VI. IKERGAZTE
NAZIOARTEKO IKERKETA EUSKARAZ

2025eko maiatzaren 28, 29 eta 30a
Bilbo, Euskal Herria

ANTOLATZAILEA:
Udako Euskal Unibertsitatea (UEU)

INGENIARITZA ETA ARKITEKTURA

Modalitate anitzeko ereduen eta hizkuntza ereduen zentzumen espazialaren azterketa

Oier Ijurco, Oier Lopez de Lacalle eta *Gorka Azkune*

165-172 or.
https://dx.doi.org/10.26876/ikergazte.vi.03.20

Modalitate anitzeko ereduen eta hizkuntza ereduen zentzumen espazialaren azterketa

Oier Ijurco, Oier Lopez de Lacalle, Gorka Azkune

HiTZ Zentroa - Ixa taldea, Euskal Herriko Unibertsitatea UPV/EHU

oier.ijurco@ehu.eus

Laburpena

Zentzumen espaziala, espazio eta harreman fisikoari buruzko gure ezagutza, giza kognizioren funtsezko ezaugarria da. Ikerketa honek gaur egungo hainbat eredu ebaluatu eta alderatzen ditu, zentzumen espazialari dagokionez. Hiru eredu mota hartzen dira kontuan: testu hutseko ereduak, sarrera bezala testua erabiltzen duten modalitate anitzeko ereduak, eta testutik abiatuta irudiak sortzen dituzten ereduak. Zentzumen espaziala gizakiontzat ulertzeko oso erraza den alderdi bat bada ere, hizkuntza ereduentzat zeregin zaila da. Ikerketa honen emaitza gisa, aurkitu dugu egungo ereduek nabarmen hobetu dutela horrelako zereginetan BERT edo RoBERTa bezalako eredu ez hain berriekin alderatuta. Azken finean, ikerketa honekin espazio-harremanei buruzko arrazoiketan adimen artifizialeko sistemen garapena azpimarratzen dugu, mundua giza mailan ulertzeko funtsezko urratsa dena.

Hitz gakoak: Adimen Artifiziala, Zentzumen espaziala, Hizkuntza Ereduak, Modalitate Anitzeko Ereduak

Abstract

Understanding spatial commonsense, our knowledge about physical space and relationships, is a fundamental aspect of human cognition. This investigation evaluates and compares various current models to determine their understanding of spatial commonsense reasoning. Three types of models are taken into account in this work: text-only models, multimodal models with text inputs, and text-to-image models generating visual representations from text. Spatial commonsense is something very simple to understand for humans, but it has been a tough task for language models in the past. Our research shows that current models have improved significantly at these kinds of tasks. Ultimately, this work contributes to the advancing development of AI systems in reasoning about spatial relationships, which is an essential step toward human-level understanding of the world.

Keywords: *Artificial Intelligence, Spatial Commonsense, Language Models, Multimodal Models*

1 Sarrera eta motibazioa

Ulermen espaziala objektuei erreferentzia egiteko gaitasuna da hauen erlazio espaziala kontuan hartuz. Erlazio espazialak, inplizituak edo esplizituak izan daitezke. Erlazio espazial esplizituetan, objektuen arteko lotura argi adierazten da. Aldiz, inplizituetan, lotura testuinguruaren bidez ulertzen da. Adibidez, erlazio espazial esplizitu bat *Katu bat mahai baten gainean* esanez adierazi daiteke. Bertan, katuak mahaiarekiko duen posizioa argi definitzen da. Aldiz, erlazio espazial inplizitu bat ondorengo esaldian ikus daiteke: *Emakume bat autoa gidatzen*, non ulertzen den emakumea auto barruan dagoela nahiz eta hori zuzenean ez adierazi.

Arrazoimen espaziala funtsezkoa da informazio espaziala ulertzeko eta prozesatzeko, eta gaitasun horiek beharrezkoak dira eguneroko bizitzako oinarrizko zeregin kognitiboak kudeatzeko (Pellegrino *et al.*, 1984). Hizkuntza eredu handiek (HEH) zentzuzko arrazoiketa atazetan urte luzez errendimendu handia lortu badute ere, ez dira gai izan ataza espazialei modu eraginkorrean aurre egiteko, gure mundu fisikoari buruzko ulermen eta arrazoiketa gaitasuna falta baitzaie (Liu *et al.*, 2022).

Zentzumen espazialaren arrazoiketa oso sinplea da gizakientzat, eta informazio hori intuizioz erabiltzen dugu eguneroko bizitzan. Nahiz eta hau oso sinplea izan gizakientzat, oso zaila izan daiteke hizkuntza ereduentzat (HE), informazio hau, normalean, ez baita esplizitu egiten testu idatzian. Arazo honi ingelesez **reporting bias** (Gordon eta Van Durme, 2013) bezala ezagutzen da. Fenomeno honen ondorioz, testua idaztean ezohikoa dena

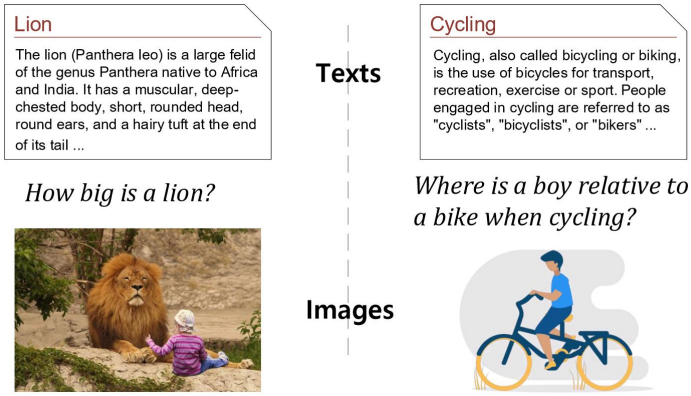
azpimarratzen dugu arruntagoa denari garrantziarik eman gabe.

Azpimarratze horrek esan nahi du testu-datuekin entrenatutako ereduek errealitatearen irudikapen alboratuak ikas ditzaketela ezohiko gertaerak modu esplizituagoan adierazten direlako. Ondorioz, eredu horiek munduari buruzko ezagutza inplizitua ulertzea eskatzen duten zereginekin zailtasunak dituzte.

Lan honetan aztertu dugun ataza bitan bana daiteke, objektu pareak edo objektuen arteko egoera bat emanik, objektuen arteko altuera eta tamaina erlatiboa zehaztea alde batetik, eta posizio erlazioak zehaztea bestetik. Aztertutako ataza horiek zentzumen espaziala eskatzen dute zuzen bete ahal izateko, eta lehen aipatu bezala, oso zaila izan daiteke hainbat hizkuntza ereduentzat.

Informazio espaziala nabarmenagoa da irudietan (Cui *et al.*, 2020), testu hutsean baino. 1. irudian ikusten denez, idatzitako bi testuek ez dute informazio espazial askorik ematen. Aldiz, irudiek lehoiaren zein neskaren tamainaren erreferentzia zuzena ematen dute; mutil bat bizikletan doala irudikatzeak laguntzen duen modu berean, bizikleta gainean esertzen dela jakiteko.

1. Irudia: Lehoiari eta bizikletan ibiltzeari buruzko testu eta irudiak. Irudiek, testu hutsak baino informazio espazial gehiago ematen dute. Irudia Liu *et al.* (2022)-tik lortu da.



Hala ere, testuzko zein ikusizko informazioa barneratzen duten ikusmen-hizkuntza ereduen (IHE) agerpenak ikuspegi berri bat eskaintzen du. IHEek irudietan zein testuan trebatzen direnez, teorian ezagutza espazial hobea izan beharko lukete. Hau, erlazio espazial esplizituak ematen dituen ikusmen-sarrera aprobetxatzeko duten gaitasunagatik gertatzen da.

Lan honekin, hizkuntza eredu eta modalitate anitzeko ereduen zentzumen espaziala neurtu dugu. Eredu mota desberdinak alderatu ditugu, haien gaitasunak eta ahuleziak aztertuz.

2 Arloko egoera eta ikerketaren helburuak

Esperimentu asko egin dira hizkuntza ereduen zentzumen espaziala aztertzeko. Objektu-eskalei dagokienez, objektu bikoteen altuera eta tamaina (Liu *et al.*, 2022) alderatzen dituzten datu multzo asko sortu dira. Datu multzo horiek, nahiz eta normalean baiezko edo ezezko erantzun sinpleko galderez sortuta egon, erronka oso zailak dira ereduentzat. Zentzumen espaziala ez ezik, arrazonamendu konplexuago baten beharra ere badagoelako gertatzen da hau.

Gaiari buruzko aurreko lanek antzeko konparaketak egin dituzte ereduen artean, baina erabilitako ereduak nahiko zaharkituta geratu dira (Liu *et al.*, 2022; Zhang *et al.*, 2022). Beraz, analisi hauek HEH eta IHE berriagoak erabiliz eguneratzea dugu helburu, erlazio espazialen ezagutza, eredu mota ezberdinen arabera nola aldatzen den modu integralagoan ulertuz.

Lan honen helburu nagusia uneko hizkuntza eredu eta modalitate anitzeko ereduen zentzumen espazialaren arrazoiketarako gaitasunak ebaluatzea da, maskaratutako hizkuntza eredu (MHE) zaharragoekin konparatuz. Horretarako, erlazio espazialen ulermena neurtu beharra dago, hala nola objektuen tamaina erlatiboak, haien posizioak eta elkarrekintzak. Lan honetan, eredu mota ezberdinen konparaketa egiten da, besteak beste, testu soileko ereduak, testu sarrerak soilik erabiltzen dituzten modalitate anitzeko ereduak eta irudikapen bisualak sortzen dituzten

testu-irudi ereduak.

3 Ikerketaren muina

3.1 Datu multzoak eta atazaren formulazioa

Lehenengo datu multzoa, TNWIT, (Liu *et al.*, 2022) lanean aurkeztu zen, ikerketa hau inspiratu duen lana. Datu multzo berari dagokionez, hiru alderdi nagusi hartzen ditu kontuan zentzumen espaziala ebaluatzeko: objektuen altuera, objektuen tamaina eta objektu pareen arteko posizio erlazioa. Jendeak objektuen eskalak multzo lausoetan kategorizatzeko joera duen aurkikuntza kognitiboan inspiratuta (Hersh eta Caramazza, 1976), datu multzoak eguneroko bizitzako 25 objektu arrunt ditu, 5 taldetan sailkatuta. Sailkapen horretatik tamainak konparatzeko datu multzoa eraiki zuten. Era beretsuan sortu zuten objektuen arteko altuera erlatiboa konparatzeko azpimultzoa.

Bi ataza horietarako, 250 objektu pare sortzen dira, adibidez *(bird, ant)*, eta pare bakoitzerako etiketa bat ezartzen da. Etiketa 0 bada, lehenengo objektua bigarrengoa baina baxuagoa/txikiagoa dela adieraziko du. Aldiz, etiketa 1koa bada, lehenengo objektua bigarrengoa baina altuagoa/handiagoa dela esan nahiko du. *(bird, ant)* objektu parearentzat adibidea 1. taulan ikus daiteke. Erantzunen banaketan desorekarik egon ez dadin, datu multzoak *(ant, bird)* bezalako kontrako pareak ere baditu, beraz, 500 instantzia egongo dira guztira altuera eta tamainaren atazentzat.

Posizio erlazioari dagokionez, datu multzoak ekintza bat planteatzen du adibide bakoitzerako. Ekintzarekin batera, pertsona bat eta objektu bat ere aipatuko ditu. Ataza honetan, ekintza hori emanda, pertsona objektuarekiko zein posizio-harremanetan dagoen adieraztea izango da. Ondorengo lau posizio harremanak hartzen dira kontuan: Goian, behean, barruan eta ondoan. *(man, car, inside, driving)* objektu eta egoerarentzat adibidea 1. taulan ikus daiteke. Adibideek ez dute preposiziorik, *sit on a chair* bezala. Ataza honetarako datu multzoak 448 instantzia ditu.

Bigarren datu multzoa, ViComTe, (Zhang *et al.*, 2022) lanean aurkeztua, objektuen arteko 5 ikusizko ezaugarri desberdin aztertzen ditu: kolorea, forma, materiala, tamaina eta zerikusi bisuala. Lan honetarako, datu multzotik tamaina aztertzen duen ataza hartuko dugu eta TNWITko formatu berdinean jarri. Datu multzo honek, TNWITek bezala objektuak tamainaren arabera taldekatzen ditu, eta talde horietatik objektu pareak sortu. *(bird, ant)* objektu parearentzat adibidea 1. taulan ikus daiteke. Denera, ViComTe datu multzoak 6.808 instantzia ditu.

1. Taula: Ataza eta ebaluazio bakoitzerako instantzien adibideak. Altuera eta tamainaren kasuan, erabilitako objektu parea *(txoria, inurria)* da. Posizio erlazioa aztertzen duen atazarako, ondorengo sarrera izango da: *(man, car, inside, driving)*.

	Altuera/Tamaina	Posizio erlazioa
Ebaluazio estandarra	Is a **bird** taller/larger than an **ant**?	A **man** is **driving** the **car**. Is the **man inside** the **car**?
Ebaluazio zorrotza	Is a **bird** taller/larger than an **ant**? + Is an **ant** taller/larger than a **bird**?	A **man** is **driving** a **car**. Where is the **man** with respect to the **car**?

Altueraren atazetarako, bi objektu eta zein den altuagoa zehazten duen etiketa bat emanda, lehenengo objektua bigarrena baino altuagoa den esatean datza zeregina. Adibidez, sarrera Q(*bird*, *ant*, *1*) bada, aurreikusitako irteera *Egiazkoa* izango litzateke. Honek esan nahi du, lehenengo objektua, txoria, bigarrena, inurria baina handiagoa dela betetzen dela. Tamainaren atazarako, formulazio berdina egingo da, baina erlazioa *Taller* izan beharrean *Larger* izango da.

Testutik abiatuta irudiak sortzen dituzten ereduen kasuan, objektuak testu moduan pasako ditugu sarrera bezala, eta sortutako irudian agertzen diren objektuen tamaina eta altuerak konparatuko ditugu instantzia bat zuzena edo okerra den erabakitzeko. Sortutako irudiak eskuz ebaluatuko direnez, eredu mota honetarako TNWIT datu multzoan bakarrik probatzea erabaki dugu, ViComTe datu multzoak instantzia kopuru handiegia duelako eta tamainaren ataza TNWIT datu multzoan agertzen delako jada ere.

Posizio erlazioa aztertzen duen atazarako, galderaren formulazioa aldatzen dugu tamaina eta altueraren atazarekin alderatuta. Ataza honetan, pertsona bat, objektu bat, pertsonaren posizioa objektuarekiko eta ekintza bat

adieraziko dira ataza formulatzeko. Modu honetan, sarrerak itxura hau izan dezake: Q(*man*, *car*, *inside*, *driving*), autoa gidatzen duen gizona barruan dagoela inplikatuz. Kasu honetan, irteera *Egiazkoa* izatea espero da.

Testutik irudiak sortzen dituzten ereduen kasuan, pertsona, objektua eta egoera testu moduan erabiliko dira sarrera bezala, eta ereduak sortutako irudian pertsona eta objektuaren arteko eralzio posizionala egokia den ala ez erabakiko dugu, kasu honetan ere eskuz ebaluatuz.

Ereduak bi modutan ebaluatu ditugu: ebaluazio estandarra eta ebaluazio zorrotza.

Ebaluazio estandarrean, ereduei galdera bakar bat egingo diegu aldi bakoitzean. Erantzun bakoitza banan-banan aztertuko dugu, eta ereduak ematen duen erantzunaren arabera, erantzuna zuzena edo okerra den ikusiko dugu. Erantzun zuzen guztien batura egin eta instantzia kopuru osoarekin zatituko da ereduaren asmatze-tasa kalkulatzeko. 1. taulan ikus daitekeen bezala, altuera eta tamainaren atazetarako, *(bird, ant)* objektu parearentzat ondorengo galdera planteatu dugu: *Is a bird taller/larger than an ant?* Esperotako emaitza baiezkoa izango da bi atazetarako kasu honetan. Posizio erlazioari dagokionez, lehenengo egoera planteatuko dugu, eta ondoren baiezko edo ezezko galdera bat egingo dugu. 1. taulan ikus daiteke *(man, car, inside, driving)* egoera duen adibiderako galdera nolakoa izango zen: *A man drives the car. Is the man inside the car?* Esperotako emaitza hemen ere baiezkoa izango zen kasu honetan.

Ebaluazio zorrotzaren ideia nagusia ebaluatutako ereduak benetan galdera ulertzen dutela ziurtatzea da. Altuera eta tamainaren kasuan, instantzia bat ongi erantzuten bada ere, honen aurkakoa ere zuzen erantzun beharko da erantzuna zuzen bezala kontatzeko. 1. taulan ikus daiteke honen adibide bat. Horrek esan nahi du, eredu batek zuzen erantzuten badu hegazti bat inurri bat baino handiagoa dela, inurri bat ez dela hegazti bat baino handiagoa ere jakin beharko luke. Bi galdera horiek egoki erantzuten baditu bakarrik kontatuko da erantzun bat zuzena bezala. Posizio erlazioa aztertzen den atazan, ebaluazio zorrotzeko galderak ez dira baiezko edo ezezkoak izango, finkatutako 4 posizio-erlazioen artean aukeraketa egin beharko du eredu bakoitzak emaitza egokia aukeratzeko. Honen adibide bat ere ikus daiteke 1. taulan.

ViComTe datu multzoa bi aldiz ebaluatuko dugu. Lehenengo, instantzia guztiak *Larger* erlazioarekin ebaluatuko ditugu, eta ondoren, erlazioa *Smaller* aldatuz ebaluatuko dugu. Modu honetan, ereduek sarrerarekiko duten alborapena (Gallegos *et al.*, 2024) ikusi al izango dugu. Adibidez, *(Elephant, Book)* objektu parea izanda, ondorengo bi galderak egingo ditugu: *Is an elephant larger than a book?* eta *Is an elephant smaller than a book?* Bi zutabe izango ditugu emaitza taulan, bat *Larger* erlaziorako, eta bestea *Smaller* denerako.

Flan-T5 ereduaren tamaina desberdinak ebaluatu ditugu ere. Hauek, TNWIT datu multzoan probatu ditugu ebaluazio estandarra erabiliz, eredua eskalatzeak duen eragina aztertzeko.

3.2 Ereduak

Ereduak hiru kategoria nagusitan banatu ditugu: testu soileko ereduak, modalitate anitzeko ereduak, eta testutik irudiak sortzen dituzten ereduak.

Testu soileko ereduen kasuan, maskaratutako eredu zaharragoak diren BERT-Large (Devlin, 2018) eta RoBERTa-Large (Liu *et al.*, 2019) erabiliko ditugu. Kategoria honetan ere Flan-T5 (Chung *et al.*, 2024), Mixtral-8x7b (Jiang *et al.*, 2024) eta Vicuna-7b (Chiang *et al.*, 2023) ereduak izango ditugu. Modalitate anitzeko ereduei dagokionez, Llava-v1.5-7b (Liu *et al.*, 2024) erabiliko dugu, Vicuna ereduetik abiatuta irudiekin doitutako ikusmen-hizkuntza eredua. Azkenik, testutik abiatuta irudiak sortzen dituzten eredu bezala Stable Diffusion-XL-Turbo (Sauer *et al.*, 2024) erabiliko dugu. Aukeratutako eredu guztiak irekiak dira, eta haien tamainak esperimentuen baliabide-mugak kontuan hartuta hautatu dira.

Testu bakoitzaren sarrerari dagokionez, 1. taulan agertzen diren galderak sortu ditugu. Instrukzio modura, galdera bakoitza baino lehen *Answer the following question with just a 'Yes' or a 'No'* pasa diegu ereduei. Baiezko edo ezezkoa ez den ataza bakarra ebaluazio zorrotzeko posizio erlazioa denez, ondorengo instrukzioa sortu dugu: *Answer the following question using one of the following: above, below, inside, beside*. Stable Diffusionen kasuan, galderarik erantzun behar ez duenez, objektu pareak (*A bird and an ant*) edo egoerak (*A man is driving a car*) pasa dizkiogu sarrera gisa. Maskaratutako hizkuntza ereduei dagokionez, ataza bakoitzeko erlazioa maskaratu dugu eta bilatzen ditugun hitzen artean probableena zein den ikusi dugu, adibidez: *A bird is [MASK] an ant* edo *A man is driving a car. The man is [MASK] the car*.

Flan-T5 ereduaren kasuan, 5 tamaina desberdin probatu ditugu: `small`, `base`, `large`, `XL`, eta `XXL`. Modeloen tamainak, parametro kopuruei dagokionez, 77 milioitik, txikiena izanik, 11.300 milioira iristen dira. Emaitzen

tauletan Flan-T5-XXL ereduaren emaitzak erakutsiko ditugu, baina parametroak eskalatzearen analisi bat egingo da 5 tamaina desberdinen artean ere.

Emaitzetako tauletan 3 eredu familia zehaztuko ditugu ere bai: maskaratutako hizkuntza ereduak (MHE), hizkuntza eredu handiak (HEH) eta ikusmen-hizkuntza ereduak (IHE).

3.3 Emaitzak

Ebaluazio estandarra Emaitzak taula bakar batean erakutsiko ditugu. Alde batetik, TNWIT datu multzoaren atazetan ereduek lortutako emaitzak jasoko ditugu. Ataza hauek Altuera, Tamaina eta Objektuen arteko posizio erlazioa dira. Taula bertan, ViComTe datu-multzorako emaitzak jasoko ditugu, *Handiagoa* eta *Txikiagoa* erlazioak erabiliz bi objektuen arteko tamaina konparaketak egiteko.

2. Taula: Ereduen asmatze-tasak TNWIT eta ViComTe datu multzoetan. †-ak 4 biteko kuantizazioa aplikatu zaiela ereduei esan nahi du. *-ak maskaratutako hizkuntza ereduetarako ausazko oinarri-lerroa %25-koa dela esan nahi du.

Familia	Eredua	TNWIT Estandarra			ViComTe Estandarra	
		Altuera	PosErl	Tamaina	Handiagoa	Txikiagoa
	Ausaz	50,00	50,00	50,00	50,00	50,00
MHE	$\text{BERT}_{\text{Large}}$	56,20	$23,21^*$	52,40	46,53	46,53
	$\text{RoBERTa}_{\text{Large}}$	48,20	$28,57^*$	74,20	62,78	62,78
HEH	$\text{Flan-T5}_{\text{XXL}}$†	93,40	81,47	**98,40**	**94,00**	84,93
	$\text{Mixtral}_{\text{8x7b}}$†	**93,60**	76,11	96,00	88,17	**87,60**
	$\text{Vicuna}_{\text{7b}}$†	52,80	66,90	47,00	70,25	47,38
IHE	Llava_{7b}	64,20	66,52	81,00	76,40	53,03
	SD-Turbo	78,40	**90,63**	80,00	-	-

2. taulan ikus daiteke altuera eta tamainaren atazetan emaitza hoberenak Flan-T5 eta Mixtral ereduek lortzen dituztela, %90 baino gehiagoko asmatze-tasak lortuz. Beraz, esan dezakegu eredu hauek atazak ulertu eta ondo erantzuteko gai izan direla. Aldiz, Vicuna eta MHEak izan dira emaitza okerrenak lortu dituzten ereduak. Hauetako batzuk ez dira gai izan ausazko oinarri-lerroa hobetzeko ere kasu batzuetan. 2. taulak argi erakusten du posizio erlazioa aztertzen den ataza izan dela zailena ereduentzat orokorrean. Hala ere, ataza honetan nabarmentzen den eredua Stable Diffusion-XL-Turbo da, %90,63-ko asmatze-tasarekin, beste ereduak nabarmen gaindituz. Flan-T5 eta Mixtralek ere emaitza onak lortzen dituzte ataza honetan, %81,47 eta %76,11ko asmatze-tasak hurrenez hurren, baina ez daude testutik irudirako ereduak erakusten duen gaitasunetik gertu.

ViComTe datu multzoko emaitzei dagokionez, eredu handienak dira errendimendu onenak lortzen dituztenak, 2. taulan ikus daitekeenez. Emaitza hauek Flan-T5 eta Mixtral bezalako ereduek TNWIT datu multzoan lortutako errendimendu ona azpimarratzen dute. Erlazio aldaketaren eragina ikus daiteke ere bai. Erlazioa *Handiagoa* denean, ereduek emaitza hobeagoak lortzen dituzte *Txikiagoa* erlazioa erabiltzean baino. Erabilitako erlazioa *Handiagoa* denean, Flan-T5 eredua da asmatze-tasa altuena lortzen duena, %94. Aldiz, erlazioa *Txikiagoa* denean, Mixtral da eredu hoberena, %87.60ko asmatze-tasarekin. Llava eta Vicuna arteko konparaketa zuzen bat egiten badugu, eredu multimodala da orokorrean emaitza hobeagoak lortzen dituen eredua.

Lehen aipatu bezala, ereduen alborapena argi ikus daiteke gure emaitzetan. Erlazioa *Handiagoa* izatetik *Txikiagoa* izatera pasatzen denean, ereduen errendimendua jaitsi egiten da orokorrean. Hori oso nabarmena da Vicuna eta Llava ereduentzat, haien asmatze-tasak %20 baino gehiago jaisten baitira erlazio aldaketa egitean. Mixtral eredua da alborapen nabarmen hau erakusten ez duen bakarra, bi erlazioekin emaitza oso antzekoak lortuz.

Ebaluazio zorrotza Bigarren ebaluazioko emaitzak 3. taulan erakusten dira. Berriro ere, altuera eta tamaina neurtzen den atazetan errendimendu onena duten ereduak Flan T5 eta Mixtral dira. Altueraren atazarako, eredu hauen emaitzak %87 ingurukoak dira eta tamainaren atazan %90a baino gehiagokoak. Gainera, bi hauek izan dira ebaluazio zorrotzak gutxien kaltetu dituen ereduak, %6,6koa izanik galera handiena kasu honetan. Ebaluazio estandarrean bezala, BERT, RoBERTa, eta Vicuna izan dira atazako emaitza okerrenak lortu dituzten ereduak. Emaitzek erakusten dute Vicuna izan dela ebaluazio zorrotzak kalte gehien egin dion eredua, bere asmatze-tasak %43,2 eta %36,6 beheratuz. Posizio erlazioa aztertzen duen atazari dagokionez, eredu guztiek izan dute galera

handia ebaluazio zorrotza aplikatzean. Azken ataza honetan, Flan-T5 izan da emaitza onena lortu duen eredua %67.41ko asmatze-tasa lortuz. Galera hau esperotakoa da, hirugarren ataza honen konplexutasuna, aurreko biena baina altuagoa delako batez ere.

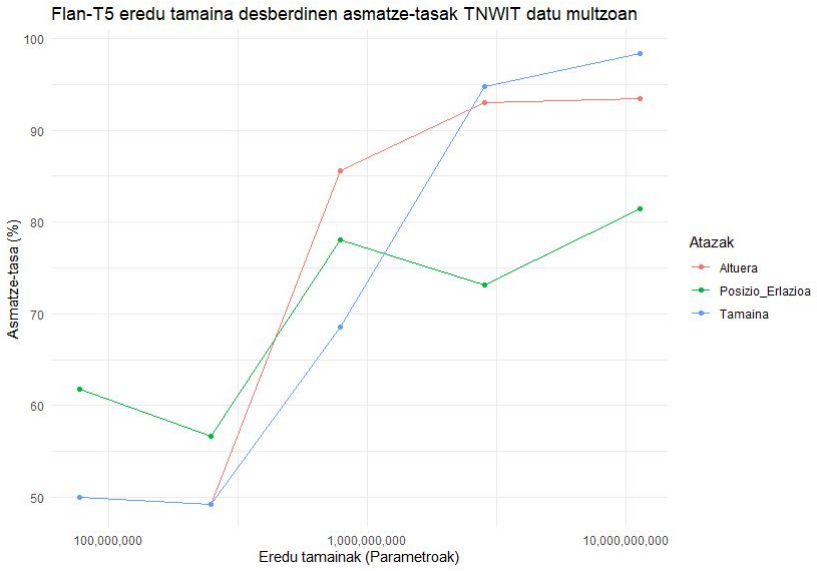
3. Taula: Ebaluazio zorrotzeko emaitzak TNWIT eta ViComTe datu multzoetan. Zenbaki gorriek, ebaluazio estandarretik ebaluazio zorrotzera pasatzean izandako beherakada erakusten dute. †-ak 4 biteko kuantizazioa aplikatu zaiela ereduei esan nahi du.

Familia	Eredua	TNWIT Zorrotza			ViComTe Zorrotza	
		Altuera	PosErl	Tamaina	Handiagoa	Txikiagoa
	Ausaz	25,00 (-25,0)	25,00 (-25,0)	25,00 (-25,0)	25,00 (-25,0)	25,00 (-25,0)
MHE	$BERT_{Large}$	16,00 (-40,2)	23,21 (-0,0)	16,80 (-35,6)	16,06 (-33,2)	16,06 (-33,2)
	$RoBERTa_{Large}$	0,40 (-47,8)	28,57 (-0,0)	52,40 (-21,8)	24,94 (-37,8)	24,94 (-37,8)
HEH	Flan-$T5_{XXL}$†	86,80 (-6,6)	**67,41** (-14,1)	**97,20** (-1,2)	**91,63** (-2,4)	71,42 (-13,5)
	$Mixtral_{8x7b}$†	**87,60** (-6,0)	42,85 (-33,3)	92,40 (-3,6)	84,28 (-3,9)	**82,70** (-4,9)
	$Vicuna_{7b}$†	9,60 (-43,2)	12,40 (-54,5)	10,40 (-36,6)	53,06 (-17,2)	15,83 (-31,6)
IHE	$Llava_{7b}$	28,80 (-35,4)	31,69 (-34,8)	62,00 (-19,0)	66,60 (-9,8)	31,64 (-21,4)

ViComTe datu multzoaren ebaluazio zorrotzean zentratuz, 3. taulak erakusten du Flan T5 dela errendimendu onena duen eredua erlazioa *Handiagoa* denean, %91,63ko asmatze tasarekin. Erlazioa *Txikiagoa* denean, Mixtral da eredurik onena, %82,7ko asmatze tasa lortuz. Berriz ere, MHEak eta Vicuna dira emaitza okerrenak lortzen dituzten ereduak, galera handia izanik ebaluazio zorrotzaren eraginez. Halaber, erlazioa *Handiagoa* izatetik *Txikiagoa* izatera pasatzen denean alborapen maila baxuena erakusten duen eredua Mixtral izan da. Eredu honek lortutako asmatze-tasak oso berdintsuak izan dira, %84,28 *Handiagoa* denean eta %82,7 *Txikiagoa* denean. Berriz ere Vicuna eta Llava alderatzen baditugu, ikus dezakegu Llavak oraindik ere Vicuna gainditzen duela harreman gisa handiagoa zein txikiagoa erabiliz.

Beraz, ViComTe datu multzoa ebaluatzeko moduak ereduen alborapen argia erakutsi du sarrerekiko. 2. eta 3. tauletan ikusten den bezala, objektuen arteko erlazioa *Handiagoa* denean errendimendua hobeagoa da erlazioa *Txikiagoa* denean baino, eta horrek iradokitzen du ereduek hitz zehatz batzuekiko berezko lehentasuna dutela. Hala ere, Mixtral ereduaren alborapena oso baxua da beste eredu batzuekin alderatuta, hau bere arkitekturagatik edo entrenatu den moduagatik izan liteke.

2. Irudia: Flan-T5 eredu desberdinen asmatze tasak TNWIT datu multzoko atazetan, parametro kopurua handitzeak duen eragina erakutsiz.



Tamainaren eragina Eredu handienak erredimendu orokor onena erakutsi badute ere, Flan T5en eredu-tamainen arteko konparaketa egin da. 2. irudian ikus dezakegunez, ereduaren bertsio txikienek %50eko asmatze-tasa lortzen dute; hau da, beti berdina erantzunez eta adibide erdietan asmatuz. Ereduak handitzen doazen heinean, emaitzak hobetzen dira, galderak modu zehatzagoan erantzunez. Horrek modeloen tamainen garrantzia erakusten du. Parametro kopurua handiagoa denean, eredua objektuak eta haien arteko erlazioak ulertzen hasten da, eta egoki erantzuteko gai bihurtzen da.

4 Ondorioak

Lan honetan, egungo hizkuntza eta eredu multimodalak ebaluatzen ditugu hauen zentzumen espazialari dagokionez. Arrazonamendu espazialerako datu multzoak identifikatu eta biltzen ditugu, datu multzoetan eredu mota ezberdinak ebaluatuz eta ataza ezberdinetarako emaitzak lortuz.

Emaitzek erakusten dute egungo ereduek aurrerapauso handia izan dutela hauen zentzumen espazialaren ulermenean. BERT edo RoBERTa bezalako eredu zaharragoek ez dute errendimendu ona lortzen ataza bakar batean ere, aldiz, gaur egungo eredu handiagoek, Mixtral edo Stable Diffusion adibidez, objektuen arteko konparazioak eta egoera jakin batean dauden objektuen posizio-erlazioak ulertzeko gai dira.

Parametro kopuru handia duten ereduek emaitza onak izan badituzte ere, argi esan dezakegu testutik irudirako ereduak izan direla objektuen arteko posizio-erlazioak hobekien ulertu dutenak. Irudiekin trebatuta egoteak lagundu egiten dio ereduari beste eredu batek ondorioztatuko ez lituzkeen egoerak eta ekintzak ulertzen.

Ereduen alborapenari dagokionez, emaitzek erakusten dute alborapen nabaria dagoela galderaren formulazioa aldatzen denean. Erlazioa *Handiagoa* denean, orokorrean emaitza hobeagoak lortzen dira *Txikiagoa* erlazioa erabiltzean baino. Honako hau, ereduak entrenatu diren moduagatik izan daiteke, bertan lehenengo erlazioa adibide gehiagoetan ikusi dutelako bigarrenarekin alderatuz. Mixtral izan da gure emaitzen arabera erlazioen artean alde txikia erakutsi duen eredu bakarra.

Eredu multimodalen modalitate anitzeko entrenamenduak, ereduen ulermen espazialean laguntzen duela ondorioztatzen dugu emaitzetatik ere. Nahiz eta sarrera-modalitatea testuala bakarrik izan, Vicuna eredua oinarri hartuta entrenatzen den Llava bezalako eredu batek Vicuna-7b bezalako eredu bat hobetu dezakeela ikusi dugu. Eredu multimodalek testuarekin batera erabiltzen dituzten irudiek lagundu egiten diete zentzumen espaziala eskatzen duten atazetan.

Lehen aipatu bezala, testu-irudi ereduak dira posizio erlazioa aztertzen den atazan errendimendu hoberena lortzen duen eredu mota. Irudiek testu soilak baino informazio gehiago dute, eta horri esker, eredu mota horiek egoerak hobeto uler ditzakete. Irudiak sortzean arazo nabarmen bat izan dugu: bi objektuetako bat sortzea falta izan zaio kasu batzuetan. Dena den, ereduak objektuak behar bezala sortzen zituenean, errepresentazio zuzenak pantailaratzeko gai izan zen, gehienetan altuera eta tamaina egokiak lortuz, eta posizio erlazioak egoki adieraziz.

Dena kontuan hartuz, gaur egungo ereduek nabarmen hobeto betetzen dituzte zentzumen espaziala eskatzen duten atazak, eredu zaharragoekin alderatuz gero. Hobekuntza horren arrazoi nagusietako bat egungo ereduetan izan den parametro-eskalatzea izan da. Parametro gehiago izateari esker, ereduek barne-irudikapen aberatsagoak eraiki ditzakete, giza kognizio bisualaren eta arrazoiketaren gaitasun handiagoa lortuz.

5 Etorkizunerako planteatzen den norabidea

Ikerketa lan honek egungo ereduen zentzumen espaziala ebaluatzeko oinarrizko pausoak ematen ditu, eta hurrengo ikerketa-lerroetan jarrai daiteke:

- **Datu Multzoak**: Datu multzo gehiago erabil litezke zentzumen espaziala eta mundu fisikoaren ulermenaren datu sorta zabalagoa biltzeko. Honek, ingurune, testuinguru eta objektu-interakzio askotarikoak dituzten datu-multzoak izan litzake, egoera konplexuagoak aztertuz.
- **Ereduak**: Eredu gehiago eta handiagoak ebaluatu litezke haien eskalagarritasuna eta orokortze-gaitasunak aztertzeko. Horren barruan sartzen dira azken belaunaldiko ereduak probatzea, parametro-tamaina handiagokoak eta ulermen multimodalerako diseinatutako arkitekturak.
- **Atazak**: Zeregin gehiago gehitu litezke ereduen zentzumen espaziala eta mundu fisikoari buruz duten ulermena modu integralean ebaluatzeko. Horrek zeregin berriak diseinatzea ekar lezake, kausalitateari, interakzio fisikoei eta denbora-sekuentziei buruz arrazoitzea eskatzen dutenak.

- **Ebaluazioa**: Irudien ebaluazio automatikoa ezar liteke testutik irudirako ereduen ebaluazio errazagoa eta objektiboagoa egiteko. Zentzuzkotasun bisualeko atazetan modeloen errendimendua automatikoki ebaluatzeko metrikak eta tresnak garatzeak eraginkortasuna eta fidagarritasuna hobetuko lituzke.

Erreferentziak

Chiang, Wei-Lin, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, eta others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)* 2.6.

Chung, Hyung Won, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, eta others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research* 25.1–53.

Cui, Wanqing, Yanyan Lan, Liang Pang, Jiafeng Guo, eta Xueqi Cheng. 2020. Beyond language: Learning commonsense from images for reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, ed. by Trevor Cohn, Yulan He, eta Yang Liu, 4379–4389, Online. Association for Computational Linguistics.

Devlin, Jacob. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* .

Gallegos, Isabel O, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, eta Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics* 1–79.

Gordon, Jonathan, eta Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, 25–30.

Hersh, Harry M, eta Alfonso Caramazza. 1976. A fuzzy set approach to modifiers and vagueness in natural language. *Journal of Experimental Psychology: General* 105.254.

Jiang, Albert Q, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, eta others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088* .

Liu, Haotian, Chunyuan Li, Qingyang Wu, eta Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36.

Liu, Xiao, Da Yin, Yansong Feng, eta Dongyan Zhao. 2022. Things not written in text: Exploring spatial commonsense from visual signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2365–2376.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, eta Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Pellegrino, James W, David L Alderton, eta Valerie J Shute. 1984. Understanding spatial ability. *Educational psychologist* 19.239–253.

Sauer, Axel, Dominik Lorenz, Andreas Blattmann, eta Robin Rombach. 2024. Adversarial diffusion distillation. In *European Conference on Computer Vision*, 87–103. Springer.

Zhang, Chenyu, Benjamin Van Durme, Zhuowan Li, eta Elias Stengel-Eskin. 2022. Visual commonsense in pretrained unimodal and multimodal models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 5321–5335.

6 Eskerrak eta oharrak

Artikulu hau gradu amaierako lan baten laburpen bat izan da. Proiektu honek honako finantziazioa izan du:

- LUMINOUS: Language Augmentation for Humanverse (Europar Batasuna: Horizon Europe HORIZON-CL4-2023HUMAN-01-21-101135724) 21-101135724)