დევიდ ვაინკაუფი*

სამუშაო დოკუმენტი დიდი ენობრივი მოდელების (“LLMs”) შესახებ**

დიდი ენობრივი მოდელები (“LLMs”) წარმოადგენს განსაკუთრებით მასშტაბურ და კომპლექსურ მანქანური სწავლების სისტემებს, რომლებსაც შეუძლია მომხმარებლის შეკითხვებზე პასუხად შექმნან მკაფიო და დამაჯერებლად ჟღერადი — თუმცა არა აუცილებლად ჭეშმარიტი — ტექსტი. წინამდებარე ნაშრომი გთავაზობთ “LLM”-ების სიღრმისეულ და კომპლექსურ ანალიზს, რომლის მიზანია მონაცემთა დაცვის საზედამხედველო ორგანოებს დაეხმაროს ამ ახალი ტექნოლოგიით გამოწვეული სირთულეების რეგულირებასა და მათზე ეფექტიან რეაგირებაში. ნაშრომში “LLM” გაანალიზდა სამი პერსპექტივიდან: (1) ტექნოლოგიური პერსპექტივა — ის, თუ როგორ ფუნქციონირებს და როგორ ვითარდება “LLM”; (2) “LLM”-ის განვითარებასთან დაკავშირებული რისკები მონაცემთა დაცვის სფეროში; და (3) აღნიშნული რისკების შესამცირებლად ან აღმოსაფხვრელად ჩამოყალიბებული საუკეთესო პრაქტიკის მაგალითები.

***საკვანძო სიტყვები**: დიდი ენობრივი მოდელები, ხელოვნური ინტელექტი, ტექნიკური ანალიზი, პირადი ცხოვრების ხელშეუხებლობა, პერსონალურ მონაცემთა დაცვა, რისკები, საუკეთესო პრაქტიკა, შემამსუბუქებელი ღონისძიებები.*

---

* დოქტორი, ტექნოლოგიური პოლიტიკის უფროსი მრჩეველი, კანადის კონფიდენციალობის კომისრის ოფისი.

** წინამდებარე დოკუმენტი წარმოადგენს „ტექნოლოგიების სფეროში მონაცემთა დაცვის საერთაშორისო სამუშაო ჯგუფის“ (“IWGDPT”) — ე.წ. „ბერლინის ჯგუფის“ — მიერ მომზადებული სამუშაო ნაშრომის რედაქტირებულ და შემოკლებულ ვერსიას. ნაშრომის ორიგინალი ვერსია გამოქვეყნდა 2024 წლის 27 დეკემბერს და ხელმისაწვდომია შემდეგ ბმულზე: https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Berlin-Group/20241206-WP-LLMs.pdf?__blob=publicationFile&v=2 [25.11.2025].

1. შესავალი

დიდი ენობრივი მოდელები ("LLM") წარმოადგენს მათემატიკურ მოდელებს, რომლებიც შემუშავებულია ხელოვნური ინტელექტისა ("AI") და მანქანური სწავლების ("ML") ტექნიკების გამოყენებით, ბუნებრივ ენასთან დაკავშირებული ამოცანების შესასრულებლად. ბოლო წლებში ამ სფეროში მიღწეული ტექნოლოგიური პროგრესი არსებითად დაჩქარდა, რის შედეგადაც ზოგიერთმა "LLM"-მა გარკვეული ამოცანების ფარგლებში ადამიანური და მეტიც — ექსპერტული დონის შესაძლებლობებიც კი გამოავლინა. სფეროში მიღწული პროგრესი უმთავრესად განპირობებულია მოდელების არქიტექტურისა და ტრენინგის ტექნიკების გაუმჯობესებით; აგრეთვე მოდელების ზომის, სასწავლო მონაცემთა კორპუსების მოცულობისა და გამოთვლითი რესურსების ხელმისაწვდომობის სწრაფი ზრდით.

მიუხედავად მათი შესაძლებლობებისა, დიდი ენობრივი მოდელები არ წარმოადგენს ტექნოლოგიურ პანაცეას. "LLM"-ის ხელოვნურ ინტელექტზე დაფუძნებული მიდგომა, რომელიც ენობრივი ამოცანების ავტომატიზაციას ემსახურება, წარმოშობს რიგ რისკებს პირადი ცხოვრების ხელშეუხებლობისა და მონაცემთა დაცვის მიმართულებით. ზოგიერთი რისკი უკავშირდება საბაზისო ტექნოლოგიის დიზაინს, ზოგი კი — პერსონალური მონაცემების დამუშავების პრაქტიკას. ამასთანავე, მნიშვნელოვანი რისკები უკავშირდება ენის აღქმისა და ენობრივი ამოცანების შესრულების ხელოვნურ ინტელექტზე დაფუძნებული მათემატიკური მიდგომების სტრუქტურულ შეზღუდვებს.

წინამდებარე ნაშრომის მიზანია დიდი ენობრივი მოდელების სიღრმისეული ანალიზი პირადი ცხოვრების ხელშეუხებლობისა და მონაცემთა დაცვის პერსპექტივიდან. ვინაიდან "LLM"-ები წარმოადგენს კომპლექსურ ტექნოლოგიებს, რომლებიც პირადი ცხოვრების ხელშეუხებლობისა და მონაცემთა დაცვის კუთხით მრავალფეროვან რისკებს წარმოშობს, მათი ანალიზიც მოითხოვს მრავალმხრივ შეფასებას. ამდენად, აუცილებელია "LLM"-ების გაანალიზება არა მხოლოდ ტექნოლოგიურ ჭრილში, ე.ი. იმის მიხედვით, თუ როგორ ფუნქციონირებენ ისინი ტექნიკურად, არამედ პირადი ცხოვრების ხელშეუხებლობისა და მონაცემთა დაცვის რისკების, ისევე როგორც იმ საუკეთესო პრაქტიკების გათვალისწინებით, რომლებიც მიზნად ისახავს აღნიშნული რისკების შემცირებას ან აღმოფხვრას. მხოლოდ აღნიშნული სამი პერსპექტივის — ტექნოლოგიის, მონაცემთა დაცვის რისკებისა და საუკეთესო პრაქტიკის — ერთობლივი გააზრებით შეუძლიათ მონაცემთა დაცვის საზედამხედველო ორგანოებს ამ ახალი გამოწვევის ეფექტიანი რეგულირება და მათზე ადეკვატური რეაგირება.

ნაშრომი შედგება სამი ნაწილისგან. პირველ ნაწილში წარმოდგენილია დიდი ენობრივი მოდელების ტექნიკური ანალიზი, შესაბამისი აქცენტით "LLM"-ების განვითარების თითოეულ ეტაპზე გამოყენებული სხვადასხვა კომპონენტის როლზე. მეორე ნაწილში განხილულია დიდი ენობრივი მოდელების მიერ წარმოშობილი სხვადასხვა რისკი მონაცემთა დაცვისა და

პირადი ცხოვრების ხელშეუხებლობის კუთხით. მესამე ნაწილში კი განხილულია საუკეთესო პრაქტიკა, რომელიც მიზნად ისახავს "LLM"-ებთან დაკავშირებული გარკვეული რისკების პრევენციას ან შემსუბუქებას, აქცენტით იმ ძირითად საკითხებზე, რომლებიც საჭიროებს გააზრებას როგორც ტექნოლოგიების თავდაპირველი შემუშავების, ისე მათი გამოყენების დროს.

დათქმა

წინამდებარე ნაშრომში არ არის მოცემული სამართლებრივი რჩევები. ამასთან, ნაშრომში გამოთქმული შეხედულებები აუცილებლად არ ასახავს „ტექნოლოგიების სფეროში მონაცემთა დაცვის საერთაშორისო სამუშაო ჯგუფის" ("IWGDPT") ცალკეული წევრების ოფიციალურ პოლიტიკას ან პოზიციას.

2. რა არის დიდი ენობრივი მოდელები ("LLMs")?

დიდი ენობრივი მოდელები წარმოადგენს განსაკუთრებით მასშტაბურ და კომპლექსურ მანქანური სწავლების სისტემებს, რომლებსაც შეუძლიათ მომხმარებლის შეკითხვებზე პასუხად შექმნან მკაფიო და დამაჯერებლად ჟღერადი — თუმცა არა აუცილებლად ჭეშმარიტი — ტექსტი. "LLM"-ები შედგება ასობით მილიარდი, ხოლო ზოგ შემთხვევაში — ტრილიონობით პარამეტრისგან, რომელიც განაწილებულია სხვადასხვა სტრუქტურულ კომპონენტზე. თითოეული ასეთი კომპონენტი ასრულებს კონკრეტულ ფუნქციას და სისტემას ახალ შესაძლებლობებს მატებს. აღნიშნული კომპონენტების მაგალითებია ლექსიკონი, სიტყვების რიცხვით ვექტორებად გარდაქმნა ("embeddings"), მოდელის შინაარსი, ანუ ე.წ. კონტექსტის ფანჯარა, მრავალთავიანი თვითყურადღების მექანიზმი და წინ მიმართული ნერვული ქსელები.

ეს კომპონენტები ერთობლივად ქმნის ე.წ. „ტრანსფორმერის" აგებულებას (არქიტექტურას). ხელოვნური ინტელექტის ("AI") მოდელებს, მათ შორის დიდ ენობრივ მოდელებს, რომელთა სტრუქტურა ეფუძნება აღნიშნულ აგებულებას, ჩვეულებრივ „ტრანსფორმერულ" მოდელებად მოიხსენიებენ. „ტრანსფორმერის" მოდელის დეტალური ტექნიკური ანალიზი გთხოვთ იხილოთ წინამდებარე სტატიის ორიგინალ ვერსიაში.[1]

დიდი ენობრივი მოდელების გაწვრთნის (ტრენინგის) ციკლი მნიშვნელოვნად განსხვავდება მანქანური სწავლების სხვა პროგრამების გაწვრთნის პროცესისგან. კერძოდ, ერთსაფეხურიანი გაწვრთნის მოდელის

[1] იხ.: *IWGDPT*, *სამუშაო დოკუმენტი დიდი ენობრივი მოდელების შესახებ (LLMs),* დეკემბერი 27, 2024, <https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Berlin-Group/20241206-WP-LLMs.pdf?__blob=publicationFile&v=2> [25.11.2025].

ნაცვლად, "LLM"-ების შემთხვევაში, როგორც წესი, იყენებენენ ორ ეტაპიან პროცესს, რომელიც სწავლის რამდენიმე სხვადასხვა მეთოდს აერთიანებს. გაწვრთნის პირველ ეტაპს ეწოდება „წინასწარი წვრთნა" ("pre-training"), ხოლო მეორეს — „დაზუსტება / შესაბამისობაში მოყვანა" ("fine-tuning / alignment").

ნაშრომის შემდგომ ნაწილებში განხილულია დიდი ენობრივი მოდელების განვითარების და გაწვრთნის პროცედურა თითოეული ეტაპის ცალკეული ანალიზის გზით, მათ შორის გამოყენებული სასწავლო მეთოდების აღწერით და სისტემის საერთო ფუნქციონირებაში თითოეული ფაზის წვლილის წარმოჩენით.

მნიშვნელოვანია აღინიშნოს, რომ ნაშრომში განხილული ეტაპები სრულად არ ფარავს "LLM"-ების განვითარების ყველა ფაზას. მაგალითად, ბევრი "LLM" გამოყენებამდე გადის ე.წ. "red-teaming"-ის ეტაპს, რომლის ფარგლებშიც უსაფრთხოების სპეციალისტებისა და სხვა დარგობრივი ექსპერტებისგან შემდგარი ჯგუფი ცდილობს, გამოავლინოს სისტემის სუსტი მხარეები და მისი ბოროტად გამოყენების შესაძლო სცენარები. მიუხედავად ამისა, "LLM"-ების განვითარების ციკლის აღნიშნული ეტაპები იძლევა საშუალებას, განვიხილოთ მათი მრავალი უნიკალური მახასიათებელი, რაც ხელს უწყობს სისტემის საერთო ფუნქციონირების უკეთ გააზრებას.

2.1. 1-ელი ეტაპი: წინასწარი წვრთნა

დიდი ენობრივი მოდელების გაწვრთნის პირველი ეტაპის მიზანია, შეიქმნას ზოგადი დანიშნულების მოდელი, რომელსაც ექნება პირველადი, ჯერ კიდევ დაუმუშავებელი უნარი — კონკრეტული თემის ფარგლებში ტექსტის მიმდევრობაში უწყვეტად იწინასწარმეტყველოს მომდევნო სიტყვა ან სიტყვითი ერთეული ("token"). ამ მიზნის მისაღწევად მოდელი იწვრთნება განსაკუთრებით დიდი მოცულობის ბუნებრივი ენის მონაცემების გამოყენებით, რომლებიც, როგორც წესი, მოიცავს ინტერნეტში ხელმისაწვდომი ვებგვერდებიდან შეგროვებულ ტექსტებს და/ან ელექტრონულ წიგნებს.

წინასწარი წვრთნა ეფუძნება ე.წ. თვითსწავლებაზე ("self-supervised learning") დაფუძნებულ მიდგომას. ეს მეთოდი ჰგავს ზედამხედველობითი სწავლების მეთოდს ("supervised learning"), თუმცა იმ მნიშვნელოვანი განსხვავებით, რომ სწორი პროგნოზის დასადგენად საჭირო „მითითებები" არ ეფუძნება ცალკე, ადამიანის მიერ მომზადებულ „ეტიკეტირებულ" მონაცემებს. ამის ნაცვლად, სწორი პროგნოზისათვის საჭირო ინფორმაცია — ე.წ. "ground truth" — უშუალოდ სასწავლო მონაცემებშია უკვე მოცემული. რადგან ბუნებრივი ენა თავისთავად შეიცავს იმის მაგალითებს, თუ რომელი სიტყვა მოჰყვება წინამორბედს, მოდელს შეუძლია ისწავლოს მომდევნო სიტყვის წინასწარმეტყველება დამატებითი, ადამიანის მიერ შექმნილი ეტიკეტების გარეშე.

წინასწარი წვრთნა მოიცავს ნაბიჯების თანმიმდევრობას, რომლებიც განმეორებით სრულდება იმდენ ხანს, რამდენიც გათვალისწინებულია

სასწავლო პროცესისთვის. სასწავლო ალგორითმი ასრულებს შემდეგ მოქმედებებს:

1. სასწავლო მონაცემებიდან არჩევს ტექსტის ერთ მონაკვეთს;
2. არჩეული ტექსტი (ბოლო სიტყვის გარეშე) შეჰყავს მოდელში, რათა მიიღოს მომდევნო, ე.ი. იმ ბოლო სიტყვის პროგნოზი, რომელიც არ შეიყვანა;
3. გამოთვლის პროგნოზის შეცდომას პროგნოზირებული შედეგისა და რეალურად არსებული ბოლო სიტყვის შედარების საფუძველზე; და
4. შეცდომის შემცირების მიზნით ასწორებს მოდელში არსებული თითოეული პარამეტრის მნიშვნელობას ე.წ. „უკუდიფერენცირების“ (“backpropagation”) მეთოდის გამოყენებით.

წინასწარი წვრთნის დასრულების შედეგად მიღებული მოდელის აღსაწერად გამოიყენება ტერმინი „ფუძემდებლური მოდელი“ (“foundation model”).[2] თუმცა აღნიშნული ტერმინი გარკვეულწილად საკამათოა. ტერმინის შემქმნელებმა განმარტეს, რომ იგი შეირჩა იმ მიზნით, რომ „ასახოს ამ მოდელების დაუსრულებელი, თუმცა მნიშვნელოვანი ნაწილი, ვინაიდან ისინი „წარმოადგენს საერთო საძირკველს, რომლის საფუძველზეც ადაპტაციის გზით შეიქმნება მრავალი განსხვავებული, ამოცანაზე ორიენტირებული მოდელი“.[3] ამ ხედვის საპირისპიროდ, კრიტიკოსები მიიჩნევენ, რომ აღნიშნული ტერმინი არაზუსტად ასახავს ამ მოდელების კავშირს ბუნებრივ ენასთან. ხელოვნური ინტელექტის ერთ-ერთმა მკვლევარმა აღნიშნულთან დაკავშირებით განაცხადა: „ეს მოდელები სინამდვილეში ჰაერში აშენებული სასახლეებია — მათ არავითარი საფუძველი არ აქვთ.“[4]

აკადემიური კვლევის მიღმა შესაძლებელია წინასწარი წვრთნის შედეგის უფრო პრაქტიკული და უბრალო ენით აღწერაც. ერთ-ერთი პრაქტიკოსის შეფასებით, წინასწარი წვრთნის შედეგად მიღებული მოდელი ფუნქციონირებს როგორც სისტემა, რომელიც ინტერნეტში არსებულ ტექსტურ მასალებზე დაყრდნობით ახდენს ტექსტის ავტომატურ გენერირებას, სიღრმისეული ანალიზის გარეშე.[5]

2.2. მე-2 ეტაპი: შესაბამისობაში მოყვანა

ზოგადი დანიშნულების „ფუძემდებლური მოდელის“ შექმნის შემდეგ წვრთნის მომდევნო ეტაპი უკავშირდება მოდელის ქცევის დახვეწას, რათა მისი

[2] იხ.: *Bommasani R., Hudson D., Adeli E. et al.*, On the Opportunities and Risks of Foundation Models, August 2021, <https://doi.org/10.48550/arXiv.2108.07258> [25.11.2025].

[3] იქვე, გვ. 3 (n. 2) და გვ. 7.

[4] იხ. ავტორის ციტატა: *Malik J.* in: *Knight W.*, A Stanford Proposal Over AI's 'Foundations' Ignites Debate, *Wired*, September 2021, <https://www.wired.com/story/stanford-proposal-ai-foundations-ignites-debate/> [25.11.2025]. აღნიშნული მოსაზრების ვიდეოჩანაწერი ხელმისაწვდომია შემდეგ ბმულზე: <https://www.reddit.com/r/MachineLearning/comments/pd4jle/d_jitendra_maliks_take_on_foundation_models_at/> [25.11.2025].

[5] იხ.: *Karpathy A.*, Let’s build GPT: from scratch, in code, spelled out, January 2023, <https://www.youtube.com/watch?v=kCc8FmEb1nY> [25.11.2025], at 1:51:45.

პასუხები უკეთ შეესაბამებოდეს ადამიანის ღირებულებებს. სასურველი ქცევა, როგორც წესი, გამოიხატება სამი ძირითადი კომპონენტით, რომლებიც ცნობილია „სამი H"-ის სახელწოდებით: "LLM"-მა უნდა იმოქმედოს ისე, რომ მისი პასუხები იყოს სასარგებლო ("helpful"), გულწრფელი ("honest") და უსაფრთხო ("harmless").[6]

აღნიშნული ამოცანა მნიშვნელოვან გამოწვევებს უკავშირდება. წინასწარი წვრთნის შემდეგ მოდელს შეუძლია მხოლოდ ტექსტის მიმდევრობაში მომდევნო სიტყვის პროგნოზირება. მიუხედავად იმისა, რომ ეს უნარი გარკვეულწილად უკავშირდება წვრთნის მეორე ეტაპზე ფორმირებულ შესაძლებლობებს, იგი არსებითად განსხვავდება მათგან. კერძოდ, მომდევნო სიტყვის პროგნოზირება არ უტოლდება ადამიანის მიერ გაცემული მითითებების შესრულებას, ცრუ ინფორმაციის თავიდან აცილებას ან მომხმარებლისთვის ტოქსიკური და მავნე შინაარსის მიწოდებისგან შეზღუდვის უნარებს, რომლებიც "LLM"-მა უნდა განივითაროს წვრთნის მეორე ეტაპზე. მაგალითად, როდესაც მომხმარებლის მიერ გაცემულ მითითებას დავალების სახით ვთავაზობთ "OpenAI"-ის მესამე თაობის გენერაციულ მოდელს — "GPT-3"-ს, რომელსაც გავლილი აქვს მხოლოდ წვრთნის პირველი ეტაპი (წინასწარი წვრთნა), იგი, როგორც წესი, უბრალოდ აგრძელებს მსგავსი მითითებების ჩამონათვალს. იმავე მოდელის "fine-tuned" ვერსია — "InstructGPT" — ანუ მოდელი, რომელსაც გავლილი აქვს წვრთნის მეორე ეტაპი და რომელიც წარმოადგენს "OpenAI"-ის პოპულარული "ChatGPT"-ის წინამორბედს, იმავე დავალების პასუხად უკვე იძლევა პირდაპირ და შინაარსობრივად გამართულ პასუხს.

საილუსტრაციოდ იხილეთ ცხრილი 1.

მითითება:	**გთხოვთ, აუხსენით მთვარეზე დაჯდომის (პილოტირებული მისია მთვარისაკენ) ფაქტი 6 წლის ბავშვს რამდენიმე წინადადებით.**
პასუხები:	"GPT-3"-ის პასუხი: „აუხსენით მიზიდულობის თეორია 6 წლის ბავშვს. აუხსენით თანაფარდობის თეორია 6 წლის ბავშვს რამდენიმე წინადადებით. აუხსენით დიდი აფეთქების თეორია 6 წლის ბავშვს. აუხსენით ევოლუცია 6 წლის ბავშვს."
	"InstructGPT"-ის პასუხი:

[6] იხ.: *Askell A., Bai Y., Chen A. et al.*, A General Language Assistant as a Laboratory for Alignment, 2021, <https://doi.org/10.48550/arXiv.2112.00861> [25.11.2025].

	„ადამიანები გაფრინდნენ მთვარეზე, გადაიღეს ფოტოები იმისა, რაც იქ ნახეს, და ეს ფოტოები დედამიწაზე გამოგზავნეს, რათა ჩვენ ყველას შეგვეძლოს მათი ნახვა.“

ცხრილი 1: *ერთი და იმავე მითითებაზე “GPT-3”-ისა და “InstructGPT”-ის მიერ გაცემული განსხვავებული პასუხების მაგალითი. წყარო: OpenAI, Aligning language models to follow instructions, <https://openai.com/research/instruction-following> [25.11.2025].*

ზოგადად, “fine-tuning”-ის ეტაპზე წვრთნის პროცესი იყოფა ორ საფეხურად (ქვეეტაპად). პირველი ეფუძნება ზედამხედველობით სწავლას (“supervised learning”), ხოლო მეორე — განმტკიცებაზე / წახალისებაზე დაფუძნებულ სწავლებას (“reinforcement learning”).

2.2.1. ზედამხედველობითი სწავლება

დიდი ენობრივი მოდელების გაწვრთნის ქვემოთ განხილული ეტაპი გარკვეულწილად ჰგავს წინასწარი წვრთნის ეტაპს, თუმცა იმ არსებითი განსხვავებით, რომ ის მაგალითები, რომელთა საფუძველზეც მოდელი იწვრთნება, ადამიანის მიერ წინასწარ არის შერჩეული და დამუშავებული. აღნიშნული მაგალითები მოიცავს როგორც იმ ტიპის მითითებებს — ანუ დავალებებს — რომელთა მიღებაც “LLM”-ს ევალება, ისე იმ პასუხებს, რომლებიც მან ასეთ შემთხვევებში უნდა გადასცეს მომხმარებელს. სწორედ ამიტომ ეწოდება აღნიშნულ ეტაპს „ზედამხედველობითი“ სწავლება. სასწავლო მასალა მოიცავს “LLM”-სა და მომხმარებელს შორის ამოცანაზე ორიენტირებული ურთიერთქმედების სრულ მაგალითებს, მათ შორის როგორც მომხმარებლის მიერ გაცემულ მითითებას, ისე შესაბამის, „სწორ“ პასუხს, რომელიც მოდელმა უნდა გასცეს.

როგორც წესი, ამ ეტაპზე გამოყენებული სასწავლო მონაცემების მოცულობა მნიშვნელოვნად ნაკლებია იმ მონაცემებთან შედარებით, რომლებიც გამოიყენება წინასწარი წვრთნის ეტაპზე. ეს განპირობებულია როგორც პრაქტიკული, ისე სამეცნიერო მიზეზებით. პრაქტიკული კუთხით, ზედამხედველობითი სწავლებისთვის სპეციალურად მორგებული სასწავლო მონაცემთა ბაზების შექმნა მნიშვნელოვნად მეტ დროსა და რესურსს მოითხოვს, ვიდრე თვითსწავლებაზე დაფუძნებული სწავლებისთვის ინტერნეტში ხელმისაწვდომი ტექსტების ან ელექტრონული წიგნების მასობრივი გამოყენება, განსაკუთრებით თანამედროვე ციფრული მასალის დიდი მოცულობის გათვალისწინებით. ამასთან, მანქანური სწავლების თვალსაზრისით, ამ ეტაპზე ნაკლები, თუმცა მაღალი ხარისხის მონაცემები რეალურად უფრო ეფექტიანია. კვლევებმა აჩვენა, რომ ზედამხედველობითი

სწავლების ეტაპი ხასიათდება ე.წ. „ნიმუშების ეფექტიანობით“ ("sample efficiency"), რაც ნიშნავს, რომ მოდელის წარმატებით გასაწვრთნელად კონკრეტულ ამოცანაზე — კერძოდ, მომხმარებლის მითითებების რეგულარულ რეჟიმში შესრულებისთვის — საჭირო მონაცემთა მოცულობა შედარებით მცირეა.[7] ამგვარად, პირველადად გაწვრთნილ მოდელზე დაყრდნობით, ზედამხედველობითი სწავლება შესაძლებელს ხდის "LLM"-ის პარამეტრების მიზნობრივად გასწორებასა და გაუმჯობესებას, რის შედეგადაც მოდელის თავდაპირველი, დაუმუშავებელი ენობრივი შესაძლებლობები გარდაიქმნება უფრო პირდაპირ და მიზანმიმართულ ქცევად. როგორც წესი, აღნიშნული ეტაპის დასრულების შემდეგ "LLM" უფრო ეფექტიანად მუშაობს.

2.2.2. წახალისებაზე დაფუძნებული სწავლება

ხელოვნური ინტელექტის მიერ მომხმარებლის მიერ გაცემული დავალების უშუალოდ შესრულება თავისთავად არ ნიშნავს, რომ იგი შესრულდება პასუხისმგებლიანად ან ეთიკურად. მართალია, ზედამხედველობითი სწავლება — ანუ წვრთნის მეორე ეტაპი — საშუალებას აძლევს "LLM"-ს მომხმარებელს გასცეს უფრო ეფექტიანი პასუხები, მაგრამ, როგორც წესი, ამ ეტაპზე მოდელს ჯერ კიდევ არ აქვს ფორმირებული ისეთი „ღირებულებები“, როგორებიცაა გულწრფელობა და უვნებლობა (ზიანის არმიყენება). ამ დამატებით ღირებულებებთან უკეთ შესაბამისობის მისაღწევად "LLM"-ები გადიან წვრთნის შემდგომ ეტაპს, რომელიც ეფუძნება ე.წ. წახალისებაზე დაფუძნებულ სწავლებას ("reinforcement learning").

წახალისებაზე დაფუძნებული სწავლება წარმოადგენს მანქანური სწავლების ისეთ ფორმას, რომლის ფარგლებშიც მოდელი იწვრთნება დინამიკურ გარემოში მიღებულ უკუკავშირზე დაყრდნობით, რაც გარკვეულწილად „ცდისა და შეცდომის“ ("trial and error") პროცესს ჰგავს. ზედამხედველობითი ან თვითსწავლებაზე დაფუძნებული მეთოდებისგან განსხვავებით, ამ შემთხვევაში ხელოვნური ინტელექტის მოდელის გაწვრთნა არ ხდება ე.წ. „სწორი ქცევის“ მაგალითების განმეორებითი ჩვენების გზით. აღნიშნულის ნაცვლად, სწავლების პროცესში ცენტრალურ როლს ასრულებს წახალისებისა და დასჯის მექანიზმი. კერძოდ, მოდელი ჯილდოვდება იმ ქცევისათვის, რომელიც მას განსაზღვრული მიზნის მიღწევაში ან მისკენ დაახლოებაში ეხმარება, ხოლო ისჯება იმ ქმედებებისათვის, რომლებსაც საპირისპირო შედეგამდე მიჰყავს. სხვადასხვა სტრატეგიის გამოცდისა და მიღებული დადებითი ან უარყოფითი უკუკავშირის საფუძველზე, საკუთარი პარამეტრების განახლების გზით, მოდელი ეტაპობრივად აყალიბებს ოპტიმალურ „ქცევით პოლიტიკას“, რომელიც შესაბამის გარემოში მაქსიმალურ წახალისებას უზრუნველყოფს. ამდენად, წახალისებაზე დაფუძნებული სწავლება სხვა ტიპის მანქანურ სწავლებასთან შედარებით უფრო ღია და კვლევითი ხასიათისაა. სწორედ ამისთვის გამოიყენება იგი

[7] იხ.: *Khandelwal U., Clark K., Jurafsky D. et al.*, Sample Efficient Text-Summarization Using a Single Pre-Trained Transformer, 2019, <https://doi.org/10.48550/arXiv.1905.08836> [25.11.2025].

ხშირად სტრატეგიაზე დაფუძნებული ამოცანებისთვის, როგორებიცაა თამაშები, მაგალითად, "Go"[8] ან "StarCraft".[9]

"LLM"-ების შემთხვევაში ის „თამაში", რომლის შესრულებაზეც მოდელი იწვრთნება, გულისხმობს მომხმარებლის მითითებებზე ეთიკურ და სათანადო რეაგირებას. ერთი შეხედვით შეიძლება ჩანდეს, რომ ეს ამოცანა სტრატეგიული თამაშების — როგორიცაა "Go" ან "StarCraft" — სწავლების პროცესის ანალოგიურია; თუმცა უფრო სიღრმისეული ანალიზი ცხადყოფს, რომ ეთიკური გადაწყვეტილებების მიღება არსებითად განსხვავებული ხასიათისაა. სწორედ ეს განსხვავებები აჩენს რიგ სერიოზულ გამოწვევებს "LLM"-ების კონტექსტში წახალისებაზე დაფუძნებული სწავლების გამოყენებისას.

მთავარი სირთულე მდგომარეობს იმაში, რომ სტრატეგიული თამაშებისგან განსხვავებით, რომლებშიც მკაფიოდ არის განსაზღვრული, თუ რას წარმოადგენს „გამარჯვება" — ე.ი. „თამაშის მოგება" (მაგალითად, მაღალი ქულის მიღება ან მოწინააღმდეგის დამარცხება) — ეთიკაში არ არსებობს ზუსტი და ერთმნიშვნელოვანი განმარტება იმისა, თუ რა იქნება „გამარჯვება/მოგება". ეთიკური ქმედება არ ხორციელდება წინასწარ განსაზღვრული საბოლოო მიზნის ან შედეგის მისაღწევად. მას არ აქვს რაიმე გარე მიზანი, რომლისკენაც მიმართულია ქცევა. პირიქით, ეთიკური მოქმედება ღირებულია თავად საკუთარი თავისთვის — უბრალოდ იმიტომ, რომ „სწორად მოქმედება სწორია". როგორც არისტოტელე განმარტავს ამ განსხვავებას, „წარმოების [მაგ., სტრატეგიული თამაშის] მიზანი განსხვავებულია თავად ამ მოქმედებისგან, ხოლო ქმედების [ეთიკური მოქმედების] მიზანი ვერ იქნება სხვა, რადგან სწორედ ეთიკური მოქმედება თავად არის მიზანი."[10]

ამ თავისებურებიდან გამომდინარე, ეთიკური მოქმედების კრიტერიუმები თავისთავად ბუნდოვანი და არაზუსტია. ისინი არ ექვემდებარება იმგვარ სიზუსტეს, რომელიც დამახასიათებელია მათემატიკისთვის ან საბუნებისმეტყველო მეცნიერებებისთვის. ეს ქმნის სერიოზულ პრობლემას წახალისებაზე დაფუძნებული სწავლებისთვის, ვინაიდან ზუსტად განსაზღვრული მიზნის არარსებობის პირობებში რთულია დადგინდეს, უნდა დაჯილდოვდეს თუ დაისაჯოს მოდელის მიერ არჩეული კონკრეტული ქმედება ან სტრატეგია. რადგან ეთიკურ მოქმედებას გარე მიზანი არ აქვს, წახალისებაზე დაფუძნებულ სწავლებას არ შეუძლია უბრალოდ განსაზღვროს ერთი კონკრეტული ობიექტური კრიტერიუმი, რომლითაც "LLM"-ის პასუხების შეფასება იქნებოდა შესაძლებელი.

[8] იხ.: *Silver D., Huang A., Maddison C. et al.*, Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, 2016, 484–489, <https://doi.org/10.1038/nature16961> [25.11.2025].

[9] იხ.: *Vinyals O., Babuschkin I., Czarnecki W. et al.*, Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, vol. 575, 2019, 350–354, <https://doi.org/10.1038/s41586-019-1724-z> [25.11.2025].

[10] იხ.: *Aristotle*, *Nicomachean Ethics*, 1140b8.

წახალისებაზე დაფუძნებული სწავლების მეორე მნიშვნელოვანი გამოწვევა "LLM"-ების კონტექსტში უკავშირდება ეთიკური ღირებულებების მრავალფეროვნებას. ეთიკის „თამაში", რომლის შესრულებაზეც მოდელი იწვრთნება, არ მოიცავს მხოლოდ ერთ ღირებულებას (ან „სიქველეს", არისტოტელეს ტერმინოლოგიით), არამედ სამის ერთობლიობას. იმისათვის, რათა მომხმარებლის მითითებებზე ეთიკურად და სათანადოდ მოახდინოს რეაგირება, "LLM"-მა ერთდროულად უნდა დაიცვას გამოსადეგობის, გულწრფელობისა და უვნებლობის ღირებულებები.

ეს კი დამატებით სირთულეს ქმნის, რადგან აღნიშნული ღირებულებები ხშირად ერთმანეთს ემთხვევა და ზოგ შემთხვევაში ურთიერთსაწინააღმდეგო ხასიათსაც იძენს, განსაკუთრებით მაშინ, როდესაც ისინი უკიდურესობამდე არის მიყვანილი. მათი თანდაყოლილი ბუნდოვანების გამო, ეს ღირებულებები არ არის ერთმანეთთან სრულად თავსებადი ან, მანქანური სწავლების ტერმინოლოგიით რომ ვთქვათ, მათი ერთდროულად „მაქსიმიზება" შეუძლებელია. გამოსადეგობის, გულწრფელობისა და უვნებლობის ღირებულებებს შორის არსებობს წინააღმდეგობა, რომლის დროსაც ერთი ღირებულების ზედმეტად დაცვა - წინ წამოწევა - იწვევს სხვების შესუსტებას. ეს კიდევ უფრო ართულებს იმ ეთიკური მიზნის ტექნიკურად (პროგრამულად) ფორმირებას, რომლის საფუძველზეც შესაძლებელი იქნებოდა "LLM"-ების წვრთნა წახალისებაზე დაფუძნებული სწავლების გზით. შესაბამისად, პრობლემას წარმოადგენს არა მხოლოდ ეთიკური ღირებულებების პროგრამულად განსაზღვრა, არამედ ასევე იმის გადაწყვეტა, თუ რა პროპორციით უნდა იქნეს გამოყენებული თითოეული ღირებულება მომხმარებლის მითითებაზე პასუხის ფორმულირებისას.

ამგვარად, ჩნდება საკვანძო შეკითხვა: როგორ უნდა განისაზღვროს „გამარჯვება" ანუ „მოგება" ეთიკაში, როდესაც საქმე ეხება წახალისებაზე დაფუძნებულ სწავლებას "LLM"-ების კონტექსტში? ეთიკური კრიტერიუმების ბუნდოვანებისა და გამოსადეგობის, გულწრფელობისა და უვნებლობის ღირებულებებს შორის არსებული შეუთავსებლობის გათვალისწინებით, როგორ შეიძლება ჩამოყალიბდეს ზუსტი მიზანი ან ობიექტი, რომლის საფუძველზეც მოდელი უფრო ეთიკურად ქცევას ისწავლიდა?

ხანგრძლივი პერიოდის განმავლობაში ეს პრობლემა მნიშვნელოვნად აფერხებდა "LLM"-ების ეფექტიან გამოყენებას. გარდატეხა მოხდა სპეციალური ტექნიკის განვითარებით, რომელმაც შესაძლებელი გახადა წახალისებაზე დაფუძნებული სწავლების გამოყენება "LLM"-ების ისეთი ამოცანის შესასრულებლად თუ გასაწვრთნელად, როგორიცაა ეთიკური შეფასება. აღნიშნული მიდგომა ცნობილია, როგორც „ადამიანის უკუკავშირზე დაფუძნებული წახალისებით სწავლება" ("reinforcement learning from human feedback" – RLHF).[11]

[11] იხ.: *Christiano P., Leike J., Brown T. et al.*, Deep Reinforcement Learning from Human Preferences, 2017, <https://arxiv.org/abs/1706.03741> [25.11.2025]; *Ziegler D., Stiennon N., Wu J. et al.*, Fine-Tuning Language Models from Human Preferences, 2019, <https://arxiv.org/pdf/1909.08593.pdf> [25.11.2025]; ასევე იხ.: *Stiennon N., Ouyang L., Wu J. et al.*, Learning to summarize from human feedback, *34th Conference on Neural Information Processing Systems* (NeurIPS 2020),

“RLHF”-ის მუშაობის პრინციპი იმაში მდგომარეობს, რომ იგი პირდაპირ, პროგრამულად არ ცდილობს ეთიკური კრიტერიუმების განსაზღვრას. ამის ნაცვლად, “RLHF” იყენებს მანქანური სწავლების შესაძლებლობებს იმისათვის, რომ, შემფასებელთა ჯგუფის (ადამიანებისგან შემდგარი) პრეფერენციების მოდელირების გზით, არაპირდაპირ „აღმოაჩინოს“, ე.ი. განსაზღვროს ასეთი კრიტერიუმების მახასიათებლები. ზოგადად, ეს ტექნიკა ეფუძნება ხუთსაფეხურიან პროცესს:

1. ადამიანის შემფასებელთა ჯგუფს ეძლევა დავალება, შეაფასოს ერთი და იმავე მითითებაზე “LLM”-ის მიერ გენერირებული რამდენიმე პასუხი და დაალაგოს ისინი ყველაზე ეთიკურიდან ყველაზე ნაკლებად ეთიკურამდე, ე.ი. იმის მიხედვით, თუ რამდენად კარგად ახერხებს თითოეული პასუხი გამოსადეგობის, გულწრფელობისა და უვნებლობის ღირებულებების დაბალანსებას;
2. მიღებული მითითებების, პასუხებისა და შემფასებლების მიერ შედგენილი რეიტინგების საფუძველზე იქმნება ზედამხედველობითი სწავლებისთვის განკუთვნილი მონაცემთა ბაზა, სადაც სწორედ ეს რეიტინგი ასრულებს ეტიკეტის როლს;
3. აღნიშნული მონაცემების გამოყენებით იწვრთნება მოდელი, რომელიც სწავლობს იმ არაპირდაპირ მახასიათებლებს, რომლებიც განსაზღვრავს „მომგებიან“ პასუხს ეთიკის „თამაშში“, ანუ იმ კრიტერიუმებს, რომლებიც რეალურად ასახავს გამოსადეგობის, გულწრფელობისა და უვნებლობის ღირებულებებს;
4. ამგვარად გაწვრთნილი „პრეფერენციების მოდელი“ გამოიყენება, როგორც წახალისების ფუნქცია “LLM”-ისთვის (წახალისებაზე დაფუძნებული სწავლების გარემოში); და
5. საბოლოოდ, “LLM” დამატებით იწვრთნება იმგვარად, რომ მისი ქცევა შეესაბამებოდეს გამოსადეგობის, გულწრფელობისა და უვნებლობის ღირებულებებს — მოდელი ჯილდოვდება იმ პასუხებისთვის, რომლებიც პრეფერენციების მოდელის კრიტერიუმებს აკმაყოფილებს, ხოლო დასანქცირდება იმისთვის, რაც ამ კრიტერიუმებს არ შეესაბამება.

მიუხედავად იმისა, რომ “RLHF” საშუალებას იძლევა ეთიკური მიზანი განისაზღვროს წახალისებაზე დაფუძნებული სწავლებისთვის, ეთიკის „ავტომატიზაციის“ ეს მიდგომა მნიშვნელოვან ნაკლოვანებებს შეიცავს. მისი მთავარი ნაკლოვანება არის შემდეგი — ტექნიკა ვერ უზრუნველყოფს იმას, რომ ადამიანებისგან შემდგარი შემფასებელთა ჯგუფის მიერ მიღებული გადაწყვეტილებები რეალურად იყოს სათანადო ან ეთიკური. მხოლოდ ის ფაქტი, რომ ადამიანთა ჯგუფს ევალება შეფასების გაკეთება, თავისთავად არ ნიშნავს, რომ მიღებული შედეგები ეთიკურია. შემფასებლები შესაძლოა თავად

<https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf> [25.11.2025].

იყვნენ მიკერძოებულნი ან მიდრეკილნი შეცდომებისკენ, რის შედეგადაც "RLHF" უბრალოდ ხელახლა დაამკვიდრებს მათ არაეთიკურ მიდგომებს, ამჯერად უკვე — ერთი შეხედვით „ობიექტური" მათემატიკური პროცესის ფორმით.

გარდა ამისა, შემფასებლების მიუკერძოებელი ჯგუფის შემთხვევაშიც კი, ის პირობები, რომლებშიც მათ უწევთ დასკვნების გამოტანა, შესაძლოა იყოს იძულებითი ან ექსპლუატაციური, რაც უარყოფითად აისახება მათი შეფასების ხარისხზე. მაგალითად, "Time Magazine"-ის ცნობით, "OpenAI"-მ ამ პროცესში დაასაქმა კენიელი შემფასებელთა ჯგუფი, ჯგუფის წევრებს შესრულებული სამუშაოსთვის კი საათში 2 დოლარზე ნაკლებს უხდიდა.[12]

"RLHF"-თან დაკავშირებული კრიტიკის საპასუხოდ შემუშავდა კიდევ ერთი მიდგომა — „ხელოვნური ინტელექტის უკუკავშირზე დაფუძნებული წახალისებით სწავლება" ("reinforcement learning from AI feedback" – RLAIF).[13] ეს ტექნიკა იმავე ზოგად პროცესს მისდევს, რომელსაც "RLHF", თუმცა ორი მნიშვნელოვანი განსხვავებით: (1) ადამიანის ნაცვლად, რამდენიმე პასუხის შეფასებას თავად "LLM" ახორციელებს; და (2) ეთიკური ღირებულებების ზოგადი ნაკრების ნაცვლად, "LLM"-ს გადაეცემა ე.წ. „კონსტიტუცია", რომელიც მოიცავს პრინციპების ერთობლიობას და ამ პრინციპებს გამოყენების შესაბამის ნიმუშებს. სწორედ ამ უკანასკნელი თავისებურების გამო "RLAIF"-ს ხშირად მოიხსენიებენ, როგორც „კონსტიტუციურ ხელოვნურ ინტელექტს" ("constitutional AI").

თუმცა, მიუხედავად იმისა, რომ "RLAIF"-მა შესაძლოა გაზარდოს შედეგების მასშტაბურობა, იგი, "RLHF"-ის მსგავსად, კვლავ მოიცავს მნიშვნელოვან ნაკლოვანებებს. კერძოდ, ისევე როგორც "RLHF" ვერ უზრუნველყოფს ადამიანებისგან შემდგარ შემფასებელთა ჯგუფის გადაწყვეტილებების რეალურ ეთიკურობას, "RLAIF" ასევე ვერ იძლევა გარანტიას, რომ "LLM"-ის მიერ განხორციელებული შეფასებები არ იქნება მიკერძოებული ან სხვაგვარად არასწორი. მეტიც, ეს რისკი შესაძლოა, კიდევ უფრო მძაფრად გამოვლინდეს "RLAIF"-ის შემთხვევაში, ვინაიდან მოდელს ეთიკური შეფასების გაკეთება ევალება იმ ეტაპზე, როდესაც იგი თვითონაც ჯერ კიდევ სრულად არ არის ამ მიმართულებით გაწვრთნილი.

3. მონაცემთა დაცვასა და პირადი ცხოვრების ხელშეუხებლობასთან დაკავშირებული რისკები

დიდი ენობრივი მოდელების გამოყენება უკავშირდება მნიშვნელოვან რისკებს პირადი ცხოვრების ხელშეუხებლობის, პერსონალური მონაცემების დაცვისა და მონაცემთა უსაფრთხოების თვალსაზრისით. ამ რისკების ნაწილის პრევენცია ან შემსუბუქება შესაძლებელია, ხოლო ნაწილი შესაძლოა თავად

[12] *Perrigo B.*, Open AI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, *Time*, January 18, 2023, <https://time.com/6247678/openai-chatgpt-kenya-workers/> [25.11.2025].

[13] იხ.: *Bai Y., Kadavath S., Kundu S. et al.*, Constitution AI: Harmlessness from AI Feedback, December 2022, <https://arxiv.org/abs/2212.08073> [25.11.2025].

სისტემების ბუნებრივი მახასიათებლიდან გამომდინარეობდეს. წინამდებარე ქვეთავში მიმოვიხილავთ იმ ძირითად რისკებს, რომლებიც "LLM"-ების გამოყენებასთან არის დაკავშირებული. ამასთან, გასათვალისწინებელია, რომ "LLM"-ები წარმოშობს ისეთ საფრთხეებსაც, რომლებიც, მიუხედავად იმისა, რომ უშუალოდ არ უკავშირდება პირადი ცხოვრების ხელშეუხებლობას ან მონაცემთა უსაფრთხოებას, მნიშვნელოვან გავლენას ახდენს ინდივიდებზე და შესაძლოა მოქცეული იყოს მონაცემთა დაცვის ორგანოების „მომხმარებელთა დაცვის მანდატში". ასეთ საფრთხეებს მოიცავს, მათ შორის, ინფორმაციის მანიპულირება, მონაცემთა დამუშავების მასშტაბების ზრდა, დეზინფორმაცია და მცდარი ინფორმაციის გავრცელება. აღნიშნულ ზიანზე ასევე იქნება საუბარი წინამდებარე ქვეთავის ბოლოს.

3.1. მონაცემთა დამუშავების მასშტაბის ზრდა

მასშტაბური სასწავლო მონაცემების საჭიროების გათვალისწინებით, ბევრი "LLM"-ის განმავითარებელი ქმნის სისტემებს, რომლებიც განურჩევლად და უწყვეტად აგროვებს მონაცემებს ინტერნეტიდან.[14] მიუხედავად იმისა, რომ ზოგიერთი კომპანია ამგვარად შეგროვებულ მონაცემებს გამოყენებამდე ამოწმებს და „ასუფთავებს", კარგი პრაქტიკის მაგალითად შეიძლება დასახელდეს "Google"-ის C4 მონაცემთა ბაზა, რომელიც ან საერთოდ არ ახორციელებს ამ შემოწმებას, ან ვერ ახერხებს შემოწმების პროცესის ხარისხის უზრუნველყოფას ინფორმაციის მოცულობის მნიშვნელოვნად შეზღუდვის გარეშე.[15] ამის შედეგად, სასწავლო მონაცემთა ბაზები შესაძლოა შეიცავდეს არაზუსტ, მიკერძოებულ ან დისკრიმინაციულ ინფორმაციას, მათ შორის – პერსონალურ მონაცემებს იმ პირების შესახებ, რომლებიც სრულად არ არიან ინფორმირებულნი იმ ფაქტის შესახებ, რომ მათი მონაცემები გამოიყენება "LLM"-ების გაწვრთნის პროცესში.

3.2. მონაცემთა სუბიექტის უფლებების ხელყოფა

დიდი ენობრივი მოდელების ბუნება მნიშვნელოვნად ართულებს მონაცემთა სუბიექტების მიერ გარკვეული უფლებების პრაქტიკულად განხორციელებას, განსაკუთრებით კი პერსონალური მონაცემების გასწორების ან მათი წაშლის მოთხოვნის უფლებას, მაშინ, როდესაც ასეთი მონაცემები უკვე ინტეგრირებულია სასწავლო მონაცემთა ბაზებში. მართალია, ზოგიერთი მონაცემთა ბაზა შედარებით მკაცრად მოწმდება როგორც მონაცემთა

[14] მაგალითად იხ.: *Hines K.*, OpenAI Launces GPTBot With Details on How to Restrict Access, *Search Engine Journal*, Aug. 7, 2023, <https://www.searchenginejournal.com/openai-launches-gptbot-how-to-restrict-access/493394/> [25.11.2025]; *Schaul K. et al*., Inside the secret list of websites that make AI like ChatGPT sound smart, *Washington Post*, Apr. 19, 2023, <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/> [25.11.2025].

[15] *Center for Countering Digital Hate*, Misinformation on Bard, Google's New AI Chat, April 5, 2023, <https://counterhate.com/research/misinformation-on-bard-google-ai-chat/> [25.11.2025].

მოპოვების, ისე მათი გამოყენების აუცილებლობის თვალსაზრისით, თუმცა განსაკუთრებით ვებწყაროებიდან მასობრივად შეგროვებული მონაცემები ხშირად შეიცავს ისეთ პერსონალურ მონაცემებს, რომელთა დამუშავებაც არ არის აუცილებელი კონკრეტული მიზნის მისაღწევად. გარდა ამისა, ასეთ მონაცემთა ბაზებში შესაძლოა, მოხვდეს პერსონალური მონაცემები, რომლებიც საჯაროდ ხელმისაწვდომი გახდა მონაცემთა უსაფრთხოების დარღვევის – ინციდენტის – შედეგად, ან წარმოადგენს კონკრეტული პირების შესახებ არაზუსტ ან ცილისმწამებლურ ინფორმაციას.

3.2.1. თავის სხვა პიროვნებად წარმოჩენა (იმპერსონაცია) და გამოძალვა

დიდი ენობრივი მოდელების შესაძლებლობები შესაძლოა, განზრახ ბოროტად იქნეს გამოყენებული. ამის მაგალითი შესაძლოა, იყოს კონკრეტული პირის პერსონალური მონაცემების ბოროტად გამოყენება ან ყალბი პერსონალური მონაცემების გენერირება, რომელთა უარყოფა ან გადამოწმება პრაქტიკაში მნიშვნელოვან გამოწვევებს უკავშირდება. აღნიშნულმა შეიძლება სერიოზულად დააზიანოს ინდივიდის ფსიქიკური ჯანმრთელობა, სოციალური ურთიერთობები, რეპუტაცია და სხვა მნიშვნელოვანი ინტერესები.

3.2.2. თაღლითობა

ინდივიდებს შეუძლიათ "LLM"-ების გამოყენება ავტომატიზებული ტექსტური შეტყობინებების, ელექტრონული წერილებისა და სხვა მასობრივი კომუნიკაციის საშუალებების შესაქმნელად. ამასთან, "LLM"-ების მიერ გენერირებული ტექსტი შესაძლოა, გამოყენებულ იქნეს ხელოვნურად შექმნილ აუდიო და ვიდეომასალასთან ერთად, რაც შესაძლებელს ხდის უფრო დამაჯერებელი იმიტაციის ან სხვა პირის სახელით კომუნიკაციის განხორციელებას. აღნიშნულის შედეგად იზრდება არა მხოლოდ თაღლითური აქტივობების საერთო მოცულობა, არამედ მნიშვნელოვნად ფართოვდება თაღლითობაში ჩართულ პირთა წრეც, ვინაიდან "LLM"-ები საშუალებას აძლევს იმ პირებსაც კი, რომელთაც კონკრეტულ ენაში ან კომუნიკაციის გარკვეულ ფორმებში შეზღუდული უნარები აქვთ, შექმნან ბუნებრივად ჟღერადი და დამაჯერებელი ტექსტები, რომლებიც სხვა შემთხვევაში უფრო იოლად ამოიცნობა, როგორც თაღლითური.

3.3. მონაცემთა უსაფრთხოების რისკები

ჰაკერებს შეუძლიათ "LLM"-ების გამოყენება მავნე პროგრამული კოდის ("malware code") სხვადასხვა ვერსიის შემუშავებისა და დახვეწის მიზნით, აგრეთვე „ფიშინგისა" და „მიზნობრივი ფიშინგის" ("spear-phishing") კამპანიების განხორციელებისთვის, მათ შორის ბიზნესებზე მიმართული თაღლითური ელექტრონული წერილების შესაქმნელად, რომელთა მიზანია

მომხმარებელთა ანგარიშის მონაცემების მოპოვება ან ელექტრონული კომუნიკაციის კომპრომეტირება.[16] გარდა ამისა, შესაძლოა, წარმოიშვას "LLM"-ებისთვის დამახასიათებელი ახალი სახის საფრთხეებიც, მათ შორის სასწავლო მონაცემთა ნაკრებებში შეტანილი ინფორმაციის მიზანმიმართული ამოღება („data mining") ან მონაცემთა ბაზების სტრატეგიული და განზრახ „დაბინძურება" არასაიმედო ან მავნე მონაცემებით.

3.4. მიკერძოება

დიდი ენობრივი მოდელები შეიძლება მარტივად იქცეს მიკერძოების წყაროდ, როგორც სასწავლო მონაცემთა ბაზებში არსებული მიკერძოებული ინფორმაციის ინტეგრირების შედეგად, ისე იმ ალგორითმების მეშვეობით, რომლებიც თავად ვითარდება მიკერძოებული მიმართულებით. შედეგად, მიკერძოება აისახება როგორც სისტემის შიდა პროცესებში, ისე "LLM"-ის მიერ გენერირებულ შედეგებში. მართალია, მიკერძოება შესაძლოა, არსებობდეს წინასწარ შერჩეულ და დამუშავებულ მონაცემთა ბაზებშიც, თუმცა განსაკუთრებით მაღალი რისკი წარმოიშობა იმ შემთხვევებში, როდესაც მონაცემთა ბაზები ყალიბდება ინტერნეტიდან მასობრივად შეგროვებული ინფორმაციის საფუძველზე. ასეთ პირობებში სასწავლო მონაცემთა ნაკრები მუდმივად იზრდება და, როგორც წესი, რეგულარულად არ მოწმდება სიზუსტის, მიკერძოების, გამოყენების მიზანშეწონილობისა და სხვა არსებითი კრიტერიუმების მიხედვით.

3.5. დეზინფორმაცია

დიდი ენობრივი მოდელები ხელს უწყობს დეზინფორმაციის უფრო დიდი მოცულობით გენერირებას, რომლის გავრცელებაც შესაძლებელია ადვილად, დაბალი დანახარჯით და გაცილებით მაღალი სიჩქარით. მსგავსი დეზინფორმაციის პოტენციური გავლენა განსაკუთრებით საგანგაშოა ისეთ სფეროებში, როგორებიცაა არჩევნები, პოლიტიკა და მედია – განსაკუთრებით ჯანმრთელობასა და უსაფრთხოებასთან დაკავშირებულ საკითხებში – ისევე როგორც სხვა კრიტიკულად მნიშვნელოვან სფეროებში.

[16] მაგალითად იხ.: *SlashNext*, The State of Phishing 2023, *SlashNext Security*, Oct. 2023, <https://slashnext.com/state-of-phishing-2023/> [25.11.2025]; *Groll E.*, ChatGPT Shows Promise of Using AI to Write Malware*, CyberScoop*, December 6, 2022, <https://cyberscoop.com/chatgpt-ai-malware/> [25.11.2025]; *Hassold C.*, Executive Impersonation Attacks Targeting Companies Worldwide, *Abnormal Blog*, February 16, 2023, <https://abnormalsecurity.com/blog/midnight-hedgehog-mandarin-capybara-multilingual-executive-impersonation> [25.11.2025]; *Center for Strategic and International Studies*, A Conversation on Cybersecurity with NSA's Rob Joyce*, YouTube*, April 11, 2023, <https://youtu.be/MMNHNjKp4Gs?t=530> [25.11.2025]. (8:50 mark).

3.6. **მისინფორმაცია / ჰალუცინაციები**

მისინფორმაცია დეზინფორმაციის მსგავს პრობლემებს წარმოშობს ერთი მნიშვნელოვანი განსხვავებით - მისინფორმაციის გამავრცელებელმა პირებმა შეიძლება დაიჯერონ, რომ ის ინფორმაცია, რომელსაც ავრცელებენ, ზუსტია. მისინფორმაცია შეიძლება გენერირებული იყოს მომხმარებლის მიერ მიწოდებული პარამეტრებიდან ან დიდი ენობრივი მოდელის მიერ გენერირებული არაზუსტი ინფორმაციიდან.

ზოგადად, ჰალუცინაციების ორი სახეობა არსებობს. პირველი და უფრო აშკარა ჰალუცინაციები, რომლებიც გამოწვეულია ცრუ ან შეცდომაში შემყვანი შინაარსის მინიშნებებით. მაგ., რაც მოხდა მეტა "Galactica"-ს დიდი ენობრივი მოდელის შემთხვევაში. თავდაპირველად იგი განსაზღვრული იყო, როგორც „სამეცნიერო ცოდნის" გაღრმავების მექანიზმი,[17] თუმცა ფუნქციონირება შეწყვიტა სამი დღის შემდეგ, რაც აღმოჩნდა, რომ ის ქმნიდა სამეცნიერო ჟღერადობის, მაგრამ სრულიად ცრუ ვიკიპედიის სტატიებს ისეთ გამოგონილ თემებზე, როგორიცაა „ნაკადის კონდენსატორი" ან „სტრიფ-სეინფელდის თეორემა".[18]

ჰალუცინაციების მეორე სახეობა დიდი ენობრივი მოდელისგან წარმოიშობა, რომლის შესახებ ინფორმაცია მომხმარებლისთვის უცნობია. ეს ჰალუცინაცია უფრო მავნეა და მისი აღმოჩენა რთულია. არსებობს მრავალი მაგალითი,[19] მაგრამ ერთი ცნობილი შემთხვევა ეხება პირის წინააღმდეგ წაყენებულ ცრუ სისხლის სამართლის ბრალდებას. კითხვაზე „რომელ გახმაურებულ საქმეებში მონაწილეობდნენ სამართლის პროფესორები?",[20] "ChatGPT"-მ წარმოადგინა ცრუ ნარატივი, რომელშიც დამტკიცებული იყო, რომ კონკრეტული პროფესორი სტუდენტის სექსუალურ შევიწროებაში იყო ბრალდებული. თუმცა კიდევ უფრო შემაშფოთებელი აღმოჩნდა ის, რომ ცრუ ნარატივის დასამოწმებლად მითითებული იყო ცრუ წყაროები, რომლებიც ამყარებდა აღნიშნულ ნარატივს.

4. კონფიდენციალობის პრინციპი და ტექნიკური ზომები

დიდი ენობრივი მოდელისთვის დამახასიათებელი მონაცემთა დაცვისა და კონფიდენციალობის რისკები ახალი არ არის. დიდი ენობრივი მოდელები და, ზოგადად, გენერირებადი ხელოვნური ინტელექტი, ხელოვნური ინტელექტის სხვა ფორმებისგან განსხვავებით, გამოირჩევა დამუშავებული მონაცემების მასშტაბით, მოდელის შემუშავებისა და მასში გამოყენებული

[17] See *Taylor R., Kardas M., Cucurell G. et al.*, Galactica: A Large Language Model for Science, 2022, <https://arxiv.org/abs/2211.09085> [25.11.2025].

[18] See *Davis E. and Sundstrom A.*, Experiment with GALACTICA, 2022, <https://cs.nyu.edu/~davise/papers/ExperimentWithGalactica.html> [25.11.2025].

[19] See *Marcus G. and Davis E.*, Large Language Models like ChatGPT say The Darnedest Things, 2023, <https://garymarcus.substack.com/p/large-language-models-like-chatgpt> [25.11.2025].

[20] See *Volokh E.*, Large Libel Models: ChatGPT-3.5 Erroneously Reporting Supposed Felony Pleas, Complete with Made-Up Media Quotes?, *Reason*, 2023, <https://reason.com/volokh/2023/03/17/large-libel-models-chatgpt-4-erroneously-reporting-supposed-felony-pleas-complete-with-made-up-media-quotes/> [25.11.2025].

ტექნიკის სირთულით, შემდგომში აღნიშნული მოდელის დანერგვის უპრეცედენტო მასშტაბითა და ტემპით.

ამ ნაწილში ჩვენ განვიხილავთ კონფიდენციალობის პრინციპის გამოყენებას დიდი ენობრივი მოდელისთვის, ასევე, გენერირებად ხელოვნურ ინტელექტთან დაკავშირებულ მონაცემთა დაცვისა და კონფიდენციალობის რისკებს.

4.1. კონფიდენციალობის პრინციპი

4.1.1. კანონიერი საფუძველი

პერსონალური მონაცემების დამუშავების პროცესში ხელოვნური ინტელექტის სისტემების შემქმნელებსა და განმათავსებლებს უნდა ჰქონდეთ მონაცემთა დამუშავების კანონიერი საფუძველი და მათი მოქმედებები შესაბამისობაში უნდა იყოს სხვა მოქმედ კანონმდებლობასთან (მაგ., საავტორო უფლებების შესახებ კანონი). ევროკავშირის „მონაცემთა დაცვის ძირითადი რეგულაციის“ მე-6 მუხლი გვთავაზობს ექვს სამართლებრივ საფუძველს, განსაკუთრებული კატეგორიის მონაცემებისთვის მე-9 მუხლით გათვალისწინებული დამატებითი მოთხოვნებით.

გენერირებადი ხელოვნური ინტელექტის საცდელი მონაცემების მისაღებად უმნიშვნელოვანესია, აღინიშნოს, რომ საჯაროდ ხელმისაწვდომი მონაცემები უმეტეს იურისდიქციაში კვლავ ექცევა მონაცემთა დაცვისა და პირადი ცხოვრების ხელშეუხებლობის კანონმდებლობის ფარგლებში, როგორც ეს ხაზგასმულია პირადი ცხოვრების ხელშეუხებლობის გლობალური ასამბლეის (“GPA”) საერთაშორისო აღსრულების თანამშრომლობის სამუშაო ჯგუფის (“IEWG”) ბოლოდროინდელ განცხადებაში.[21] აგრეთვე, აშშ-ის ფედერალურ სასამართლოებსა და დიდ ბრიტანეთში საავტორო უფლებების დაცვის თაობაზე მისაღებ გადაწყვეტილებებს[22] შეიძლება მნიშვნელოვანი წონა ჰქონდეს კანონიერების პრინციპთან მიმართებით, თუ ჩაითვლება, რომ ვებგვერდიდან მოპოვებული საცდელი მონაცემები არღვევს საავტორო უფლებებისა და ინტელექტუალური საკუთრების კანონმდებლობას. მონაცემთა დაცვის საზედამხედველო ორგანოები, რა თქმა უნდა, ეყრდნობიან სასამართლოების მიერ მიღებულ გადაწყვეტილებებს.

4.1.2. მიზნის შეზღუდვა

დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელებმა და განმათავსებლებმა უნდა

[21] *Global Privacy Assembly (GPA) International Enforcement Cooperation Working Group*, Joint statement on data scraping and the protection of privacy, August 2023, <https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf> [25.11.2025].

[22] *David E.*, Getty lawsuit against Stability AI to go to trial in the UK, *The Verge*, December 4, 2023, <https://www.theverge.com/2023/12/4/23988403/getty-lawsuit-stability-ai-copyright-infringement> [25.11.2025].

უზრუნველყონ, რომ პერსონალური მონაცემები დამუშავდეს კონკრეტული, მკაფიო და ლეგიტიმური მიზნებისთვის. გარდა ამისა, მონაცემები არ უნდა დამუშავდეს ფიზიკური პირების გონივრული მოლოდინის ფარგლებს გარეთ ან შეუთავსებელი მიზნებისთვის.

4.1.3. მონაცემთა მინიმიზაცია

დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელებმა და განმათავსებლებმა მონაცემთა დამუშავება უნდა შეზღუდონ იმით, რაც „აუცილებელია" მათი მიზნისთვის. რაც უფრო დიდია დამუშავებული პერსონალური მონაცემების მოცულობა, მით უფრო დიდია რისკები.

ნებისმიერი ინფორმაციის ადრეული შეზღუდვა მნიშვნელოვანია მონაცემთა სუბიექტის უფლების დაცვისთვის. ამ მიზნით სისტემის შემქმნელები უნდა ეცადონ, რომ მონაცემთა ნაკრებებში პერსონალური მონაცემების ნებისმიერ შემთხვევაზე გამოიყენონ მონაცემთა მინიმიზაცია. გავრცელებული მიდგომაა მონაცემთა გამორიცხვისა და ანონიმიზაციის პროცედურის გამოყენება, თუმცა, ამ ტექნიკის გამოყენების შემთხვევაშიც კი, შეიძლება რთული იყოს იმის სრულად უზრუნველყოფა, რომ მონაცემთა ნაკრებები არ შეიცავდეს პერსონალურ ინფორმაციას. იმ შემთხვევებში, როდესაც წინასწარ შეგროვებული მესამე მხარის მონაცემთა ნაკრებები გამოიყენება სისტემის გამოსაცდელად, ასევე მნიშვნელოვანია პერსონალური ინფორმაციის ამოღება დამუშავების შემდგომ ეტაპებზე.

4.1.4. გამჭვირვალობა

დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელებმა და განმათავსებლებმა უნდა განახორციელონ გამჭვირვალობის უზრუნველმყოფი ზომები, განსაკუთრებით მონაცემთა სუბიექტებთან მიმართებით, რომლებსაც აქვთ ინფორმაციის მიღების უფლება. ეს უნდა მოიცავდეს ინფორმაციას იმის შესახებ, თუ რა მონაცემი, როგორ, როდის და რატომ გროვდება და გამოიყენება სისტემის სწავლების პროცესში, მათ შორის: საცდელი მონაცემების წყაროებს, პერსონალური ინფორმაციის წაშლის წინასწარ და შემდგომ დამუშავების ზომებს და გენერირებული ტექსტის სანდოობას.

4.1.5. უსაფრთხოება

დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელებმა და განმათავსებლებმა უნდა დანერგონ უსაფრთხოების ზომები. პერსონალური მონაცემები უსაფრთხოდ უნდა იყოს შენახული შენახვის, შემუშავების, ასევე დანერგვის შემდგომი პერიოდის განმავლობაში, რათა გათვალისწინებულ იქნეს ისეთი რთული

საკითხები, როგორებიცაა სისტემაზე ე.წ. "სწრაფი ინექციის" შეტევები, მოდელის ინვერსიის შეტევები[23] და მონაცემთა გაჟონვა.

4.1.6. ანგარიშვალდებულება

დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელებმა და განმათავსებლებმა უნდა უზრუნველყონ, რომ მათ შეუძლიათ მონაცემთა დაცვის მოთხოვნებთან შესაბამისობის დემონსტრირება. ანგარიშვალდებულება, ფაქტობრივად, მეტაპრინციპია, რომელიც გარანტის ფუნქციას ასრულებს.

4.1.7. სიზუსტე

დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელებმა და განმათავსებლებმა უნდა უზრუნველყონ, რომ მათ მიერ დამუშავებული ნებისმიერი პერსონალური მონაცემი იყოს ისეთივე ზუსტი, სრული და განახლებული, როგორც ეს აუცილებელია იმ მიზნებისთვის, რომლისთვისაც ისინი უნდა იქნეს გამოყენებული. ეს განსაკუთრებით ეხება პერსონალურ მონაცემებს, რომლებიც გამოიყენება დიდი ენობრივი მოდელების ან გენერირებადი ხელოვნური ინტელექტის სწავლებისთვის.

ამ პრინციპის მხარდასაჭერად სისტემა შეიძლება განახლდეს (მაგალითად, მოდელის დახვეწით ან გადამზადებით) იმ შემთხვევებში, როდესაც აღმოჩენილი იქნება არაზუსტი ან მოძველებული მონაცემები, როგორებიცაა სისტემის საცდელი მონაცემები. გარდა ამისა, მომხმარებელს უნდა ეცნობოს მონაცემების სიზუსტესთან დაკავშირებული ნებისმიერი ცნობილი პრობლემის ან შეზღუდვის შესახებ. ეს შეიძლება მოიცავდეს იმ შემთხვევებს, როდესაც საცდელი მონაცემები დროში შეზღუდულია (ანუ შეიცავს ინფორმაციას მხოლოდ გარკვეულ თარიღამდე); როდესაც შინაარსზე შეიძლება უარყოფითად იმოქმედოს გარკვეულმა წყაროებმა; ან როდესაც არსებობს კონკრეტული თემები ან მინიშნებები, რომლებიც, როგორც წესი, არაზუსტ შედეგებამდე მიგვიყვანს.

4.1.8. მონაცემთა სუბიექტის უფლებები

მონაცემთა სუბიექტების უფლებები მონაცემთა დაცვის ქვაკუთხედია. დიდი ენობრივი მოდელებისა და გენერირებული ხელოვნური ინტელექტის სისტემების შემქმნელები და განმათავსებლები, რომლებიც ამუშავებენ პერსონალურ მონაცემებს, ვალდებულნი არიან, უზრუნველყონ, რომ პირებს შეეძლოთ თავიანთ მონაცემებზე წვდომა, მათი გასწორება, წაშლა და

[23] *Veale M., Binns R. and Edwards L.*, Algorithms that remember: model inversion attacks and data protection law, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, October 2018, <https://doi.org/10.1098/rsta.2018.0083> [25.11.2025].

დამუშავებაზე უარის თქმა, სხვა უფლებებთან ერთად. ეს ასევე მნიშვნელოვანია განსაკუთრებული კატეგორიის პერსონალურ მონაცემებთან და ბავშვთა უფლებების დაცვის საკითხთან დაკავშირებით.

4.2. ტექნიკური საკითხები

როდესაც საქმე დიდი ენობრივი მოდელის სწავლებას ეხება, არსებობს ტექნიკური ჩარევის რამდენიმე ეტაპი და ტიპი, რომელთა განხორციელებაც შესაძლებელია რისკის შესამცირებლად. ამ ნაწილში ჩვენ ყურადღებას გავამახვილებთ ჩარევის ზოგიერთი ვარიანტის გამოყენების უპირატესობებსა და ნაკლოვანებებზე.

4.2.1. მეთვალყურეობა და წინასწარი დამუშავება

დიდი ენობრივი მოდელები ფუნქციონირებისთვის საჭიროებს მასშტაბური რაოდენობით ტექსტურ მონაცემებს და, მათი დამახსოვრების უნარის გათვალისწინებით,[24] მნიშვნელოვანია მოდელების ფუნქციონირების უზრუნველყოფა იმავე რისკის დაზღვევით, როგორც მათი მომზადებისას. მონაცემთა ნაკრებების შეგროვებისა და მეთვალყურეობის პროცესში შესაძლებელია გადაწყვეტილებების მიღება და ნაბიჯების გადადგმა აღნიშნული რისკის შესამცირებლად.

წყაროს მეთვალყურეობა: განხილვის საგანია, თუ რა ტიპის მონაცემები გამოიყენება მოდელის სწავლებისთვის.[25] ერთი მარტივი განსხვავება – არის თუ არა ეს ინფორმაცია პერსონალური მონაცემი, რადგან ამ ტიპის მონაცემებს აშკარა გავლენა აქვს კონფიდენციალობაზე, როდესაც ისინი გამოიყენება სისტემის გამოცდისთვის, მონაცემთა სუბიექტის სურვილის გათვალისწინების გარეშე. გასათვალისწინებელია საჯაროდ ხელმისაწვდომი და ღია მონაცემები. მიუხედავად იმისა, რომ ყველა საჯარო (ან ღია) მონაცემი საჯაროდ ხელმისაწვდომია, არ უნდა განიხილებოდეს ისე, თითქოს ის მისი ფართო გამოყენებისთვის კვლავ ნებადართულია. განსხვავება მდგომარეობს მონაცემების ხელმისაწვდომობის განზრახვასა და მოლოდინში: საჯარო მონაცემები ეხება მონაცემებს, რომლებიც შეიქმნა ფართოდ გაზიარებისა და გამოყენების მიზნით, მაგალითად, სამთავრობო მონაცემთა ნაკრებები და/ან ვიკიპედიის ტექსტური ნაწილი, მაშინ, როდესაც ზოგიერთი საჯაროდ ხელმისაწვდომი მონაცემი შეიძლება გაზიარებულიყო კონკრეტული შინაარსით გამოყენებისა და მოხმარების მიზნით, მაგალითად, სოციალური მედიის პოსტები და პროდუქტის მიმოხილვები. კვლევამ აჩვენა, რომ აკადემიური კვლევის კონტექსტშიც კი, სოციალური მედიის მომხმარებლები შეიძლება არ გრძნობდნენ თავს კომფორტულად, როდესაც მათი მონაცემები

[24] *Carlini N., Tramer F., Wallace E. et al.*, Extracting Training Data from Large Language Models, *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650, <https://arxiv.org/abs/2012.07805> [25.11.2025].
[25] "Train" used here encompasses both initial training and any fine tuning or additional training steps.

გამოიყენება მათი თანხმობის გარეშე,[26] მაშინაც კი, თუ ის საჯაროდ ხელმისაწვდომია. ამ მოდელების რისკის შესამცირებლად მნიშვნელოვანია მონაცემების მიღება იმ წყაროებიდან, რომლებშიც მონაცემების გამოყენების კონფიდენციალობის მოლოდინები შეესაბამება დიდი ენობრივი მოდელისთვის დასახულ მიზანს.

წინასწარი დამუშავება (მგრძნობიარე მონაცემების წაშლა): მონაცემთა ნაკრებების შედგენის შემდეგი ნაბიჯი გულისხმობს ამ მონაცემების წინასწარ დამუშავებას მოდელების მომზადებისთვის გამოყენებამდე. ამ ეტაპზე შესაძლებელია ავტომატიზებული მექანიზმების გამოყენება მგრძნობიარე ინფორმაციის აღმოსაჩენად და წასაშლელად, მაგალითად, პერსონალური ინფორმაციის, ჯანმრთელობის შესახებ ინფორმაციის და სხვა ინფორმაციის შესახებ. აღნიშნული მექანიზმები შეიძლება მოიცავდეს ამ ინფორმაციის აღმოჩენას და მისი ადამიანის მიერ განხილვისთვის მონიშვნას, ინფორმაციის ავტომატურ წაშლას, ჩანაცვლებას ან დაფარვას.

4.2.2. დიფერენციალური კონფიდენციალობა

დიდი ენობრივი მოდელის მომზადებისას შესაძლებელია კონფიდენციალობის გამაძლიერებელი ტექნოლოგიების, მათ შორის, როგორიცაა ("DP"), [27] გამოყენება იმ მოდელების მოსამზადებლად, რომლებიც პირადი ხასიათისაა. ეს შეიძლება გაკეთდეს სხვადასხვა ეტაპზე (მაგ., საცდელი მონაცემები, გამოცდა, შედეგი), განხილვის სხვადასხვა ერთეულით (მაგ., განხილვის დონე, ჯგუფის დონე) და სხვადასხვა პირობებში (მაგ., ცენტრალური, ადგილობრივი, განაწილებული). თითოეულ მათგანს აქვს უნიკალური სარგებელი და ხარჯები, რომლებსაც ამ ნაწილში განვიხილავთ. სხვადასხვა მიდგომის, მათი განხორციელებისა და მოსაზრებების უფრო ყოვლისმომცველი წარმოდგენისთვის მკითხველს ვურჩევთ გაეცნოს სტატიას: „როგორ განვახორციელოთ ML DP-Fy".[28]

---

[26] *Fiesler C. and Proferes N.*, 'Participant' Perceptions of Twitter Research Ethics, *Social Media + Society,* Vol. 4(1), 2018, <https://doi.org/10.1177/2056305118763366> [25.11.2025].

[27] The definition of differential privacy that is being used in this section is the one proposed in *Dwork C., McSherry F., Nissim K. and Smith A.*, Calibrating noise to sensitivity in private data analysis, *Procedures of the Third Conference on Theory of Cryptography (TCC)*, 265–284, <http://dx.doi.org/10.1007/11681878_14> [25.11.2025]:

> ჩვენ ვამბობთ, რომ ორი მონაცემთა ნაკრები D და D′ მეზობლები არიან, თუ ისინი განსხვავდებიან ზუსტად ერთი ჩანაწერით; უფრო ზუსტად, ერთი მონაცემთა ნაკრები არის მეორის ასლი, მაგრამ ერთი ჩანაწერით დამატებული ან ამოღებული. დავუშვათ ε არის დადებითი სკალარი. მექანიზმი A უზრუნველყოფს ε-დიფერენციალურ კონფიდენციალობას, თუ ნებისმიერი ორი მეზობელი მონაცემთა ნაკრებისთვის D და D′ და ნებისმიერი S ⊆ Range(A)-სთვის,
>
> $$P[A(D) \in S] \leq \exp(\varepsilon) \times P[A(D') \in S]$$

[28] *Ponomareva N., Vassilvitskii S., Xu Z. et al.*, How to DP-fy ML: A Practical Tutorial to Machine Learning with Differential Privacy, *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23*), 2023, 5823–5824, <https://arxiv.org/abs/2303.00654> [25.11.2025].

კონფიდენციალობის ერთეული: შესაბამისი ერთეულის განსაზღვრა კრიტიკულად მნიშვნელოვანია იმის უზრუნველსაყოფად, რომ არსებობდეს შესაბამისი დონე, რადგან ის განსაზღვრავს, თუ რა ჩაითვლება მონაცემთა ნაკრებად. განხილვის დონის მონაცემთა ბაზა უზრუნველყოფს დაცვას მონაცემთა ნაკრებში შემავალი თითოეული ნიმუშისთვის, ხოლო ჯგუფური დონის მონაცემთა ბაზა უზრუნველყოფს დაცვას აბსტრაქციის უფრო მაღალ დონეზე (მაგ., მომხმარებლის დონეზე, დოკუმენტის დონეზე და ა.შ.). დიდი ენობრივი მოდელებისთვის, შესაძლოა უკეთესი იყოს ჯგუფური დონის მონაცემთა ბაზა, რადგან ამ მოდელების გამოცდისას გამოყენებული თანმიმდევრობა არა მხოლოდ მოდელის მუშაობაზე იმოქმედებს, არამედ კონფიდენციალობის გარანტიებზეც და ამ ორი ფაქტორის გამოყოფა შეიძლება უფრო სასარგებლო იყოს. გარდა ამისა, განხილვის დონეზე მოქმედებათა გამეორების მაღალი ალბათობა, სავარაუდოდ, მნიშვნელოვნად შეასუსტებს გარანტიებს. თუმცა მაინც მნიშვნელოვანია ყურადღებით განვიხილოთ, დაჯგუფების რომელ დონეზეა მიზანშეწონილი კონფიდენციალობის ერთეულის განსაზღვრა. მაგალითად, მიუხედავად იმისა, რომ შეიძლება გვინდოდეს მომხმარებლის დონის "DP"-ის მიწოდება, იმის გათვალისწინებით, რომ ამ მოდელების წვრთნისთვის ხშირად გამოყენებული მონაცემები ინტერნეტიდან საჯაროდ ხელმისაწვდომი ტექსტია, ამის გაკეთება შესაძლოა შეუძლებელი იყოს, რადგან შეუძლებელია იმის იდენტიფიცირება, თუ რომელი ნიმუშები იყო წარმოდგენილი მომხმარებელთა მიერ.

განხორციელების ეტაპი: არსებობს დეტალიზაციის რამდენიმე დონე, რომელიც დაკავშირებულია იმასთან, თუ როდის შეიძლება "DP"-ის იმპლემენტაცია. გამარტივების მიზნით, ეს ქვეთავი მას მხოლოდ მოდელის ტრენინგის დონეზე განიხილავს. მოდელის ტრენინგის ეტაპზე "DP"-ის გამოყენება იძლევა გარანტიას, რომ მეორე მხარე ვერ შეძლებს განასხვაოს მოდელები, რომლებიც მოიცავს ან არ მოიცავს კონკრეტულ განხილვას ტრენინგის მონაცემებში. ენობრივი მოდელებისთვის ყველაზე მიზანშეწონილი მიდგომები ეხება გრადიენტული მოქმედების საკითხს, რომელშიც ყველაზე ხშირად გამოყენებული ალგორითმია: "DP-SGD". მიზანშეწონილია, რომ დიფერენციალური კონფიდენციალობის დანერგვით დაინტერესებულმა პირებმა გაითვალისწინონ ექსპერტების აზრი ამ თემაზე ან, სულ მცირე, გამოიყენონ არსებული სამეცნიერო რესურსები.[29]

4.2.3. პოსტ-დამუშავება და მანქანური სწავლების გაუქმება

აქტუალური გახდა მიდგომა, რომელსაც მანქანური სწავლების გაუქმება ეწოდება და ორიენტირებულია უკვე გაწვრთნილი მოდელების მოდიფიცირებაზე, რათა მათ შეძლონ სასწავლო მონაცემების კონკრეტული ნაწილების „დავიწყება". ჩვენი კვლევა ორ მთავარ მიდგომაზეა

[29] მაგ., იქვე.

ფოკუსირებული: „ზუსტი“ სწავლების გაუქმება და „მიახლოებითი“ სწავლების გაუქმება:[30]

- ზუსტი სწავლების გაუქმების მიზანია დიდი ენობრივი მოდელის მიზნობრივი სასწავლო მონაცემების სრულად ჩამოშორება მისგან, თავდაპირველი სასწავლო მონაცემების მრავალ ქვესიმრავლედ დაყოფით. როდესაც მონაცემთა წერტილები იდენტიფიცირებულია წასაშლელად, საჭიროა მხოლოდ იდენტიფიცირებულ მონაცემთა წერტილებთან დაკავშირებული ქვემოდელის ხელახლა პროგრამირება.[31] ეს აჩქარებს სისტემის ხელახალი გამოცდის პროცესს, რაც სხვა შემთხვევაში ძვირადღირებული პროცედურა იქნებოდა;
- მიახლოებითი სწავლების გაუქმება უშუალოდ მოდელზეა ორიენტირებული. შეცვლილი მონაცემებით ხელახალი მომზადების ნაცვლად, ის მოდელის მოცულობას შემდგომში არეგულირებს, რათა შეამციროს სამიზნე სწავლების მონაცემები. მიუხედავად იმისა, რომ ინფორმაციის ჩამოშორება ნაკლებად ზუსტია, იგი გარკვეულ შემთხვევებში შეიძლება ნაკლებად რთულ პროცედურას წარმოადგენდეს.

მიუხედავად იმისა, რომ მანქანური სწავლების გაუქმების მომხრეები ამბობენ, რომ ეფექტური მიდგომა, მისი შემუშავების შემთხვევაში, გააუმჯობესებს კონფიდენციალობას და ხელს შეუწყობს არაზუსტი ან მოძველებული მონაცემების არსებობას, აღნიშნული მონაცემების წაშლა არ შეიძლება უბრალოდ მათი მონაცემთა ბაზიდან წაშლით: მონაცემების გავლენა, როგორიცაა გავლენა მოდელის მოცულობაზე, ასევე უნდა მოიხსნას მანქანური სწავლების ძირითადი მოდელებიდან. გარდა ამისა, როგორც აღინიშნა, ბოლოდროინდელმა კვლევებმა აჩვენა, რომ მოდელის სასწავლო მონაცემებიდან მონაცემების კონკრეტული ფაქტორების ამოღებამ შეიძლება გამოავლინოს სხვა რისკები.[32] ამ ეტაპზე ეს სფერო ჯერ კიდევ შეუსწავლელია და არ არსებობს მკაფიო პასუხი იმის შესახებ, თუ რამდენად ეფექტიანი იქნება „მანქანური სწავლება“ მომავალში. ამ პროცესში მონაცემთა სუბიექტის უფლებების პატივისცემა კვლავაც გამოწვევაა.[33]

---

[30] იხ.: *Xu J., Wu Z., Wang C. and Jia X.*, Machine Unlearning: Solutions and Challenges, 2024, <https://arxiv.org/abs/2308.07061> [25.11.2025].

[31] See *Yan H., Li X., Guo Z. et al.*, ARCANE: An Efficient Architecture for Exact Machine Unlearning, 2022, <https://www.ijcai.org/proceedings/2022/0556.pdf> [25.11.2025].

[32] See *Carlini et al.*, *supra* note 25.

[33] *Zhang D., Finckenberg-Broman P., Hoang T. et al.*, Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions, *Algorithms that forget: Machine unlearning and the right to erasure, 2023,* <https://arxiv.org/pdf/2307.03941> [25.11.2025].

5. დასკვნა

დიდ ენობრივ მოდელებთან დაკავშირებული საკითხები პერსონალურ მონაცემთა დაცვის საზედამხედველო ორგანოებისთვის ერთგვარ გამოწვევად გადაიქცა. აღნიშნულ ნაშრომში განხილული ტექნოლოგია, სხვა ხელოვნური ინტელექტის სისტემებთან შედარებით, არა მხოლოდ რთულია, უნიკალური დეტალებითა და განვითარების ეტაპებით, არამედ წარმოშობს პირადი ცხოვრების ხელშეუხებლობისა და მონაცემთა დაცვის სხვადასხვა რისკს, რომელთა გააზრება დამოკიდებულია ტექნოლოგიის ფუნქციონირების გაგებაზე.

ნაშრომში შევეცადეთ, მოგვეწოდებინა დიდი ენობრივი მოდელების სიღრმისეული ანალიზი, რათა ადამიანები უკეთესად მოემზადონ დიდი ენობრივი მოდელების მიერ შექმნილი გამოწვევების დასაძლევად. პერსონალურ მონაცემთა დაცვის საზედამხედველო ორგანოების პირველი კონტაქტი დიდ ენობრივ მოდელებთან და მასთან დაკავშირებულ გენერირებად ხელოვნურ ინტელექტთან დაკავშირებით მხოლოდ დასაწყისია. მოსალოდნელია, რომ სფეროს განვითარებასთან ერთად გამოწვევებიც გაიზრდება.

ბიბლიოგრაფია:

1. *Askell A., Bai Y., Chen A. et al.*, A General Language Assistant as a Laboratory for Alignment, 2021, <https://doi.org/10.48550/arXiv.2112.00861> [25.11.2025].
2. *Aristotle*, *Nicomachean Ethics*, 1140b8.
3. *Bai Y., Kadavath S., Kundu S. et al.*, Constitution AI: Harmlessness from AI Feedback, December 2022, <https://arxiv.org/abs/2212.08073> [25.11.2025].
4. *Bommasani R., Hudson D., Adeli E. et al.*, On the Opportunities and Risks of Foundation Models, August 2021, <https://doi.org/10.48550/arXiv.2108.07258> [25.11.2025].
5. *Carlini N., Tramer F., Wallace E. et al.*, Extracting Training Data from Large Language Models, *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650, <https://arxiv.org/abs/2012.07805> [25.11.2025].
6. *Center for Countering Digital Hate*, Misinformation on Bard, Google's New AI Chat, April 5, 2023, <https://counterhate.com/research/misinformation-on-bard-google-ai-chat/> [25.11.2025].
7. *Center for Strategic and International Studies*, A Conversation on Cybersecurity with NSA's Rob Joyce*, YouTube*, April 11, 2023, <https://youtu.be/MMNHNjKp4Gs?t=530> [25.11.2025].
8. *Christiano P., Leike J., Brown T. et al.*, Deep Reinforcement Learning from Human Preferences, 2017, <https://arxiv.org/abs/1706.03741> [25.11.2025].
9. *Davis E. and Sundstrom A.*, Experiment with GALACTICA, 2022, <https://cs.nyu.edu/~davise/papers/ExperimentWithGalactica.html> [25.11.2025].

10. *David E.*, Getty lawsuit against Stability AI to go to trial in the UK, *The Verge*, December 4, 2023, <https://www.theverge.com/2023/12/4/23988403/getty-lawsuit-stability-ai-copyright-infringement> [25.11.2025].
11. *Dwork C., McSherry F., Nissim K. and Smith A.*, Calibrating noise to sensitivity in private data analysis, *Procedures of the Third Conference on Theory of Cryptography (TCC)*, 265–284, <http://dx.doi.org/10.1007/11681878_14> [25.11.2025].
12. *Fiesler C. and Proferes N.*, 'Participant' Perceptions of Twitter Research Ethics, *Social Media + Society,* Vol. 4(1), 2018, <https://doi.org/10.1177/2056305118763366> [25.11.2025].
13. *Global Privacy Assembly (GPA) International Enforcement Cooperation Working Group*, Joint statement on data scraping and the protection of privacy, August 2023, <https://ico.org.uk/media/about-the-ico/documents/4026232/joint-statement-data-scraping-202308.pdf> [25.11.2025].
14. *Groll E.*, ChatGPT Shows Promise of Using AI to Write Malware*, CyberScoop*, December 6, 2022, <https://cyberscoop.com/chatgpt-ai-malware/> [25.11.2025].
15. *Hassold C.*, Executive Impersonation Attacks Targeting Companies Worldwide, *Abnormal Blog*, February 16, 2023, <https://abnormalsecurity.com/blog/midnight-hedgehog-mandarin-capybara-multilingual-executive-impersonation> [25.11.2025].
16. *Hines K.*, OpenAI Launces GPTBot With Details on How to Restrict Access, *Search Engine Journal*, Aug. 7, 2023, <https://www.searchenginejournal.com/openai-launches-gptbot-how-to-restrict-access/493394/> [25.11.2025].
17. *IWGDPT*, *Working Paper on Large Language Models (LLMs)*, December 27, 2024, <https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Berlin-Group/20241206-WP-LLMs.pdf?__blob=publicationFile&v=2> [25.11.2025].
18. *Karpathy A.*, Let's build GPT: from scratch, in code, spelled out, January 2023, <https://www.youtube.com/watch?v=kCc8FmEb1nY> [25.11.2025].
19. *Knight W.*, A Stanford Proposal Over AI's 'Foundations' Ignites Debate, *Wired*, September 2021, <https://www.wired.com/story/stanford-proposal-ai-foundations-ignites-debate/> [25.11.2025].
20. *Khandelwal U., Clark K., Jurafsky D. et al.*, Sample Efficient Text-Summarization Using a Single Pre-Trained Transformer, 2019, <https://doi.org/10.48550/arXiv.1905.08836> [25.11.2025].
21. *Marcus G. and Davis E.*, Large Language Models like ChatGPT say The Darnedest Things, 2023, <https://garymarcus.substack.com/p/large-language-models-like-chatgpt> [25.11.2025].
22. *Perrigo B.*, Open AI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic, *Time*, January 18, 2023, <https://time.com/6247678/openai-chatgpt-kenya-workers/> [25.11.2025].

23. *Ponomareva N., Vassilvitskii S., Xu Z. et al.*, How to DP-fy ML: A Practical Tutorial to Machine Learning with Differential Privacy, *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23*), 2023, 5823–5824, <https://arxiv.org/abs/2303.00654> [25.11.2025].
24. *Schaul K. et al*., Inside the secret list of websites that make AI like ChatGPT sound smart, *Washington Post*, Apr. 19, 2023, <https://www.washingtonpost.com/technology/interactive/2023/ai-chatbot-learning/> [25.11.2025].
25. *Silver D., Huang A., Maddison C. et al.*, Mastering the game of Go with deep neural networks and tree search, *Nature*, Vol. 529, 2016, 484–489, <https://doi.org/10.1038/nature16961> [25.11.2025].
26. *SlashNext*, The State of Phishing 2023, *SlashNext Security*, Oct. 2023, <https://slashnext.com/state-of-phishing-2023/> [25.11.2025].
27. *Stiennon N., Ouyang L., Wu J. et al.*, Learning to summarize from human feedback, *34th Conference on Neural Information Processing Systems* (NeurIPS 2020), <https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf> [25.11.2025].
28. *Taylor R., Kardas M., Cucurell G. et al.*, Galactica: A Large Language Model for Science, 2022, <https://arxiv.org/abs/2211.09085> [25.11.2025].
29. *Veale M., Binns R. and Edwards L.*, Algorithms that remember: model inversion attacks and data protection law, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, October 2018, <https://doi.org/10.1098/rsta.2018.0083> [25.11.2025].
30. *Vinyals O., Babuschkin I., Czarnecki W. et al.*, Grandmaster level in StarCraft II using multi-agent reinforcement learning, *Nature*, vol. 575, 2019, 350–354, <https://doi.org/10.1038/s41586-019-1724-z> [25.11.2025].
31. *Volokh E.*, Large Libel Models: ChatGPT-3.5 Erroneously Reporting Supposed Felony Pleas, Complete with Made-Up Media Quotes?, *Reason*, 2023, <https://reason.com/volokh/2023/03/17/large-libel-models-chatgpt-4-erroneously-reporting-supposed-felony-pleas-complete-with-made-up-media-quotes/> [25.11.2025].
32. *Xu J., Wu Z., Wang C. and Jia X.*, Machine Unlearning: Solutions and Challenges, 2024, <https://arxiv.org/abs/2308.07061> [25.11.2025].
33. *Yan H., Li X., Guo Z. et al.*, ARCANE: An Efficient Architecture for Exact Machine Unlearning, 2022, <https://www.ijcai.org/proceedings/2022/0556.pdf> [25.11.2025].
34. *Zhang D., Finckenberg-Broman P., Hoang T. et al.*, Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions, *Algorithms that forget: Machine unlearning and the right to erasure*, 2023, <https://arxiv.org/pdf/2307.03941> [25.11.2025].
35. *Ziegler D., Stiennon N., Wu J. et al.*, Fine-Tuning Language Models from Human Preferences, 2019, <https://arxiv.org/pdf/1909.08593.pdf> [25.11.2025].