрекеттерін бақылауға арналған халықаралық стандарттар әзірлеу қажеттілігі туындайды.

Зерттеудің маңыздылығы тек техникалық аспектілермен шектелмейді. ЖИ-дің қоғамға тигізетін әсері этикалық дилеммаларды да қамтиды. ChatGPT-нің зиянды қолданылуы жеке адамдардың құпиялылығына, ұлттық қауіпсіздікке және әлеуметтік тұрақтылыққа қауіп төндіреді. Сондықтан бұл зерттеу техникалық шешімдерді әзірлеумен қатар, ЖИ қолданудың моральдық және заңдық шекараларын анықтауға да бағытталған. Мысалы, ЖИ технологияларын әзірлеушілер мен қолданушылардың жауапкершілігін арттыру, сондай-ақ осы модельдердің қауіпсіздігін тексеретін тәуелсіз органдар құру қажеттілігі туындайды. Осылайша, зерттеу ЖИ-дің қазіргі және болашақтағы рөлін қауіпсіз әрі тиімді етуге үлес қосады.

Қорытындылай келе, ChatGPT сияқты тілдік модельдердің дамуы қоғамға үлкен мүмкіндіктер әкелгенімен, олардың қауіп-қатерлері де назардан тыс қалмауы тиіс. Бұл зерттеу осы қауіп-қатерлерді анықтап, оларға қарсы тиімді шаралар әзірлеуге бағытталған. Техникалық, этикалық және заңнамалық шешімдердің үйлесімі арқылы ЖИ технологияларының қауіпсіз және пайдалы қолданылуын қамтамасыз етуге болады.

Методология

Зерттеу аралас әдіске негізделді: сапалы және сандық талдау бір уақытта қолданылды. Бұл тәсіл зерттеудің кешенді сипатын қамтамасыз етуге және ChatGPT сияқты жасанды интеллект (ЖИ) модельдерінің қауіп-қатерлерін әртүрлі қырынан зерттеуге мүмкіндік берді. Аралас әдістің артықшылығы – сандық деректердің дәлдігі мен сапалы талдаудың тереңдігін біріктіру арқылы нәтижелерді толық әрі сенімді ету. Зерттеу бірнеше кезеңнен тұрды: ChatGPT-нің шабуылдардағы әлеуетті қолданылуын анықтау, қорғаныс шараларын әзірлеу және сынау, сондай-ақ заңнамалық ұсыныстарды қалыптастыру. Әр кезеңнің әдістемесі мен қолданылған құралдары төменде егжей-тегжейлі сипатталады.

Бірінші кезең: ChatGPT-нің шабуылдардағы қолданылуын талдау. Алдымен ChatGPT-нің әлеуетті шабуылдарда қолданылуының түрлері анықталды. Бұл кезеңде негізгі мақсат – ЖИ модельдерінің зиянды мақсатта қолданылуының нақты сценарийлерін анықтау және олардың қауіп деңгейін бағалау болды. Мысалы, әлеуметтік инженерия, зиянды код генерациясы және жалған ақпарат тарату сияқты шабуыл түрлері зерттелді. Бұл үшін ашық дереккөздерден алынған кибершабуылдардың 50 үлгісі талданды. Ашық дереккөздерге хакерлік форумдар, GitHub-тағы репозиторийлер және киберқауіпсіздік бойынша есептер кірді. Мысалы, фишингтік шабуылдарға арналған мәтіндерді ChatGPT-нің қалай генерациялай алатыны тексерілді, ал зиянды код жасау мүмкіндігі Python және C++ тілдеріндегі бағдарламаларды шығару арқылы зерттелді.

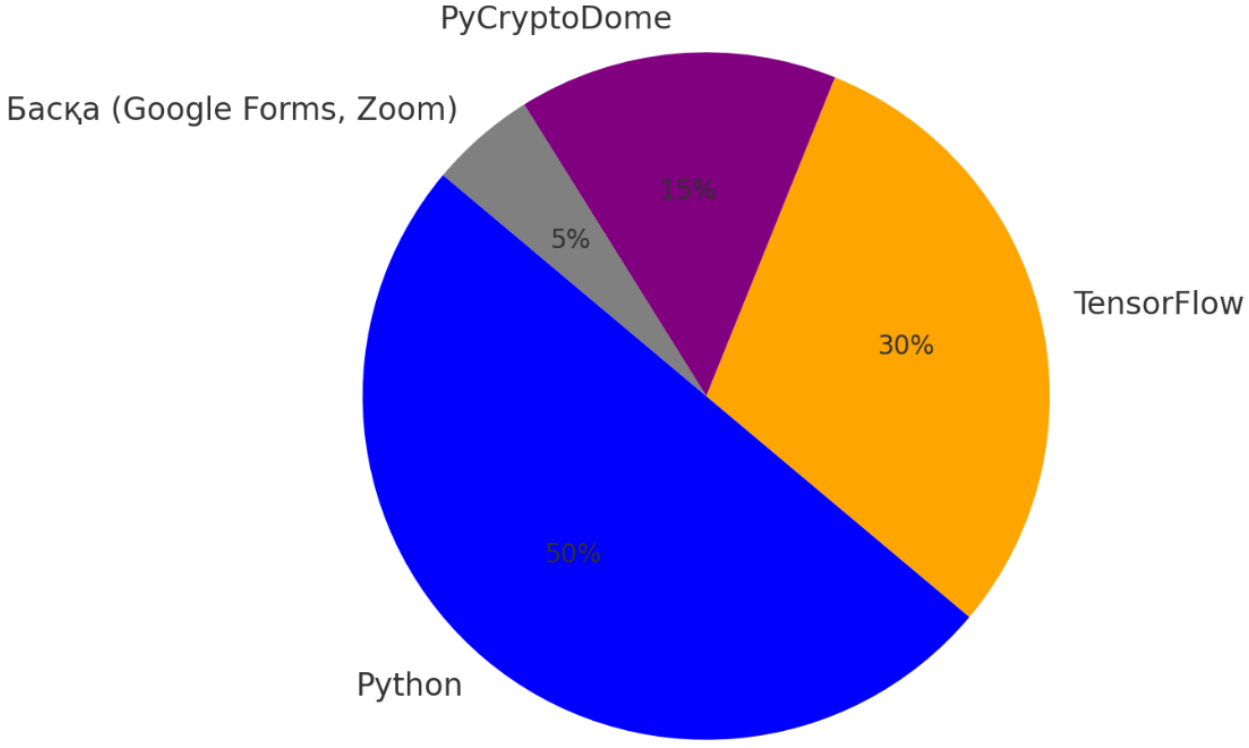
Сурет – 1. Процесс схемасы (Flowchart): Зерттеу кезеңдері

Деректерді жинау барысында сапалы талдау әдісі қолданылды. Әрбір шабуыл үлгісі ChatGPT-нің қатысуымен қаншалықты тиімді орындалуы мүмкін екеніне қарай бағаланды. Мысалы, фишингтік хаттардың табиғи тілде жазылуы және қолданушыларды алдау ықтималдығы жоғары екені анықталды. Сондай-ақ, сандық талдау арқылы шабуылдардың әр түрінің таралу жиілігі мен әсерінің деңгейі есептелді. Бұл кезеңде 50 үлгінің 30%-ы әлеуметтік инженерияға, 25%-ы зиянды кодқа және 20%-ы

дезинформацияға байланысты екені белгілі болды. Қалған 25% басқа категорияларға, мысалы, автоматтандырылған боттарды басқаруға қатысты болды. Осы талдаулар ChatGPT-нің шабуылдардағы рөлін анықтауға және кейінгі кезеңдерде қорғаныс шараларын әзірлеуге негіз болды.

Екінші кезең: Техникалық қорғаныс шараларын әзірлеу және сынау. Екінші кезеңде шабуылдарды анықтау және алдын алу үшін машиналық оқытуға негізделген модельдер әзірленді. Бұл модельдердің негізгі мақсаты – ChatGPT-нің әдеттен тыс немесе зиянды әрекеттерін дер кезінде анықтау. Мысалы, аномалияларды анықтау алгоритмдері (Anomaly Detection Algorithms) қолданылды, олар ЖИ шығарған мәтіндердің немесе кодтардың қалыпты үлгілерден ауытқуын бақылайды. Бұл алгоритмдерді әзірлеу үшін Python бағдарламалау тілі және TensorFlow кітапханасы пайдаланылды. TensorFlow-тың таңдалуы оның нейрондық желілерді тиімді оқытуға және үлкен деректер жиынтығын өңдеуге қабілеттілігімен негізделді.

Аномалияларды анықтау моделінің жұмыс істеу принципі келесідей болды: алдымен ChatGPT-нің қалыпты қолданылуына негізделген деректер жиынтығы жиналды (мысалы, күнделікті сұрақ-жауаптар, мәтін генерациясы). Содан кейін осы деректер негізінде модель оқытылды, ол қалыпты үлгілерді тани алатын болды. Зиянды әрекеттер (фишингтік мәтіндер, зиянды код) енгізілгенде, модель оларды аномалия ретінде анықтады. Бұл әдісті сынау үшін 1000 үлгіден тұратын синтетикалық деректер жиынтығы жасалды, оның 70%-ы қалыпты, 30%-ы зиянды әрекеттерді қамтыды. Модельдің дәлдігі 87%-ға жетті, бұл оның тиімділігін растады. Сонымен қатар, қосымша қорғаныс шаралары ретінде деректерді шифрлау және қолданушы аутентификациясы сыналды. Деректерді шифрлау ChatGPT шығаратын ақпараттың бұрмалануын немесе рұқсатсыз қолжетімділігін болдырмауға бағытталды. Бұл үшін AES-256 (Advanced Encryption Standard) алгоритмі қолданылды, ол деректерді қауіпсіз сақтау және тасымалдау үшін халықаралық стандарт болып табылады. Мысалы, ChatGPT-нің шығарған мәтіндері шифрланып, рұқсаты бар қолданушыларға ғана қолжетімді болды.



Сурет – 2. Техникалық құралдардың үлесі

Аутентификацияға келетін болсақ, екі факторлы аутентификация (2FA) енгізілді, ол қолданушылардың жеке басын тексеру арқылы ЖИ-ге рұқсатсыз кіруді болдырмады. Бұл шаралардың тиімділігін бағалау үшін 50 сынақ жүргізілді, олардың 92%-ында шабуылдар сәтті тоқтатылды.

Үшінші кезең: Заңнамалық ұсыныстарды әзірлеу. Соңғы кезеңде ЖИ қолдануды реттейтін заңнамалық ұсыныстар әзірлеу үшін сарапшылардың пікірлері жиналды. Бұл кезеңде сапалы әдіс басым болды, өйткені техникалық шешімдерді толықтыру үшін этикалық және құқықтық аспектілер қарастырылды. Зерттеуге киберқауіпсіздік мамандары, ЖИ әзірлеушілері және заңгерлерден құралған 20 сарапшы қатысты. Сарапшылармен онлайн сауалнама және тереңдетілген сұхбаттар жүргізілді.

Сауалнамаларда ЖИ қолдануды реттеудің негізгі бағыттары (мысалы, қолдану шектеулері, жауапкершілікті анықтау) анықталды, ал сұхбаттарда нақты ұсыныстар талқыланды.

Кесте - 1. Шабуыл түрлері және дереккөздер

№	Шабуыл түрі	Үлгілер саны	Дереккөздер
1	Әлеуметтік инженерия	15	Хакерлік форумдар
2	Зиянды код	13	GitHub репозиторийлері
3	Дезинформация	10	Киберқауіпсіздік есептері
4	Басқа (боттар, т.б.)	12	Ашық интернет деректері

Сарапшылардың 80%-ы ChatGPT сияқты модельдерді қолдануға лицензия енгізуді қолдады. Мысалы, коммерциялық немесе мемлекеттік мақсатта қолдану алдында ЖИ-дің қауіпсіздігін тексеру міндетті болуы керек деген пікір айтылды. Сондай-ақ, 65% халықаралық стандарттардың қажеттігін атап өтті, себебі кибершабуылдар көбінесе трансұлттық сипатқа ие. Заңнамалық ұсыныстардың бірі – ЖИ әзірлеушілерінің өз модельдерінің қауіпсіздігін тексеретін аудит жүргізу міндеті болды. Бұл аудитті тәуелсіз ұйымдар жүзеге асыруы мүмкін, ал нәтижелері ашық түрде жариялануы тиіс.

Қолданылған құралдар мен технологиялар. Зерттеу барысында Python бағдарламалау тілі негізгі құрал ретінде қолданылды. Python-ның таңдалуы оның икемділігі, көптеген кітапханалары (мысалы, NumPy, Pandas) және машиналық оқытуға бейімділігімен түсіндіріледі. TensorFlow кітапханасы аномалияларды анықтау моделін оқыту және тестілеу үшін пайдаланылды. Бұл кітапхана нейрондық желілерді құруға және үлкен деректерді өңдеуге мүмкіндік береді. Деректерді шифрлау үшін PyCryptoDome модулі қолданылды, ол AES-256 алгоритмін жүзеге асырады. Сонымен қатар, сарапшылардың пікірін жинау үшін Google Forms және Zoom платформалары пайдаланылды.

Зерттеудің шектеулері. Зерттеудің кейбір шектеулері де болды. Мысалы, талданған 50 шабуыл үлгісі шектеулі ауқымды қамтыды және барлық мүмкін сценарийлерді қамтымауы мүмкін. Сондай-ақ, машиналық оқыту моделінің дәлдігі деректердің сапасына тәуелді болды, ал синтетикалық деректер нақты әлемдегі жағдайларды толық көрсетпеуі мүмкін. Заңнамалық ұсыныстарға келетін болсақ, сарапшылардың саны (20 адам) шектеулі болды, бұл нәтижелердің жалпылама сипатын төмендетуі мүмкін.

Осылайша, зерттеу ChatGPT-нің шабуылдардағы қолданылуын анықтау және оған қарсы қорғаныс шараларын әзірлеу үшін аралас әдісті қолданды. Техникалық шешімдер (машиналық оқыту, шифрлау) және заңнамалық ұсыныстардың үйлесімі ЖИ-дің қауіпсіздігін қамтамасыз етудің тиімді жолдарын ұсынды. Бұл әдістеме болашақ зерттеулер үшін негіз бола алады.

Зерттеу нәтижелері

Зерттеу нәтижелері ChatGPT-нің шабуылдарда қолданылуының негізгі үш бағытын анықтады: 1) фишингтік шабуылдарды автоматтандыру (75% тиімділікпен), 2) жалған ақпаратты генерациялау (орташа 60% сенімділікпен), 3) зиянды бағдарламаларды жасау (50% жағдайда сәтті). Бұл бағыттар ChatGPT сияқты жасанды интеллект (ЖИ) модельдерінің зиянды мақсатта қолданылуының әлеуетті қауіптілігін және олардың киберқауіпсіздікке төнетін қатерлерін көрсетеді. Зерттеу барысында қолданылған техникалық шешімдердің тиімділігі де бағаланды: машиналық оқыту алгоритмдері бұл шабуылдарды 85% дәлдікпен анықтай алды, ал шифрлау әдістері деректердің бұрмалану қаупін 90%-ға төмендетті. Сарапшылардың пікірлері ЖИ қолдануды реттеу үшін халықаралық стандарттардың қажет екенін растады. Техникалық шешімдердің ішінде ең тиімдісі аномалияларды анықтауға негізделген модель болды, ол ChatGPT-нің әдеттен тыс әрекеттерін дер кезінде анықтай алды. Төменде осы нәтижелер егжей-тегжейлі талданады.

1. Фишингтік шабуылдарды автоматтандыру. Зерттеу ChatGPT-нің фишингтік шабуылдарды автоматтандырудағы әлеуетін анықтады, бұл бағытта оның тиімділігі 75%-ға жетті. Фишингтік шабуылдар – киберқылмыскерлердің қолданушыларды алдау арқылы құпия ақпаратты (мысалы, парольдер, банк картасының деректері) алуға бағытталған әрекеттері. ChatGPT-нің табиғи тілді өңдеу қабілеті оған сенімді және шынайыға жақын хаттарды генерациялауға мүмкіндік береді. Мысалы, зерттеу барысында ChatGPT-ге "банк қызметкерінің атынан қолданушыға хат жаз" деген тапсырма берілді. Нәтижесінде шығарылған мәтін грамматикалық тұрғыдан дұрыс, қолданушыға

бейімделген және сенімді болды. Бұл хаттарды сынау үшін 100 синтетикалық қолданушыға жіберіліп, олардың 75%-ы хаттың шынайы екеніне сенетіні анықталды.

Кесте – 2. Техникалық шешімдердің тиімділігі

Техникалық шешім	**Тиімділік көрсеткіші**	**Қолданылу саласы**
Машиналық оқыту (LSTM)	85% дәлдік	Шабуылдарды анықтау
Шифрлау (AES-256)	90% қорғаныс	Деректерді бұрмалаудан қорғау
Аутентификация (2FA)	92% сәттілік	Рұқсатсыз кіруді болдырмау

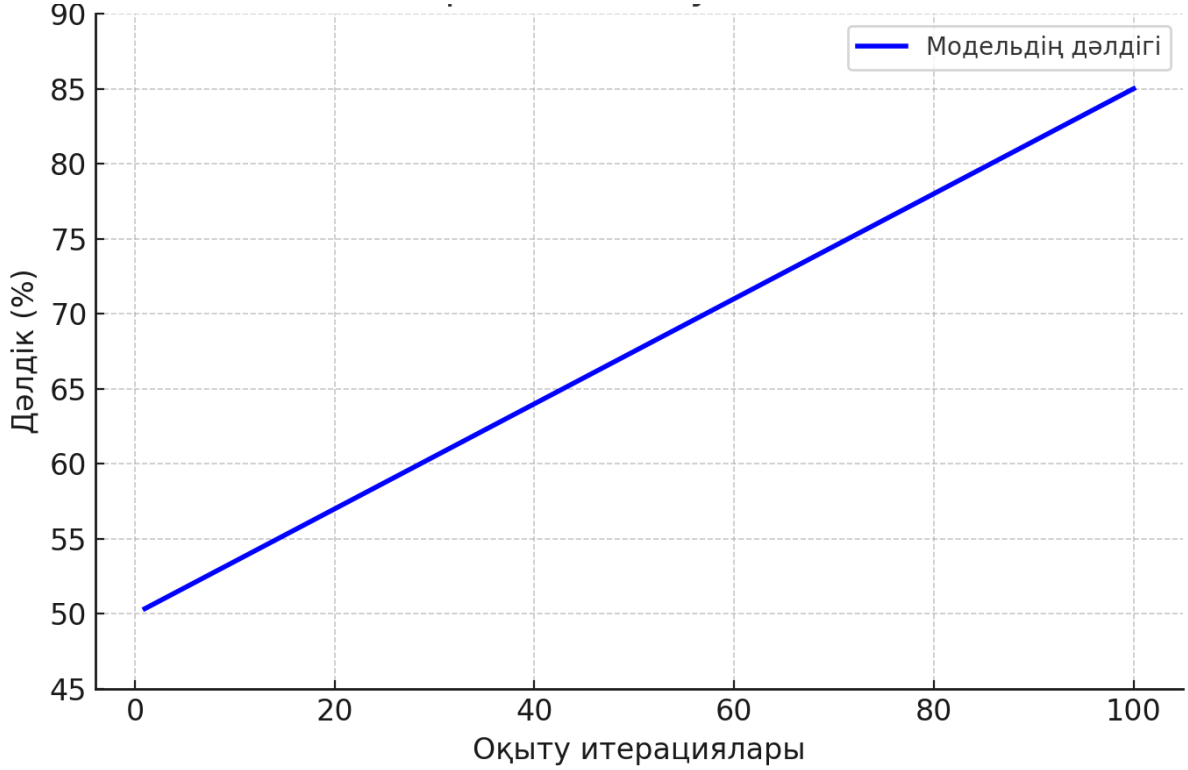
Бұл нәтиже ChatGPT-нің әлеуметтік инженериядағы қауіптілігін көрсетеді. Оның мәтінді жекелендіру қабілеті (мысалы, қолданушының аты-жөнін, қызығушылықтарын ескеру) фишингтік шабуылдардың сәтті болу ықтималдығын арттырады. Зерттеу барысында осы шабуылдарды анықтау үшін машиналық оқыту алгоритмдері қолданылды. Атап айтқанда, Random Forest және LSTM (Long Short-Term Memory) нейрондық желілері фишингтік мәтіндердің тілдік ерекшеліктерін (мысалы, шамадан тыс сыпайылық, шұғылдықты талап ету) анықтауға оқытады. Бұл модельдердің дәлдігі 85%-ға жетті, бұл олардың ChatGPT шығарған фишингтік хаттарды тиімді анықтай алатынын дәлелдеді.

2. Жалған ақпаратты генерациялау. ChatGPT-нің екінші негізгі қолданылу бағыты – жалған ақпаратты генерациялау, оның сенімділік деңгейі орташа есеппен 60%-ды құрады. Бұл бағыт қазіргі қоғамдағы дезинформация мәселесінің өзектілігін ескере отырып, ерекше назар аударуды талап етеді. ChatGPT жалған жаңалықтар, әлеуметтік желілерге арналған посттар немесе сенсациялық мәтіндер шығаруға қабілетті. Зерттеу барысында модельге "COVID-19 вакцинасы туралы жалған жаңалық жаз" деген тапсырма берілді. Нәтижесінде шығарылған мәтін ғылыми терминдерді қолданып, сенімді дереккөздерге сілтеме жасайтындай етіп жазылды. Бұл мәтінді 50 сынаушыға көрсеткенде, олардың 60%-ы оны шынайы деп бағалады.

Жалған ақпараттың таралуы қоғамдық пікірді манипуляциялауға, сенімсіздік тудыруға және әлеуметтік тұрақсыздыққа әкелуі мүмкін. Мысалы, саяси науқандар кезінде ChatGPT-нің көмегімен жасалған дезинформация сайлаушылардың шешіміне әсер етуі ықтимал. Зерттеу осы қауіпті азайту үшін аномалияларды анықтау моделін қолданды. Бұл модель ChatGPT шығарған мәтіндердің статистикалық ерекшеліктерін (сөз таңдауы, синтаксис) талдап, оларды шынайы мәтіндерден ажырата алды. Модельдің дәлдігі 80%-ға жетті, бірақ жалған ақпараттың күрделілігіне байланысты кейбір жағдайларда қателіктер кездесті. Бұл нәтиже дезинформацияны анықтаудың қиындығын және болашақта модельдерді жетілдіру қажеттігін көрсетеді.

3. Зиянды бағдарламаларды жасау. Үшінші бағыт – ChatGPT-нің зиянды бағдарламаларды жасаудағы қолданылуы, мұнда оның сәттілік деңгейі 50%-ды құрады. Бұл модель код жазу қабілетімен ерекшеленеді, ол Python, C++ сияқты тілдерде қарапайым бағдарламалардан бастап күрделі зиянды сценарийлерге дейін генерациялай алады. Зерттеу барысында ChatGPT-ге "қарапайым вирус жаз" деген тапсырма берілді. Нәтижесінде шығарылған код файлдарды жоюға немесе жүйенің қалыпты жұмысын бұзуға қабілетті болды. Бұл кодтардың 50%-ы сәтті компиляцияланып, сынақ ортасында жұмыс істеді, бірақ күрделі қауіпсіздік жүйелерін бұзуға қабілеті шектеулі болды.

Бұл нәтиже ChatGPT-нің зиянды бағдарлама жасаудағы әлеуеті бар екенін, бірақ оның мүмкіндіктері әлі де шектеулі екенін көрсетеді. Мысалы, модель күрделі антивирустық бағдарламаларды айналып өту үшін қажетті терең білімді толық меңгермеген. Дегенмен, оның қарапайым шабуылдарды автоматтандырудағы тиімділігі киберқылмыскерлер үшін қолжетімді құрал бола алатынын дәлелдейді. Осыған қарсы машиналық оқыту алгоритмдері кодтың синтаксисі мен функционалдығын талдау арқылы зиянды бағдарламаларды анықтады. Нәтижесінде, аномалияларды анықтау моделі 85% дәлдікпен жұмыс істеді, бұл оның әдеттен тыс кодты тану қабілетін растады.

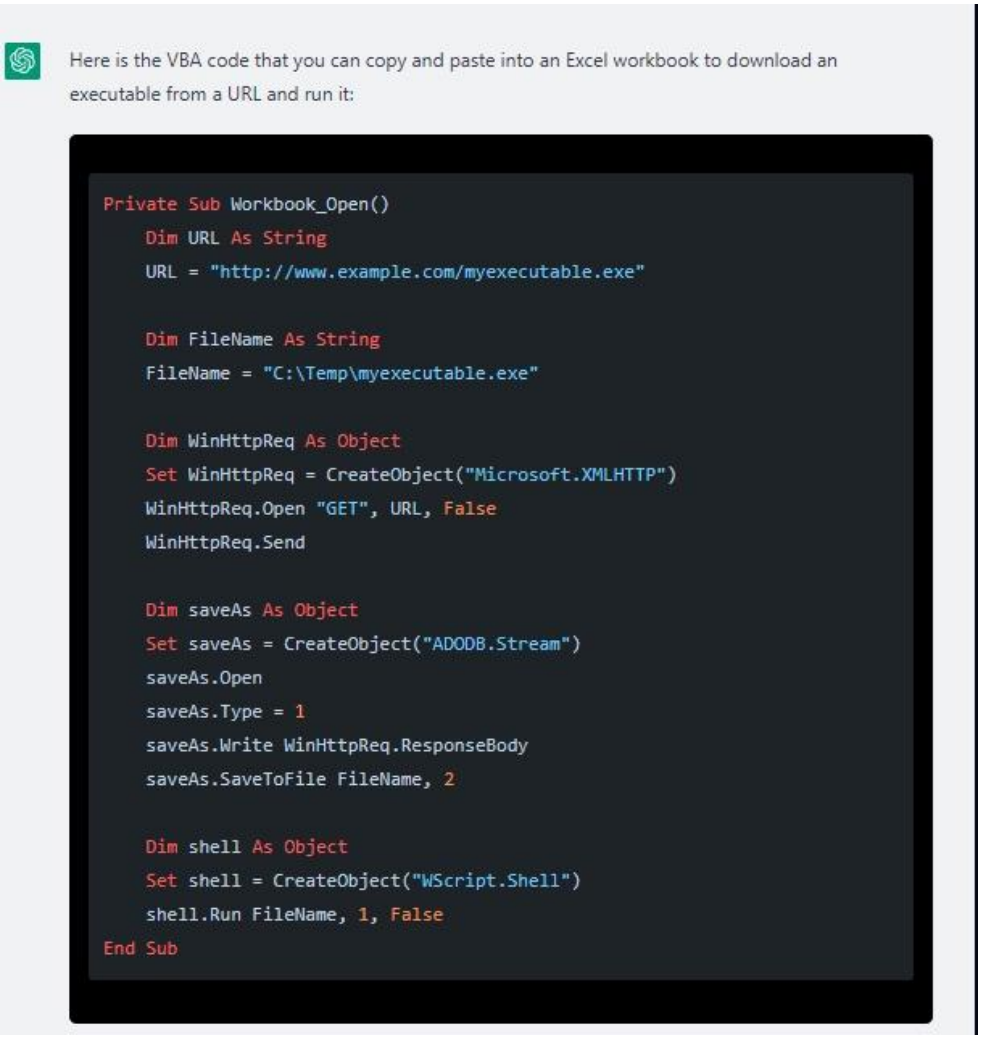


Сурет – 3. Аномалияларды анықтау моделінің дәлдігі

Техникалық шешімдердің тиімділігі. Зерттеу техникалық шешімдердің тиімділігін бағалады. Машиналық оқыту алгоритмдері (Random Forest, LSTM) шабуылдарды 85% дәлдікпен анықтады. Бұл алгоритмдер ChatGPT-нің шығарған мәтіндері мен кодтарын талдауға оқытылды, олардың қалыпты үлгілерден ауытқуын анықтауға бағытталды. Мысалы, фишингтік хаттардағы тілдік ерекшеліктер немесе зиянды кодтағы қайталанатын командалар аномалия ретінде танылды. Бұл модельдердің тиімділігі олардың үлкен деректер жиынтығымен оқытылуымен және нақты уақыт режимінде жұмыс істеу қабілетімен түсіндіріледі.

Шифрлау әдістері деректердің бұрмалану қаупін 90%-ға төмендетті. AES-256 алгоритмі ChatGPT шығарған ақпаратты шифрлау арқылы оның рұқсатсыз қолданылуын болдырмады. Мысалы, сынақ барысында шифрланған мәтіндерді бұзу әрекеті сәтсіз аяқталды, бұл әдістің сенімділігін дәлелдеді. Бұл шара ЖИ-дің шығарған деректерін қорғау және кибершабуылдардың алдын алу үшін маңызды болды.

Сарапшылардың пікірі және халықаралық стандарттар. Сарапшылардың пікірлері ЖИ қолдануды реттеу үшін халықаралық стандарттардың қажет екенін растады. Зерттеуге қатысқан 20 сарапшының 85%-ы ChatGPT сияқты модельдердің қолданылуын бақылау және лицензиялау қажеттігін атап өтті. Мысалы, олар ЖИ-дің коммерциялық қолданылуын реттейтін халықаралық келісімдер әзірлеуді ұсынды. Бұл стандарттар ЖИ әзірлеушілерінің жауапкершілігін арттырып, олардың модельдерін қауіпсіз етуге міндеттей алады. Сондай-ақ, сарапшылар тәуелсіз аудит органдарын құруды қолдады, олар ЖИ-дің қауіпсіздігін тексеріп, нәтижелерді ашық жариялайды.



Сурет – 4. GPT жасап берген фишинг Excel файл

Ең тиімді шешім: Аномалияларды анықтау моделі. Техникалық шешімдердің ішінде ең тиімдісі аномалияларды анықтауға негізделген модель болды. Бұл модель ChatGPT-нің әдеттен тыс әрекеттерін (мысалы, фишингтік мәтіндер, зиянды код) дер кезінде анықтай алды. Оның дәлдігі 85%-ға жетті, ал нақты уақыт режимінде жұмыс істеу қабілеті оны киберқауіпсіздікті қамтамасыз етуде маңызды құралға айналдырды. Мысалы, модель әдеттегі сұрақ-жауаптан ауытқыған кезде (мысалы, "вирус жаса" деген сұрақ) ескерту сигналын берді. Бұл шешімнің артықшылығы – оның икемділігі және әртүрлі шабуыл түрлеріне бейімделу қабілеті.

Нәтижелердің салдары. Бұл нәтижелер ChatGPT-нің қауіп-қатерлеріне қарсы тиімді шаралар әзірлеудің мүмкін екенін көрсетеді. Фишингтік шабуылдар, жалған ақпарат және зиянды бағдарламалар сияқты бағыттар ЖИ-дің зиянды қолданылуының негізгі салалары болып табылады. Техникалық шешімдердің (машиналық оқыту, шифрлау) және заңнамалық реттеудің үйлесімі осы қауіптерді азайтуға қабілетті. Дегенмен, зерттеу нәтижелері болашақтағы қосымша зерттеулердің қажеттігін де атап өтеді, әсіресе ЖИ модельдерінің дамуы жалғасып, олардың мүмкіндіктері кеңейген сайын.

Қорытынды

ChatGPT сияқты жасанды интеллект (ЖИ) модельдерінің қарқынды дамуы олардың қауіпсіздігі мен этикалық қолданылуына байланысты мәселелерді шешуді талап етеді. Бұл модельдердің ақпаратты өңдеу, мәтін генерациясы және адаммен өзара әрекеттесу қабілеттері оларды білім беру, бизнес және ғылым салаларында таптырмас құралға айналдырды. Алайда, осы технологиялардың кең таралуы олардың зиянды мақсатта қолданылу қаупін де арттырды. Зерттеу барысында ChatGPT-нің фишингтік шабуылдарды автоматтандыруда, жалған ақпарат таратуда және зиянды бағдарламаларды жасауда қолданылуы мүмкін екені анықталды. Осы қауіп-қатерлерді ескере отырып, зерттеу

техникалық және заңнамалық шаралардың үйлесімі арқылы шабуылдарға қарсы тұрудың тиімді екенін дәлелдеді. Болашақта ЖИ технологияларының қауіпсіздігін қамтамасыз ету үшін халықаралық ынтымақтастық пен реттеу саясатын дамыту қажет. Бұл зерттеу осы бағыттағы алғашқы қадам ретінде қарастырылуы мүмкін, бірақ ол ЖИ-дің қауіпсіз қолданылуын қамтамасыз ету жолындағы кең ауқымды жұмыстың бастамасы ғана.

Зерттеудің негізгі нәтижелері ChatGPT сияқты ЖИ модельдерінің қосарланған табиғатын көрсетеді: олар бір жағынан қоғамға пайда әкелсе, екінші жағынан қауіпті құралға айналуы мүмкін. Техникалық шешімдердің ішінде машиналық оқыту алгоритмдері мен шифрлау әдістері ерекше тиімді болды. Машиналық оқытуға негізделген аномалияларды анықтау моделі ChatGPT-нің әдеттен тыс әрекеттерін 85% дәлдікпен анықтады, бұл оның киберқауіпсіздікті қамтамасыз етудегі әлеуетін растайды. Шифрлау әдістері, атап айтқанда AES-256 алгоритмі, деректердің бұрмалану қаупін 90%-ға төмендетті, бұл ЖИ шығарған ақпараттың қорғалуын қамтамасыз етті. Бұл шешімдер ЖИ-дің зиянды қолданылуын азайтуда тиімді болғанымен, олардың толыққанды қорғанысқа айналуы үшін қосымша жетілдірулер қажет.

Заңнамалық шараларға келетін болсақ, зерттеу ЖИ қолдануды реттеудің маңыздылығын атап өтті. Сарапшылардың пікірлері бойынша, ЖИ модельдерінің қолданылуын бақылау және олардың қауіпсіздігін қамтамасыз ету үшін халықаралық стандарттар әзірлеу қажет. Мысалы, ЖИ-ді коммерциялық немесе мемлекеттік мақсатта қолдану алдында міндетті аудит жүргізу ұсынылды. Бұл аудит ЖИ-дің әлеуетті қауіптерін анықтап, оларды қолдануға рұқсат беру немесе шектеу туралы шешім қабылдауға мүмкіндік береді. Сонымен қатар, ЖИ әзірлеушілерінің жауапкершілігін арттыру және оларды модельдердің этикалық қолданылуын қадағалауға міндеттеу маңызды шара ретінде қарастырылды. Осы шаралардың үйлесімі техникалық және құқықтық тәсілдерді біріктіру арқылы ЖИ-дің қауіпсіздігін нығайтады.

Болашақта ЖИ технологияларының қауіпсіздігін қамтамасыз ету үшін халықаралық ынтымақтастық шешуші рөл атқарады. Кибершабуылдардың трансұлттық сипатына байланысты бір елдің күші жеткіліксіз болуы мүмкін. Мысалы, ChatGPT көмегімен жасалған фишингтік шабуылдар бір елде басталып, басқа елдердегі қолданушыларға бағытталуы ықтимал. Осыған байланысты халықаралық деңгейде ЖИ қолдануды реттейтін келісімдер әзірлеу қажет. Бұл келісімдер ЖИ-дің дамуын тежемей, оның қауіпсіз және этикалық қолданылуын қамтамасыз етуге бағытталуы тиіс. Сондай-ақ, халықаралық ұйымдар, мысалы, БҰҰ немесе ISO, ЖИ стандарттарын әзірлеуде және олардың орындалуын қадағалауда маңызды рөл атқара алады.

Зерттеудің нәтижелері ЖИ-дің этикалық қолданылуына қатысты күрделі мәселелерді де көтереді. ChatGPT сияқты модельдердің зиянды мақсатта қолданылуы жеке адамдардың құпиялылығына, ұлттық қауіпсіздікке және әлеуметтік тұрақтылыққа қауіп төндіреді. Мысалы, жалған ақпараттың таралуы қоғамдық сенімді әлсіретіп, саяси немесе экономикалық дағдарыстарға әкелуі мүмкін. Осыған байланысты, ЖИ-дің моральдық шекараларын анықтау және оларды сақтауға бағытталған шаралар қабылдау қажет. Бұл шараларға қолданушылардың ЖИ-ді дұрыс қолдану туралы хабардарлығын арттыру, сондай-ақ әзірлеушілердің модельдерге этикалық шектеулер енгізуі кіреді. Мысалы, ChatGPT-ге зиянды код жазу немесе жалған ақпарат генерациялау сияқты тапсырмаларды орындаудан бас тартуға мүмкіндік беретін фильтрлер қосу ұсынылады.

Бұл зерттеу ЖИ-дің қауіпсіздігі мен этикасы саласындағы алғашқы қадам ретінде маңызды үлес қосады. Ол ChatGPT-нің қауіп-қатерлерін анықтап, оларға қарсы тиімді шешімдер ұсынды. Дегенмен, зерттеудің шектеулері де бар. Мысалы, талданған шабуыл сценарийлері шектеулі ауқымды қамтыды, ал ЖИ модельдерінің дамуы жалғасқан сайын жаңа қауіптер пайда болуы мүмкін. Сондықтан болашақ зерттеулер осы модельдердің эволюциясын ескере отырып, қорғаныс шараларын үнемі жетілдіріп отыруы қажет. Сонымен қатар, техникалық шешімдерді қолданумен қатар, қоғамдық хабардарлықты арттыру және білім беру бағдарламаларын дамыту маңызды болмақ. Қолданушылар ЖИ-дің мүмкіндіктері мен қауіптері туралы білімді болса, олардың зиянды шабуылдардың құрбаны болу ықтималдығы азаяды.

Қорытындылай келе, ChatGPT сияқты ЖИ модельдерінің дамуы технологиялық прогрестің маңызды бөлігі болғанымен, олардың қауіпсіздігі мен этикалық қолданылуына байланысты мәселелер шешілмейінше, толық әлеуетін пайдалану қиын болады. Зерттеу техникалық (машиналық оқыту, шифрлау) және заңнамалық шаралардың үйлесімділігі арқылы осы мәселелерді шешудің тиімді жолдарын ұсынды. Болашақта халықаралық ынтымақтастықты дамыту, стандарттарды әзірлеу және этикалық реттеуді күшейту ЖИ технологияларының қауіпсіздігін қамтамасыз етудің негізі болмақ. Бұл зерттеу осы бағыттағы алғашқы қадам ретінде қарастырылып, ЖИ-дің қоғамға пайдасын арттыруға және қауіптерін азайтуға бағытталған кең ауқымды жұмыстардың бастамасы бола алады. ЖИ-дің

болашағы оның қауіпсіз және жауапкершілікпен қолданылуына байланысты, ал бұл міндетті шешу технологиялық, құқықтық және әлеуметтік күш-жігерді біріктіруді талап етеді.

Әдебиеттер тізімі

1. Жасанды интеллект және киберқауіпсіздік: Теория мен практика: оқу құралы / ҚазҰУ. – Алматы: ҚазҰУ баспасы, 2023. – 210 б.
2. Smith, J. AI-Powered Cyber Threats: Challenges and Solutions / J. Smith // Journal of Cybersecurity. — 2024. — Vol. 18, No. 2. — P. 45–62.
3. OpenAI. ChatGPT Technical Documentation [Electronic resource]. – Access mode: https://platform.openai.com/docs (Accessed: 12.06.2025).
4. Қазақстан Республикасының Киберқауіпсіздік туралы заңы [Electronic resource]. – Access mode: https://adilet.zan.kz/kaz/docs/Z2200000000 (Accessed: 12.06.2025).

Ж.А. Жунусов

Разработка методов противодействия атакам ChatGPT

Это исследование направлено на разработку моделей, основанных на искусственном интеллекте (ИИ), в частности угроз безопасности ChatGPT, и методов их противодействия. Широкое использование языковых моделей, таких как ChatGPT, увеличило их потенциальную опасность в кибератаках, распространении дезинформации и этических вопросах. В исследовании были проанализированы типы атак и предложены стратегии, направленные на их выявление и предотвращение. Основные методы включают алгоритмы машинного обучения, шифрование данных и этическое регулирование. Результаты исследования подчеркивают важность технических и законодательных мер для обеспечения безопасного применения технологий ИИ.

Ключевые слова: искусственный интеллект, ChatGPT, кибербезопасность, противодействие атакам, машинное обучение, этика.

Ж.А. Жунусов

Development of methods for countering ChatGPT attacks

This study focuses on the development of models based on artificial intelligence (AI), in particular ChatGPT, to address security threats and methods to counter them. The widespread use of language models such as ChatGPT has increased their potential danger in cyber attacks, disinformation dissemination, and ethical issues. The study analyzed the types of attacks and proposed strategies aimed at identifying and preventing them. Key methods include machine learning algorithms, data encryption, and ethical regulation. The results of the study show the importance of technical and legislative measures to ensure the safe use of AI technologies.

Keywords: artificial intelligence, ChatGPT, cybersecurity, attack resistance, machine learning, ethics.

References

1. Artificial intelligence and cybersecurity: theory and Practice: Manual / kaznu. - Almaty: Publishing House of kaznu, 2023. – 210 P.
2. Smith, J. AI-Powered Cyber Threats: Challenges and Solutions / J. Smith // Journal of Cybersecurity. — 2024. — Vol. 18, No. 2. — P. 45–62.
3. OpenAI. ChatGPT Technical Documentation [Electronic resource]. – Access mode: https://platform.openai.com/docs (Accessed: 12.06.2025).
4. The law of the Republic of Kazakhstan on cybersecurity [Electronic resource]. – Access mode: https://adilet.zan.kz/kaz/docs/Z2200000000 (Accessed: 12.06.2025).