# COGNITIVE LANGUAGE MODELS IN LINGUISTICS

## კოგნიტური ენობრივი მოდელები ლინგვისტიკაში

**Tinatin Mshvidobadze**
Doctor of Technical Sciences,
Associate Professor of Gori State University,
Gori, Chavchavadze st. #53, 1400, Georgia,
+995555118379 tinikomshvidobadze@gmail.com
https://orcid.org/0000-0003-3721-9252

**Abstract.** This paper assesses the fundamental utility of language models in the social sciences. A simulation perspective is presented, in which language models offer a new paradigm for role simulation and modeling of cognitive processes. We discuss limitations and considerations of internal, external, construct, and statistical validity, and provide methodological recommendations for the effective integration of language models into psychological research. This perspective reexamines the role of language models in behavioral and cognitive science and sheds light on the similarities and differences between machine intelligence and human cognition and thought.

**Keywords:** Generative Artificial Intelligence (GenAI), Large Language Models (LLM), Simulation, Modeling.

**თინათინ მშვიდობაძე**
ტექნიკურ მეცნიერებათა დოქტორი,
გორის სახელმწიფო უნივერსიტეტის ასოცირებული პროფესორი,
ქ. გორი, ჭავჭავაძის ქ. #53, 1400, საქართველო,
+995555118379 tinikomshvidobadze@gmail.com
https://orcid.org/0000-0003-3721-9252

**აბსტრაქტი.** ნაშრომში ვაფასებთ ენობრივი მოდელების ფუნდამენტურ სარგებლიანობას სოციალურ მეცნიერებებში. მოცემულია სიმულაციის პერსპექტივა, რომელშიც ენობრივი მოდელები გვთავაზობენ ახალ პარადიგმას როლების სიმულაციისა და კოგნიტური პროცესების მოდელირებისათვის. განვიხილავთ შინაგან, გარე, კონსტრუქციულ და სტატისტიკური ვალიდურობის შეზღუდვებსა და მოსაზრებებს, ვაძლევთ მეთოდოლოგიურ რეკომენდაციებს ენობრივი მოდელების ფსიქოლოგიურ კვლევაში ეფექტური ინტეგრაციისათვის. ეს პერსპექტივა ხელახლა განიხილავს ენობრივი მოდელების როლს ქცევით და კოგნიტურ მეცნიერებაში და ნათელს ჰფენს მანქანურ ინტელექტსა და ადამიანის შემეცნებასა და აზროვნებას შორის მსგავსებებსა და განსხვავებებს.

**საკვანძო სიტყვები:** გენერაციული ხელოვნური ინტელექტი (GenAI), დიდი ენობრივი მოდელები (LLM), სიმულაცია, მოდელირება.

**შესავალი.** ბოლო წლებში კოგნიტური ლინგვისტიკის თეორიები საკმარისად დახვეწილი და დეტალური გახდა, რომელთა შემოწმება შესაძლებელია კოგნიტური მეცნიერებების კონვერგენტული მეთოდების ფართო სპექტრის გამოყენებით. ენისა და შემეცნების ნეირონული საფუძველი დიდი ხანია გავლენას ახდენს კოგნიტური ლინგვისტიკის თეორიების ბუნებასა და შინაარსზე.

დიდი ენობრივი მოდელების (LLM). გამოყენების იდეა ასახავს ხელოვნური ინტელექტის უფრო ფართო ნარატივს, რომლის მიზანია ადამიანების ხელოვნური ანალოგებით ჩანაცვლება ავტონომიური ხელოვნური ინტელექტის სისტემების შექმნით, რომლებიც „ადამიანებს გაუსწრებს ყველაზე ღირებული სამუშაოს შესრულებისას“. (Dillion, 2023: 597-600). ამიტომ, განსაკუთრებით მნიშვნელოვანია ჩანაცვლების პერსპექტივისა და მისი ძირითადი ვარაუდების საფუძვლიანი შეფასება, ენობრივი მოდელების სათანადოდ განთავსება ქცევითი და სოციალური მეცნიერებების კვლევის ლანდშაფტში.

ღია კოდის ენობრივი მოდელები წარმოადგენს კოგნიტურ მოდელებს, რომლებიც ხელმისაწვდომია და მართვადია, რაც მკვლევრებს საშუალებას აძლევს გამოიკვლიონ, ჩაერიონ, დააკვირდნენ და გაზომონ ქცევები, რომელთა უშუალოდ შესწავლა ადამიანებში სხვაგვარად არაპრაქტიკულია. (Zhang, et al. 2022: 168).

მართლაც, დიდი ხანია გამოიყენება გამოთვლით ინსტრუმენტები გონებრივი პროცესების აბსტრაქციის წარმოსადგენად, როგორც საიმონი აღნიშნავს: „კომპიუტერების გამოყენება ადამიანების სიმულირებისთვის, რათა გავიგოთ, თუ როგორ მუშაობენ ადამიანები“. (Simon, 1983: 23-27).. თუმცა, ტრადიციულ სიმბოლურ ხელოვნურ ინტელექტთან ან სხვა ნეირონულ ქსელურ მოდელებთან (მაგ., CNN) შედარებით, ტრანსფორმატორზე დაფუძნებული მოდელები ავლენს მჭიდრო მახასიათებლებსა და შესაძლებლობებს, ადამიანის შემეცნებასთან დაკავშირებით.

გარდა ამისა, ფსიქოლოგიური კვლევის მონაცემებზე დაფუძნებული მოდელების დახვეწამ შეიძლება მოგვცეს ადამიანის ქცევის კიდევ უფრო ზუსტი წარმოდგენები, რომლებიც კოგნიტური დამუშავების ძირითად ასპექტებს ასახავს ( Buckner, 2023: 681-712).

ერთ-ერთი მიდგომაა ბუნებრივი ენის მოდელების შიდა მდგომარეობების, სტრუქტურების ან წარმოდგენების შესწავლა, კონკრეტული ენობრივი თვისებების ან მახასიათებლების ამოსაცნობად (მაგ., სინტაქსი, სემანტიკა) (Belinkov, 2022: 207-219). ეს ავლენს, თუ როგორ არის წარმოდგენილი და განაწილებული ინფორმაცია დონეებს შორის.

ენობრივი მოდელები ავსებენ სხვა ტიპის ხელოვნურ ნეირონულ ქსელებსაც, როგორიცაა CNN - გონების და ტვინის მოდელირებისათვის, ვიზუალური ობიექტის ამოცნობიდან სმენით აღქმამდე. (Kanwisher et al, 2023: 240-254).

მაგალითად, ვიზუალურ-სემანტიკური წარმოდგენები, როგორიცაა CLIP (კონტრასტული ენისა და სურათის წინასწარი მომზადება), ეს OpenAI-ს მიერ შემუშავებული მულტიმოდალური მოდელია, რომელიც სწავლობს სურათებისა და ტექსტის დაკავშირებას. ტრადიციულ, მხოლოდ მხედველობაზე ორიენტირებულ მოდელებთან შედარებით, CLIP უფრო ზუსტად პროგნოზირებს ტვინის აქტივობას ვიზუალური აღქმის საპასუხოდ.(Tenney, 2019: 250).

თუმცა, მიუხედავად იმისა, რომ ანალიზის მეთოდები ენობრივი მოდელების შიდა წარმოდგენების შესახებ ინფორმაციას გვაწვდის, მათ აქვთ საკუთარი შეზღუდვები, როგორც პრაქტიკული, ასევე კონცეპტუალური. რომელთაგან, ყველაზე კრიტიკული მიზეზობრიობის არარსებობაა: „კლასიფიკაციის ან დეკოდირების მაღალი შესრულება შეიძლება ასახავდეს არა მოდელის ფუნქციურ გამოყენებას მისი ძირითადი ამოცანებისათვის, არამედ ანალიზის მეთოდის უნარს, ამოიღოს ინფორმაცია. (მაგ., ზედაპირული კორელაციები)“. (Wang, 2024: 108).

პრაქტიკული შეზღუდვები მოიცავს იმ ფაქტს, რომ ღია კოდის მოდელები, როგორიცაა *Llama* და *OPT,* ჯერ არ შეესაბამება წამყვანი დახურული კოდის მოდელების შესაძლებლობებს. გარდა ამისა, მასშტაბური ექსპერიმენტების ჩატარება და ენობრივი მოდელების, განსაკუთრებით უფრო ქმედითი, დიდი მოდელების დახვეწა მოითხოვს გამოთვლით რესურსებს, რომლებიც შეიძლება იყოს შეუზღუდავი. კონცეპტუალური შეზღუდვები ეხება იმას, ასახავს თუ არა ენობრივ მოდელებში მონაცემთა ოპტიმიზაცია ნამდვილად ადამიანების მიერ განცდილ ამოცანების ოპტიმიზაციას. კერძოდ, იმ ამოცანებს, რომლებიც ჩამოყალიბებულია გადარჩენისა და რეპროდუქციის შედეგად ბუნებრივი გადარჩევის გზით.

**მეთოდები**. მოცემულ სტატიაში გამოყენებულია აღწერითი, ანალიზისა და განმარტების მეთოდები, რის საფუძველზეც გამოკვეთილია აღნიშნული კვლევის მნიშვნელოვანი საკითხები. კონცეფციის ჩამოსაყალიბებლად დავიმოწმეთ სხვადასხვა მკვლევარების შეხედულებები, მათ საფუძველზე მოვახდინეთ მსჯელობისა და დასკვნების ილუსტრირება.

ნაშრომში მოცემულია ყველაზე ხშირად გამოყენებული მეთოდებისა და ფორმალური ჩარჩოების მიმოხილვა კომპიუტერული სიმულაციური ექსპერიმენტების შესახებ ფონეტიკისა და ფონოლოგიის სფეროებში, როგორიცაა ძირითადად გამოთვლითი მოდელირებისა და მანქანათმცოდნეობის მეთოდები. აღნიშნული მეთოდები, ძირითადად გამოიყენება სინთეზურ მონაცემებთან, სადაც წარმოქმნიან უამრავ გადაწყვეტილებას.

**შედეგები და მსჯელობა.** ენობრივი მოდელების დამუშავებისას მისი დონეების კვლევა შესაძლებელია იმის გასაგებად, თუ როგორ წარმოადგენენ ისინი შინაგანად სხვადასხვა სინტაქსურ და სემანტიკურ სტრუქტურებს.

ადრეული დონეები, წარმოადგენენ დაბალი დონის სინტაქსურ მახასიათებლებს (მაგ., მეტყველების ნაწილების ტეგები, რომლებიც თითოეული სიტყვისათვის მინიჭებული განმარტებებია, მათი გრამატიკული ფუნქციის აღსანიშნავად, როგორიცაა არსებითი სახელი. ზმნა და წინდებული). ხოლო გვიანდელი დონეები კოდირებას ახდენენ უფრო რთული სემანტიკური ურთიერთობების (Tenney et al, 2019: 250). ეს ხელს უწყობს კონკრეტული ცენტრების ლინგვისტურ ამოცანებთან მიბმას, რაც ნათელს ჰფენს, თუ როგორ მიმდინარეობს ბუნებრივი ენის დამუშავების პროცესი (Manning et al, 2020: 346-354).

იმდენად, რამდენადაც ენობრივი მოდელები ენის შესწავლისა და დამუშავების ცალსახა რეალიზაციას უზრუნველყოფენ, ისინი გამოიყენება ტექსტთან დაკავშირებული ქცევების, უპირატესად ენობრივი ფენომენების გასაგებად.

განვიხილავთ სამ ძირითად განზომილებას, რომლებიც აფასებენ ენობრივი მოდელების გამოყენების სამეცნიერო და ფილოსოფიურ ვალიდურობას სოციალურ და ქცევით კვლევაში: დახასიათება (ინტელექტის ბუნება LLM-ებში), ინტერპრეტაცია (რას გვეუბნება ადამიანებსა და ენობრივი მოდელებს შორის შეთანხმება მათი მექანიზმების შესახებ) და სარგებლიანობა (ენობრივი მოდელების გამოყენების შეზღუდვები და სარგებელი კვლევაში). (ცხრილი 1.).

| განზომილება | ენობრივი მოდელები, როგორც ადამიანის შემცვლელი | ენობრივი მოდელები, როგორც ლინგვისტური სიმულატორები | ძირითადი არგუმენტი | პოტენციური მცდარი წარმოდგენა |
|---|---|---|---|---|
| **დახასიათება** | ტექნოლოგია, რომელიც ადამიანის აზროვნებას აგენერირებს | მრავალმხრივი ინსტრუმენტი, რომელიც სხვადასხვა პერსპექტივების სიმულირებას ახდენს | ენობრივი მოდელები ოპტიმიზაციას უკეთებენ ციფრული ერთეულის პროგნოზირებას, რაც ბიოლოგიური ინტელექტისაგან განსხვავებულ მანქანურ ინტელექტს ქმნის | ციფრული ერთეულის პროგნოზირება, როგორც ადამიანის ინტელექტის მცდარი წარმოდგენა |
| **ინტერპრეტაცია** | კორელაციული შესაბამისობა ადამიანებსა და ენობრივი მოდელებს შორის | ენობრივი მოდელები ადამიანის შემეცნების ასპექტების მოდელირებას ახდენენ, თუმცა ფუნდამენტურად განსხვავდებიან | არქიტექტურული შესაბამისობა | გასწორება |
| **სარგებლიანობა** | ენობრივი მოდელები შეიძლება იყოს ადამიანის გონების პირდაპირი გამოვლენის ძირითადი ინსტრუმენტი | ენობრივი მოდელები დამატებითი ინსტრუმენტებია, რომლებსაც ძირითადი შეზღუდვები აქვთ დახურული მონაცემების გამო | ენობრივი მოდელების პასუხები უნდა დადასტურდეს რეალური მონაცემებით | ჩანაცვლების მცდარი წარმოდგენა |

ცხრილი 1 | სოციალურ და ქცევით მეცნიერებებში ენობრივი მოდელების კონცეპტუალიზაცია, როგორც მონაწილეთა ჩანაცვლება სიმულატორების საშუალებით

ამრიგად, ენობრივი მოდელები ქმნიან ფონს იმის შესასწავლად, თუ როგორ შეიძლება ენობრივი მოდელები გამოყენებულ იქნას, როგორც როლური თამაშების ინსტრუმენტები და კოგნიტური მოდელები სხვადასხვა სიმულაციისთვის და გონებრივი პროცესების უკეთ გასაგებად. (Tao et al, 2024: 346).

იმისათვის, რომ ენობრივი მოდელები ქცევისა და შემეცნების წარმოდგენის სანდო ინსტრუმენტებად ჩაითვალოს, აუცილებელია პირველ რიგში მათი ვალიდურობის განხილვა. როგორც ნაჩვენებია (ცხრილი 2.).

შინაგანი ვალიდურობა გულისხმობს დამოუკიდებელ და დამოკიდებულ ცვლადებს შორის მიზეზობრივი კავშირების დადგენას, იმის უზრუნველყოფას, რომ ენობრივი მოდელების პასუხებში ცვლილებები გამოწვეულია მოთხოვნებში მანიპულაციებით და არა სხვადასხვა გარე ფაქტორებით.

გარე ვალიდურობა გულისხმობს შედეგების განზოგადებას შესწავლილი კონტექსტის მიღმა - შესაძლებელია თუ არა სიმულაციების განზოგადება რეალურ ადამიანურ შემცნებაზე, რაც ეჭვქვეშ დგება მონაცემებისა და ალგორითმების შეზღუდვებით და მოითხოვს ბუნებრივი მონაცემებით ვალიდაციას.

და ბოლოს, სტატისტიკური ინფერენციული ვალიდურობა ეხება დასკვნების დადასტურებას მონაცემებით, რაც ხაზს უსვამს შესაბამისი ნიმუშებისა და ანალიზების მნიშვნელობას. პოტენციური საფრთხეები წარმოიშობა არადამოუკიდებელი პასუხებიდან და მცირე ნიმუშებიდან, რომლებიც საჭიროებენ სიფრთხილის ზომებს, როგორიცაა თანმიმდევრულობის დაცვა.

| **ვალიდურობის ტიპი** | **განმარტება** | **ვალიდურობის საფრთხეები** |
|---|---|---|
| შინაგანი ვალიდურობა | ენობრივი მოდელები ქცევისა და რეაქციების ცვლილებები ახდენს კონკრეტული მოთხოვნის მანიპულაციებს ან მოდელის კონფიგურაციებს | **მოდელის ცვლადები:** უკონტროლო ფაქტორები, როგორიცაა შემთხვევითობა. **მოდელის პარამეტრები:** შეუძლიათ გავლენა მოახდინონ პასუხებზე, რაც გავლენას ახდენს ა ექსპერიმენტებზე. **მოთხოვნის ცვლადები:** მოთხოვნის ფორმულირებაში, სტრუქტურაში ან კონტექსტში მცირე ვარიაციებმა შეიძლება გამოიწვიოს პასუხებში განსხვავებები |
| გარეგანი ვალიდურობა | გარკვეული ხარისხით შესაძლებელია ენობრივი მოდელების მიერ გენერირებული სიმულაციების განზოგადება ადამიანის ქცევასა და შემეცნებაზე მიზნობრივ კონტექსტში | **მონაცემების მიკერძოებები და შეზღუდვები:** მონაცემები შეიძლება ზუსტად არ წარმოადგენდეს კონკრეტულ პოპულაციებს/ქვეჯგუფებს, კულტურებს. **კონტექსტთან დაკავშირებული შეზღუდვები:** კონტროლირებად გარემოში ჩატარებული სიმულაციები შეიძლება არ იყოს რეალური სამყაროს სცენარებიასთვის გამოყენებადი |
| კონსტრუქციული ვალიდურობა | რამდენად ზუსტად წარმოაჩენენ ენობრივი მოდელების სიმულაციები ადამიანის ქცევის თეორიულ | **შეუსაბამობა**: ფსიქოლოგიური კონსტრუქტები მრავალმხრივი და სუბიექტურია, პოტენციურად სრულად ან ზუსტად არ არის აღქმული |

| | | |
|---|---|---|
| | კონსტრუქტებს და შემეცნებას, რომელთა მოდელირებასაც ისინი აპირებენ | ტექსტზე დაფუძნებული პასუხებით. **ნამდვილი აღქმის ნაკლებობა:** ბუნებრივი ენის მოდელები არ არიან დაფუძნებული რეალურ კონტექსტზე და არ შეუძლიათ დროის განცდა, მათი პასუხები კი დაფუძნებულია საცდელ მონაცემებზე, რეალური კოგნიტური პროცესების გარეშე. |
| სტატისტიკური ინფერენციული ვალიდურობა | ენობრივი მოდელების მიერ გენერირებული მონაცემების ანალიზიდან გამოტანილი დასკვნების შესაბამისი სტატისტიკური მეთოდებით დადასტურების ხარისხი | **ნიმუშის შეზღუდვები:** სიმულირებული პასუხების შეზღუდულმა რაოდენობამ შეიძლება შეამციროს სტატისტიკური ტესტების სიმრავლე და პასუხები შეიძლება არ იყოს დამოუკიდებელი წინა ურთიერთქმედებების გავლენისაგან |

ცხრილი 2 | ენობრივი მოდელების გამოყენების ვალიდურობა როლების სიმულირებისა და კოგნიტური პროცესების მოდელირებისთვის

ეს ფაქტორები და ვალიდურობის პოტენციური საფრთხეები, შეიძლება საფრთხეს უქმნიდეს ადამიანის ქცევისა და შემეცნების სიმულაციების საფუძველზე სათანადო დასკვნების გამოტანის მიზნით ენობრივი მოდელებიის გამოყენების სანდოობას.

ამრიგად, ისინი ქმნიან ფონს იმის შესასწავლად, თუ როგორ შეიძლება ენობრივი მოდელები გამოყენებულ იქნას, როგორც როლური თამაშების ინსტრუმენტები სხვადასხვა სიმულაციისათვის და როგორც კოგნიტური მოდელები გონებრივი პროცესების უკეთ გასაგებად.

**დასკვნა.** ადამიანის დონის ხელოვნური ინტელექტის ბოლოდროინდელი მიღწევები აცოცხლებს კლასიკურ დებატებს გამოთვლითი არტეფაქტების როლის შესახებ ადამიანის გონებისა და შემეცნების გაგებისას.

ჩვენ შევაფასეთ ახალი ხედვა, რომელიც გულისხმობს ქცევით და სოციალურ მეცნიერებებში ადამიან/მონაწილეების ჩანაცვლებას ბუნებრივი ენის მოდელებით.

ადამიანის ინტელექტი არ არის უბრალოდ ტექსტის დამუშავებისა და ნიშნების პროგნოზირების თანმდევი პროდუქტი. ის დაფუძნებულია სენსორულ გამოცდილებაზე, გამდიდრებულია მულტიმოდალური ინტეგრაციით და ყალიბდება სუბიექტურობით. როგორც ფილოსოფოსმა მორის მერლო-პონტიმ აღნიშნა: „სხეული ჩვენი ზოგადი საშუალებაა სამყაროს შესაქმნელად“, ჭეშმარიტი გაგება წარმოიშობა განსახიერებული გამოცდილებიდან, რაც ბუნებრივი ენის მოდელებს არ გააჩნიათ. (Merleau-Ponty, 2012: 24).

ნაშრომში წარმოდგენილი ვალიდურობის სამი ასპექტის (ცხრილი 2) ბუნებრივი ენის მოდელების გამოყენების საუკეთესო პრაქტიკის გამოწვევების გათვალისწინებით. არგუმენტების ანალიზი მხარს უჭერს სიმულაციის პერსპექტივას: ბუნებრივი ენის მოდელები არის როლების სიმულაციისა და კოგნიტური პროცესების მოდელირების

ინსტრუმენტები, რომლებიც ავსებენ, მაგრამ არ ცვლიან ადამიანებს - ისევე, როგორც მუსიკალური ნოტები იპყრობს რიტმებს, მაგრამ ვერ იმეორებს ემოციასა და ინტერპრეტაციას, რომელიც მუსიკოსს შემოაქვს შესრულებისას.

სიმულაცია საშუალებას გვაძლევს შევისწავლოთ კითხვები, რომელთა გადაჭრაც შეიძლება რთული ან მოუხერხებელი იყოს ტრადიციული მეთოდების გამოყენებით. (Misra, 2024: 145). რეალური სამყაროს მონაცემებით ვალიდაცია კრიტიკულად მნიშვნელოვანია სიმულაციების სიზუსტისა და განზოგადების დასადგენად.

საბოლოო ჯამში, ენობრივი მოდელები ხელს უწყობს ენის საფუძვლად მყოფი ფსიქოლოგიისა და შემეცნების შესწავლის ახალ პარადიგმას. თავის მხრივ, ისინი ასევე აღრმავებენ ჩვენს შემეცნებას თავად ენობრივი მოდელების შესახებ, რაც მათ უფრო ინტერპრეტირებადს და ახსნადს ხდის. ეს პერსპექტივა გვაიძულებს გადახედოთ ხელოვნური ინტელექტის როლს ქცევით და კოგნიტურ მეცნიერებაში, რომლის მეშვეობითაც უკეთ შეგვიძლია გავიგოთ მსგავსებები და განსხვავებები ადამიანის ინტელექტსა და მანქანურ ინტელექტს შორის.

**გამოყენებული ლიტერატურა**

Buckner, C. (2023). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *Br. J. Philos. Sci.* 74, 681-712.

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist.* 48, 207-219.

Dillion, D., Tandon, N., Gu, Y. & Gray, K. (2023). Can AI language models replace human participants? *Trends Cogn. Sci*. 27, 597-600.

Kanwisher, N., Khosla, M. & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci*. 46, 240-254.

Misra, K. & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: the case of the missing AANNs. *arXiv:2403*.19827. p. 145.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. U. S. A*. 117, p.346-354.

Merleau-Ponty M. (2012). The Phenomenology of Perception. Transl. ed Landes. D. A. *London; New York: Routledge*. p. 24.

Simon, H. A. (1983). in Machine Learning (eds Ryszard S. Michalski, Jaime G. Carbonell, & Tom M. Mitchell) *Morgan Kaufmann*, p. 25-37.

Tenney, I., Das, D. & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv:1905*. P. 250.

Tao, Y., Viberg, O., Baker, R. S. & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, p. 346.

Wang, A., Morgenstern, J. & Dickerson, J. P. (2024). Large language models cannot replace human participants because they cannot portray identity groups. *arXiv:2402*.019 p. 108.

Zhang, S. et al. (2022). OPT: open pre-trained transformer language models. *arXiv:4502*.01. p. 168.

**REFERENCES**

Buckner, C. (2023). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *Br. J. Philos. Sci.* 74, 681-712.

Belinkov, Y. (2022). Probing classifiers: promises, shortcomings, and advances. *Comput. Linguist.* 48, 207-219.

Dillion, D., Tandon, N., Gu, Y. & Gray, K. (2023). Can AI language models replace human participants? *Trends Cogn. Sci*. 27, 597-600.

Kanwisher, N., Khosla, M. & Dobs, K. (2023). Using artificial neural networks to ask 'why' questions of minds and brains. *Trends Neurosci*. 46, 240-254.

Misra, K. & Mahowald, K. (2024). Language models learn rare phenomena from less rare phenomena: the case of the missing AANNs. *arXiv:2403*.19827. p. 145.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proc. Natl. Acad. Sci. U. S. A*. 117, p.346-354.

Merleau-Ponty M. (2012). The Phenomenology of Perception. Transl. ed Landes. D. A. *London; New York: Routledge*. p. 24.

Simon, H. A. (1983). in Machine Learning (eds Ryszard S. Michalski, Jaime G. Carbonell, & Tom M. Mitchell) *Morgan Kaufmann*, p. 25-37.

Tenney, I., Das, D. & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. *arXiv:1905*. P. 250.

Tao, Y., Viberg, O., Baker, R. S. & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus* 3, p. 346.

Wang, A., Morgenstern, J. & Dickerson, J. P. (2024). Large language models cannot replace human participants because they cannot portray identity groups. *arXiv:2402*.019 p. 108.

Zhang, S. et al. (2022). OPT: open pre-trained transformer language models. *arXiv:4502*.01. p. 168.