EKAIA (2025), artikulua prentsan/article in press
https://doi.org/10.1387/ekaia.27338
Behin-behineko bertsioa (euskara-orrazketaren faltan).

ekaia
ZIENTZIA eta TEKNOLOGIA ALDIZKARIA

ISSN: 0214-9001
e-ISSN: 2444-3255
©UPV/EHU Press

Paziente birtuala simulatzeko txatbot medikoa

(Virtual Patient Simulation for a Medical Chatbot)

Mikel Idoyaga Bazan[*]1, *Janire Arana Gonzalez*[1], *Jose Vicente Lafuente Sanchez*[2], *Elisa Espina Valiño*[2], *Koldo Gojenola Galletebeitia*[1], *Aitziber Atutxa Salazar*[1]
[1]HiTZ: Hizkuntza Teknologiako Euskal Zentroa, Ixa taldea
Bilboko Ingeniaritza Eskola, Rafael Moreno Pitxitxi pasealekua, 3 | 48013 Bilbao
Euskal Herriko Unibertsitatea UPV/EHU
[2]Medikuntza eta Erizaintza Fakultatea UPV/EHU, Bizkaia

LABURPENA: Lan honetan, mediku eta pazientearen arteko elkarrizketan paziente birtualaren rola hartuko duen gaztelaniazko txatbot baten prototipoaren sorrera aurkeztuko da, lortutako emaitzekin batera. Txatbotaren erantzunak benetako pazientearenak bezalakoak izateko helburuarekin, lan honetan baliabide ezberdinak garatu dira. Alde batetik, medikuaren galderen eta pazienteen txosten klinikoetatik hartutako erantzunen corpus bat baliatu da, eta pazienteak erantzungo lituzkeen esaldien corpus berria sortu da. Baliabide horiek erabilita, paziente birtualaren papera egin dezakeen txatbotaren lengoaia naturalaren sorkuntzarako sistema bat entrenatu da adimen artifizialeko sare neuronalen bidez. Sistema ebaluatua izan da metrika desberdinak erabiliz emaitza positiboekin, horrela medikuntzako elkarrizketa sistema automatikoak eraikitzeko bidea irekiz.
HITZ GAKOAK: adimen artifiziala, lengoaia naturalaren sorkuntza, elkarrizketa-sistema, testu klinikoen prozesamendua.

ABSTRACT: *In this work, we will present the creation of the prototype of a Spanish chatbot that will take the role of a virtual patient in the conversation between doctor and patient, together with the results obtained. With this aim, we have taken an already developed corpus of doctor/patient questions and answers taken from clinical records as a starting point, and a new corpus of patient-like sentences has been generated. Using these resources, a chatbot that can play the role of a virtual patient has been trained using artificial intelligence and natural language generation techniques, implemented with neural networks. The system has been evaluated using different metrics, paving the way for the construction of automatic medical dialogue systems.*
KEYWORDS: Artificial Intelligence, Natural Language Generation, Dialogue System, Clinical Text Processing.

[*]**Harremanetan jartzeko/Corresponding author:** Mikel Idoyaga Bazan, HiTZ Hizkuntza Teknologiako Euskal Zentroa-Ixa, Euskal Herriko Unibertsitatea (UPV/EHU, Sarriena auzoa, z/g (48091-Leioa, Euskal Herria).
https://orcid.org/0009-0007-4955-0758, mikel.idoyaga@ehu.eus

Nola aipatu/How to cite: (Adibidea:) Idoyaga, Mikel; Arana, Janire; Lafuente, Jose Vicente; Espina, Elisa; Gojenola, Koldo; Atutxa, Aitziber (2025). «Paziente birtuala simulatzeko txatbot medikoa», Ekaia, DOI: https://doi.org/10.1387/ekaia.27338

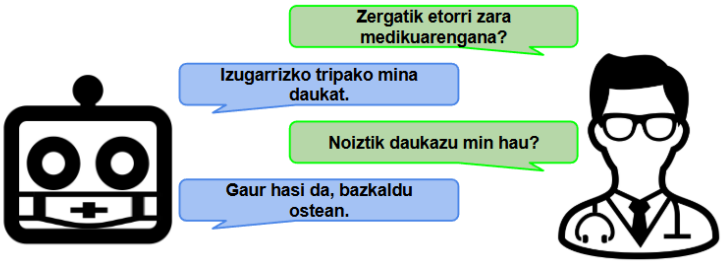
Jasoa: martxoak 5, 2025; Onartua: uztailak 10, 2025
ISSN 0214-9001-eISSN 2444-3225 / ©2025 UPV/EHU

Obra *Creative Commons Atribución 4.0 Internacional*-en lizentziapean dago

1. Sarrera

Urtero, 6. ikasturteko medikuntza eta erizaintza ikasleek Ebaluazio Kliniko Objektibo eta Egituratua (EKOE), gaitasun klinikoen eta komunikazio gaitasunen azterketa, egin behar dute, non ikasle bakoitzaren jakintza teorikoa, arrazoiketa klinikoa, eta haien komunikatzeko gaitasuna eta trebetasuna probatzen diren. Azterketa honetan, ikasle bakoitza zirkuitu batetik pasatzen da eta, bertan, agertoki kliniko bat birsortu edo simulatzen da. Ebaluazioa ahalik eta objektiboen egiteko prestatuta dago; estazio bakoitzerako, ikasleek bete beharreko eginkizun zerrenda bat prestatzen da, eta atazaren osatze-tasaren arabera puntuazio bat jasotzen dute.

Espainiako Medikuntza Fakultateetako Dekanoen Konferentzia Nazionalak (*CNDFME*) [1] 20 estazio proposatzen ditu, bakoitzak 10 minutu inguruko iraupena izanik. Estazio horien artean bi mota nagusi bereiz daitezke: paziente simulatukoak eta pazienterik gabekoak. Lehenengo motako estazioetan, aktore edo boluntario batek pazientearen rola hartu eta kasu kliniko bat simulatuko du. Ikaslea aktorearekin komunikatu beharko da kontsulta aurrera eramateko, pazientearekin duen interakzioa ebaluatzeko. Urtero aktoreak kontratatzea eta prestatzea oso prozesu neketsu eta garestia da [2]. Horren ondorioz, paziente birtualak (PB) eraikitzeko ideia proposatu da, pazienteen txosten klinikoak simulatzeko gai baitira, eta lehenengo prototipoak garatu dira [3] (ikusi 1. irudia). Sistema horiei esker, fakultateek dirua eta denbora aurreztuko dute aktoreak kontratatzen, azterketa kopurua eta aniztasuna handitu ahal izango dituzte eta ikasleek edozein momentutan praktikatzeko aukera izango dute.



1. irudia: PB eta mediku baten arteko elkarrizketaren irudi kontzeptuala.

Paziente birtualak txatbotetan oinarritzen dira. Sistema horiek adimen artifizialean oinarritutako sistema automatizatuak dira, eta erabiltzaileekin hizkuntza naturaleko elkarrizketen bidez informazioa trukatzeko diseinatuta daude [4]. Txatbot baten barnean elkarrizketa-sistema aurki daiteke, zeinen helburu nagusia gizakien eta ordenagailuen arteko komunikazio eraginkorra ahalbidetzea den, erabiltzailearen ekarpenak testu edo ahots moduan ulertuz eta erabiltzailearen eskaerei behar bezala erantzunez.

Paziente birtualak sortzeko existitzen diren txatbot metodologien artean bi mota nagusi bereizi daitezke: txatbot sortzaileak (ChatGPT[1] modukoak) eta modularrak, azken hauek egokienak izanik proiektu hau aurrera eramateko, medikuntzaren domeinuan sistemaren kontrol estuagoa ahalbidetzen baitute.

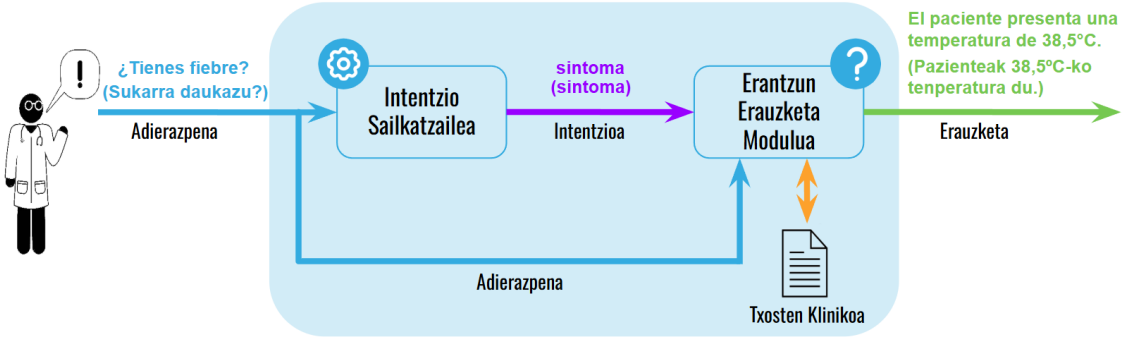
Txatbot modularrek atal desberdinak dituzte eta bakoitza elkarrizketa normal baten zati ezberdinez arduratzen da; honela, hurbilpen sortzaile batek baino aukera gehiago eskaintzen ditu erantzunak kontrolatzeko eta sortzaileen haluzinazioak ekiditeko. Beraz, proiektu honetarako aurretiko prototipo batean oinarritu gara [3], non **Lengoaia Naturalaren Ulermen** (LNU) modulua inplementatu zen (ikusi 2. irudia), honako elementuez osatua:

- ***Adierazpen Ulermena (AU)***: erabiltzailearen sarrera prozesatzen du. Azpi-modulu honek hainbat ataza burutzen ditu, baina proiekturako garrantzitsuena **intentzio sailkatzailea** (IS)

[1] https://chatgpt.com/

da, non erabiltzaileak egindako adierazpenaren intentzioa igartzen saiatzen den. Horrez gain, beste ataza batzuk ere egin ditzake; hala nola, hitz gakoak harrapatzea eta elkarrizketan zerbait jada aipatu den antzematea.

- ***Erantzun Erauzketa (EE)***: datu-base edo kanpo-dokumentuetan (kasu honetan, paziente birtual bat denez, pazienteak deskribatzen dituzten txosten klinikoetan) erabiltzailearen sarrerari erantzun egokia topatzeaz arduratzen da.



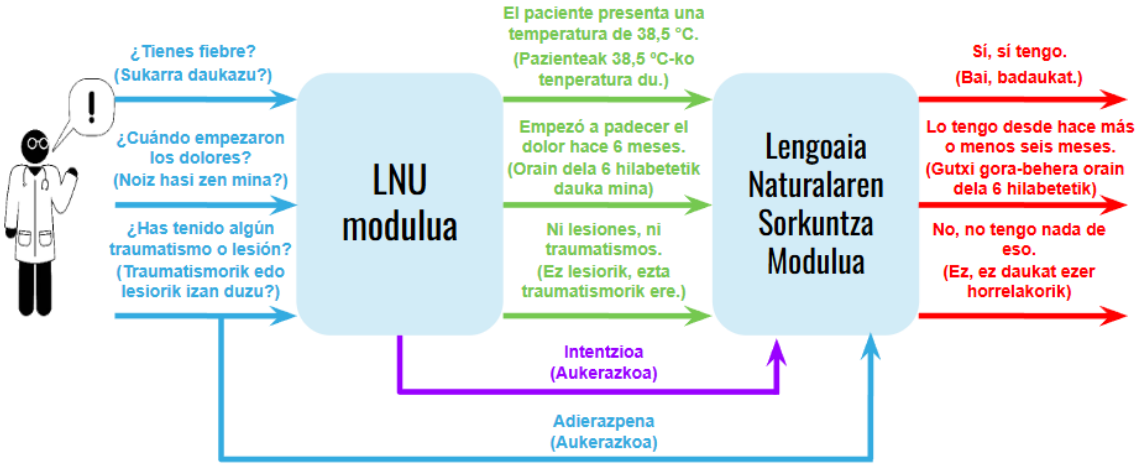
2. irudia: LNU moduluaren egitura [3].

2. irudian ikusten denez, sistema horretan, galdera baten erantzuna pazientea deskribatzen duen txosten klinikotik erauzten da eta, ondorioz, sistemak bueltatzen duen erantzuna ez da naturala, txostenetik hartutako testu-zatia baita. Hau da, medikuaren galderei erantzutean, txosten klinikoan agertzen diren hitz berberak erabiliko ditu. Beraz, txatbotak ez du benetako paziente batek erantzungo lukeen moduan erantzungo, lehen pertsona erabiliz edo hitz teknikoak saihestuz. Paziente birtuala ahalik eta errealistena izatea nahi denez, hori konpontzeko modulu berri bat gehitu nahi dugu, **Lengoaia Naturalaren Sorkuntza** (LNS, ikusi 3. irudia) modulua, hain zuzen ere. Adibidez, medikuak "Sukarra daukazu?" galdetzen badu, eta kasu klinikoan "Pazienteak 38,5 ºC-ko tenperatura du" agertzen bada, LNS-ak lehenengo pertsonara bihurtuko du, esanez "Bai, badaukat". Lan honetan, txosten kliniko batetik atera daitezkeen esaldiak paziente batek esango lituzkeen esaldietara bihurtzen dituen modulua garatuko da.

LNS modulu egokiena aurkitzeko, artikulu honetan garatu diren eredu eta teknika desberdinak azalduko dira eta horiek sortzeko beharrezkoak izan diren baliabideak deskribatuko dira. Hala ere, LNS moduluan zentratu arren, IS eta EE moduluak ere erabiliko dira LNS ereduen hobekuntzarako. Izan ere, txatbotaren parte dira eta modulu guztiek ondo funtzionatu behar dute elkar lanean. Helburu hori lortzeko, honako baliabide hauek garatu dira:

- **Esaldien egokitzapenerako corpusa**. Corpus horretan txatbot erauzle baten [3] erantzunak pazientearen elkarrizketako erantzunekin parekatu dira, betiere txosten kliniko berean oinarrituta, (*txosteneko esaldia, pazientearen esaldia*) bikoteen corpusa sortuz. Horrez gain, hainbat corpus aldaera sortu dira, hizkuntzaren egokitasuna, koherentzia eta fidagarritasuna hobetzeko. Helburua ikasketa automatikorako formatu hoberena identifikatzea da, LNS moduluaren garapenerako baliagarri izango den datu-multzo bat osatuz.

- **Lengoaia Naturalaren Sorkuntza modulua inplementatu eta ebaluatu da** adimen artifizialeko teknikak erabiliz. Horretarako, hainbat arkitektura aztertu dira, baita doiketa-teknikak eta datu gehikuntza estrategiak ere. Moduluaren helburua test klinikoen sorkuntza errealistagoa lortzea da, pazienteen elkarrizketak simulatzeko eta osasun-profesionalen galderei erantzuteko ahalmena hobetuz.



3. irudia: LNS moduluaren funtzionamendua. Lehendik egindako LNU moduluak txosteneko hitzen bidezko erantzunak lortuko ditu (irudiaren erdian), eta LNS moduluak pazientearen hitz egiteko eran bihurtuko ditu (irudiaren eskuinean).

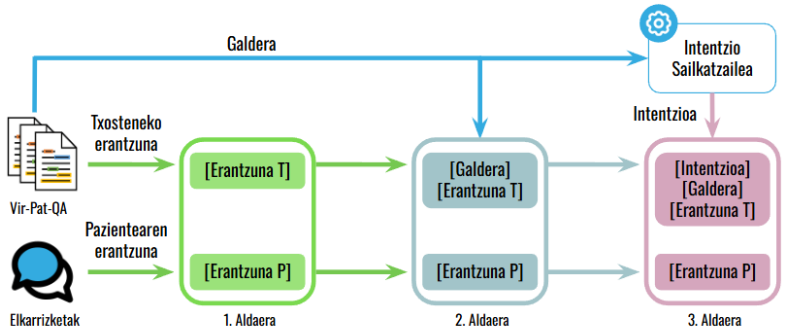
2. Baliabideak eta metodoak

Lortu nahi den LNS modulua garatzeko, txosten klinikoetatik hartutako edozein esaldi edo *EE* modulutik lortutako erantzunak paziente batek esango lituzkeen esaldietara itzultzeko gai den adimen artifizialeko eredu bat sortu behar da. Horretarako, corpus bat eraikiko da, non bi esaldiak parekatuta dauden (ikusi 2.1 azpiatala). Corpus hori oinarri hartuta, lengoaia naturalaren sorkuntzarako modulu bat garatu da, txosten klinikotik hartutako esaldiak paziente batek esango lituzkeen hizkerara itzultzeko (ikusi 2.2. azpiatala).

2.1. VIR-PAT-LNS corpusa

LNS ereduak sortu ahal izateko ikasketa gainbegiratuko sistemak erabili dira; hau da, ereduek etiketatutako datuetatik ikasiko dute, hortaz, datu multzo etiketatuak behar dira hori burutzeko. Datuen babesari buruzko legea dela eta, oso zaila da osasun esparruko datuak lortzea eta, horregatik, eskuragarri dauden corpusetatik, VIR-PAT-QA [3] corpusa aukeratu da. Corpus horrek, mediku-paziente elkarrizketetatik lortutako txosten kliniko desberdinak biltzen ditu, baita medikuek egindako galderak eta horien erantzunak dauzkaten txosteneko zatiak ere.

Gure corpusak hiru aldaera ezberdin ditu. 4. irudian ikus daiteke corpusaren aldaera bakoitza lortzeko prozesua era laburtuan.



4. irudia: VIR-PAT-LNS corpusaren aldaerak sortzeko prozesua. (T = Txostenetik erauzitako erantzuna, P = Pazienteak elkarrizketan emandako erantzuna)

Lehenengoaren abiapuntu gisa, VIR-PAT-QA corpuseko erantzunak, hori sortzeko erabilitako elkarrizketetan pazienteek emandakoekin parekatu dira. Adibidez, 3. irudian ikusten den bezala, txosteneko esaldia "Pazienteak 38,5 ºC-ko tenperatura du." bada, horri dagokion pazientearen esaldia hurrengo hau izan daiteke: "Bai, badaukat." Jatorrizko erantzunak medikuntza ikasle eta boluntarioekin grabatutako kontsulta simulatuen audioen transkripzioak eginez sortu ziren EKOE-ko azterketetan [5]. Hauek ingelesez egin zirenez, itzulpen automatikoa eta itzulpenaren akatsen zuzenketaren bidez lortu ziren guk erabilitako erantzunak [3].

VIR-PAT-QA corpusean hiru galdera mota bereizten dira: a) "erantzuna duten galderak", b) "txosten klinikoan erantzunik ez duten galderak" eta c) "erantzunik behar ez duten gelderak", azken bi mota horiek bereziak izanik. *b* motako galderetan txosten klinikoan ez dago erantzunik elkarrizketako esaldiarekin parekatzeko eta, horregatik, mota hori baztertua izan da. *c* motako galderetan, aldiz, erantzunak ez dira txosten klinikotik atera behar (erantzun generikoak espero dira) baina, hala ere, hauek mantentzea erabaki da, azken batean, galdera hauek medikuaren agurrak, baieztapenak, ezeztapenak... biltzen dituelako eta, hortaz, ez dira galdera/interakzio garrantzitsu bezala hartu behar. Hauetan, nahiz eta erantzuna ez den txosten klinikoan bilatu behar, erantzun bat eskatzen dute pazientearen aldetik (ikusi 1. taula). Hori dena kontuan izanda, 3934 [*Erantzuna, Pazientearen Erantzuna*] pare geratzen zaizkigu gure ereduak eraikitzeko.

Galdera mota	**entrenamendua**	**garapena**	**test**
a) *Erantzuna dutenak*	*2753*	*295*	*580*
~~b) *Erantzunik ez dutenak*~~	~~*1820*~~	~~*201*~~	~~*335*~~
c) *Erantzunik behar ez dutenak*	228	27	51
Guztira (b mota kenduta)	**2981**	**322**	**631**

1. taula: Mota bakoitzeko galdera kopurua entrenamendu, garapen eta test multzoetan, erantzunik gabeko galdera-motak kenduta.

Horren ostean, corpusaren hurrengo aldaera eraiki da. Izan ere, VIR-PAT-QA corpusaren txosteneko erantzunek, zuzenak izan arren, askotan ez dute testuinguru nahikorik ematen ereduak zeri buruz ari diren ulertu ahal izateko. Adibidez, medikuaren galdera "Noiz izan zenituen lehen sintomak?" bada eta txostenean hartu den zatia bakarrik "orain dela 3 hilabete" bada, LNS ereduak ez du jakingo sintomei buruz ari dela. Horregatik, ereduei testuinguru gehiago emateko aldaera bat egin da, non txosteneko eta pazienteak emango lukeen erantzunak egoteaz gain, egindako galdera ere egongo den. Esaterako, ["Badaukazu sukarrik?", "Pazienteak 38,5 ºC-ko tenperatura du.", "Bai, badaukat."], hau da, [*Galdera, Erantzuna, Pazientearen erantzuna*] motako hirukoteak agertuko dira.

Gainera, ikusita aurrekarietan medikuaren intentzioak lagungarriak izan zirela ereduen emaitzak hobetzeko [3], hirugarren aldaera bat sortu da, non txosteneko zatiei intentzioa ere lotzen zaien, aurreko corpusaren patroi berdina jarraituz. Hurrengo laukote hau izan genezake: ["Sintoma", "Badaukazu sukarrik?", "Pazienteak 38,5 ºC-ko tenperatura du.", "Bai, badaukat."], hau da, [*Intentzioa, Galdera, Erantzuna, Pazientearen erantzuna*] laukoteak (ikusi 4 irudiko. 3. aldaera). Aldaera hau garatzeko, galdera bakoitzaren intentzioa lortzeko intentzio-sailkatzaile automatikoa erabili da [3]. Eredu horrek, modu hierarkikoan antolatutako 145 intentzio desberdin (guztiak 11 kategoria nagusitik hedatuta) bereizi ditzake. Kategoria nagusi horiek, medikuen galderetan aurki daitezkeen intentzioen mota nagusiak azaltzeko balio dute, 2. taulan ikus daitekeen bezala. Intentzioen sailkapen osoaren deskribapena ulertzeko ikusi [3].

Kategoria nagusia	Deskribapena	Adibidea	Entrenamendua	Garapena	Test
baieztatu	Erantzun ez beharreko adierazpenak.	Bai	109	12	31
agur_esan	Elkarrizketa amaitzeko adierazpenak.	Agur	4	0	1
egoera	Pazientearen egoeraren galdera orokorrak.	Zelan zaude?	3	0	0
kontsulta arrazoia	Pazientearen etorreraren arrazoia	Zerk ekartzen zaitu hona?	95	10	21
besteak	Beste kategorietara moldatzen ez diren adierazpenak	Emadazu zure osasun txartela	3	0	0
pertsonala	Pazientearen bizitzaren inguruko galderak	Zein da zure izena?	400	47	100
psikiatria	Pazientearen sentimenduei buruzko adierazpenak	Pozik bizi zara?	48	4	10
kaixo_esan	Elkarrizketa hasteko moduko adierazpenak.	Aupa	15	4	3
sintoma	Pazientearen sintomei buruzko galderak	Alergiarik?	2006	214	400
tratamendua	Pazienteak izandako tratamenduei buruzko galderak	Beste medikamenturik?	228	31	58
bizitza sexuala	Pazientearen bizitza sexualari buruzko galderak	Bizitza sexual aktiboa duzu?	10	0	7
Guztira					

2. taula: Intentzio kategoria nagusien deskribapena.

2.2. Esperimentuen diseinua

Hurrengo azpiataletan eredu onena aukeratzeko burututako esperimentuak azaltzen dira. Lehenengo, 2.2.1. azpiatalean, eskuragarri dauden eredu guztietatik egokiena aurkitzeko egin diren esperimentuak biltzen dira. Esperimentu horien artean tamaina, arkitektura, aurre-entrenatutako datu desberdinak eta amaierako emaitza lortzeko paradigma desberdinak probatu dira. 2.2.2. azpiatala, berriz, proiektuan sortutako corpusetan zentratzen da; hau da, 2.1. atalean azaldutako corpusaren aldaera guztiak erabiltzen dira ereduaren azken doiketa egiteko.

2.2.1. Ereduen, aurre-entrenamenduaren eta paradigma desberdinen esperimentuak

Hasteko, VIR-PAT-LNS corpusaren gainean esperimentuak egiteko VIR-PAT-QA corpuseko distribuzio berdina erabili da, hau da, datuen %75 erabili da ereduaren entrenamendurako (*train*), %10 garapenerako (*dev*) eta %15 testerako (test), baina *b* motako galderak kenduta (ikusi 1. taula). Gure datu kantitatea txikia ez izan arren, gaur egun erabiltzen diren eredu erraldoien bezalakorik zerotik entrenatzeko ez daukagu nahiko daturik. Beraz, hori konpontzeko aurre-entrenatutako ereduak erabiltzea erabaki da, zehatzago esanda, transformadoreetan[2] [6] oinarritutako sistemak. Izan ere, aurre-entrenatutako ereduek aurretiko ezagutza zabala dute, testu eta dokumentu ugari erabiliz prestatuak izan baitira. Hori dela eta, hizkuntzaren egitura eta testuinguruak ulertzeko gai dira, eta ataza desberdinetan aplikatzeko prest daude datu kantitate handiak behar izan gabe. Eskuragarri dauden eredu guztien artean, T5 eta FLOR familia ereduak aukeratu ditugu.

2.2.1.1. mT5 kodetzaile-deskodetzailea

Aukeratutako lehenengo eredu familia T5 [7] izan da, kodetzaile-deskodetzaile arkitektura duena. Zehazki, familia honen barietate eleaniztuna aukeratu da, mT5 familia [8], gaztelaniazko testuekin entrenatuta izan delako. Hori dela eta, gaztelaniazko testuingurua eta sintaxia ulertzeko gai da eta, hortaz, gure helbururako aproposa. Gainera, mT5 familiakoen artean, gaztelaniazko testu medikuekin entrenatuta dagoen eredu bat existitzen da, *Medical mT5*, atazarako are aproposagoa [9]. Eredu horien bitartez hainbat esperimentu burutu dira:

- mT5 originala eta *Medical mT5*-en artean konparaketak egin dira, domeinu medikoan entrenatutako eredu batek gure atazan domeinu orokorrekoak baino emaitza hobeak lor ditzakeen ikusteko.
- mT5 eta *Medical mT5* ereduen tamaina desberdinak ebaluatu dira, tamainak emaitzetan duen eragina ikusteko. Kasu honetan, bi tamaina desberdin erabili dira eredu bakoitzean, *large* (L) eta *extra large* (XL). mT5-L-ek 1,2 mila milioi parametro ditu eta mT5-XL-ek 3,7

[2]Transformadoreak ikasketa sakoneko algoritmo aurreratuenak eta emaitza onenak ematen dituztenak dira.

mila milioi. *Medical mT5*-ek, berriz, 738 milioi eta 3 mila milioi parametro ditu L eta XL tamainetan, hurrenez hurren.

- *Low-Rank Adaptation of Large Language Models* (LoRa) [10] izeneko optimizazio teknika aplikatu da. Izan ere, eredu oso handiekin lanean ari garenez, konputazio-indar handia behar da, eta batzuetan karga hain handia da, ingurune batzuetan ezin dela egikaritu ere. LoRa erabiliz, doiketa-prozesua arintzen da eta baliabide-kontsumoa modu nabarmenean murrizten da, ereduak entrenatzeko eta exekutatzeko behar den tamaina eta kostua txikiagotuz. Hortaz, horretaz baliatuz, optimizazioak ereduen emaitzetan duen eragina ebaluatu egin da.

2.2.1.2. FLOR deskodetzailea

Esperimentuak osatzeko, deskodetzaile soila (*decoder-only*) eredua jarraitzen duten ereduak ere ebaluatzea erabaki da. Azken finean, mota honetako ereduek abantaila handi bat dute besteekiko: ez dira berriro entrenatu behar, hau da, kodetzaile-deskodetzaile ereduak burutu nahi den atazarako sortutako corpusaren gainean doitu egin behar dira (*fine-tuning*), baina deskodetzaile soileko ereduak gai dira doiketarik gabe ataza burutzeko.

Horretarako, mota horretako FLOR[3] (6,3 mila milioi parametroko tamaina) eredua aukeratu da, hori ere gaztelaniazko testuekin aurre-entrenatuta dagoelako, baina ez da aurkitu arkitektura berdina duen medikuntzan aurre-entrenatutako eredurik. Mota honetako ereduekin beste lau esperimentu garatu dira, lehenengo biak ereduaren aurre-entrenamenduarekin loturikoak, eta beste biak, berriz, ereduaren irteera lortzeko erabiltzen diren paradigma desberdinekin lotuta:

- Eredu hauek entrenatzeko bi modu daude: Instrukzioekin Egokituta (*Instructed*) edo ez. Instrukzioekin egokitutako ereduak gai dira agindu batetik abiatuta ataza betetzeko. Entrenamendu hori jaso ez dutenek, aldiz, ez dute aginduak modu naturalean ulertzeko gaitasunik; hortaz, testua jarraitzeko edo osatzeko joera izango dute, interpretazio espliziturik gabe. Esperimentu hauen helburua, bi entrenamendu mota hauetatik, gure atazan zeinek funtzionatzen duen hobeto ikustea da.

- Deskodetzaile soila motako eredu baten errendimendua ebaluatzeko, ikasketa-adibiderik gabeko (*Zero-shot*) [11] eta ikasketa-adibide bakarreko (*One-shot*) [12] paradigmak erabil daitezke. Ikasketa-adibiderik gabeko paradigman, ereduak ataza bat burutu behar du aurretik inolako adibiderik jaso gabe. Adibide bakarreko ikasketan, aldiz, ereduari adibide bakarra ematen zaio atazari egokitzen laguntzeko.

2.2.2. Datuen aldaera desberdinak

Azpiatal honetan, sarreraren arabera emaitzetan dauden aldaketak ikusteko, datu-mota desberdinekin esperimentuak egin dira. Hurrengo probak burutu dira:

- **Corpus aldaerak**. Ereduen doiketa-prozesua 2.2.1. atalean azaldutako hiru corpus aldaerekin burutu da (ikusi 3 taula):
  - Lehenengo aldaera: Doiketa-prozesuan ereduek bakarrik txosteneko esaldia dute sarrera gisa (Erantzuna).
  - Bigarren aldaera: Doiketa-prozesuan ereduek txosteneko esaldi bat izateaz gain, txosteneko zati hori lortzeko erabili den galdera ere badute sarrera gisa (Galdera, Erantzuna).

[3] https://huggingface.co/projecte-aina/FLOR-6.3B

– Hirugarren aldaera: Bigarren aldaerak dituen eremuetaz gain, medikuaren galderaren intentzioa ere gehitu zaio sarrerari (Intentzioa, Galdera, Erantzuna). Horrez gain, aldaera honekin beste formatu bat probatu da, non token[4] bereizle bat (</s>) jarri den eremu desberdinen artean (Intentzioa </s> Galdera </s> Erantzuna).

Aldaera	Formatua	Pazientearen erantzuna
1	[*Erantzuna*] ["Pazienteak 38,5 ºC-ko tenperatura du."]	"Bai, badaukat."
2	[*Galdera, Erantzuna*] ["Badaukazu sukarrik?", "Pazienteak 38,5 ºC-ko tenperatura du."]	
3	[*Intentzioa, Galdera, Erantzuna*] ["Sintoma", "Badaukazu sukarrik?", "Pazienteak 38,5 ºC-ko tenperatura du."] [*Intentzioa </s> Galdera </s> Erantzuna*] ["Sintoma" </s> "Badaukazu sukarrik?" </s> "Pazienteak 38,5 ºC-ko tenperatura du."]	

3. taula: VIR-PAT-LNS corpusaren aldaera desberdinen adibideak.

- **Datu gehikuntza**. Adimen artifizialak duen sorkuntza-ahalmena dela eta, datu gehikuntza (*Data Augmentation*) teknika probatu egingo da, datu gehiagorekin ereduek emaitza hobeak lortzen dituzten ikusteko. Alabaina, teknika honek konputazio-denbora modu nabarmenean luzatzea dakar eta, horregatik, lehenengo eredu onenak aukeratzeko esperimentuak jada azaldu ditugun corpusekin (ikusi 2.2.2.) egin dira. Horrela, gehikuntza aplikatuz lortutako datu-sorta berria bakarrik onenekin probatuko da. Hori sortzeko, ChatGPT API-az[5] baliatu gara datu gehiago lortzeko. Adimen artifizial sortzaileari txosteneko esaldi bakoitzetik abiatuta 5, eta bere pareko elkarrizketako zatitik 3 esaldi berri sortzea eskatu diogu. Instantzia bakoitzeko, txosteneko 6 esaldi (1 originala + 5 berri) eta elkarrizketako 4 esaldi (1 originala + 3 berri) ditugunez, konbinazio guztiekin 24 instantzia berri izango genituzke, horietatik 23 berriak izanik. Beraz, entrenamenduan 2981 instantziatik, 71544 instantzia izatera igaro gara.

  Alabaina, token kopuru osoa handitu den arren, hiztegiaren aniztasunaren neurri gisa erabiltzen den *Type-Token Ratio* (TTR)-a gutxitu egin da. Corpus originalak 31.559 token zituen eta TTR 0.1225 zen, hau da, testuan erabili diren hitzen %12,25 inguru bakarrak ziren. Berriz, datu gehikuntza aplikatu ondoren, corpusa 709.662 tokenetara zabaldu da, baina TTR 0.0121-era (%1,21) jaitsi da. Honek adierazten duenez, nahiz eta testu gehiago sortu den, hitz berrien erabilera proportzionalki murriztu egin da eta, beraz, hiztegiaren aberastasuna txikiagoa da erlatiboki. Hori gertatzen da kasu batzuetan esaldiak oso laburrak direlako eta aukera gutxi daudelako esanahi berdineko esaldia sortzeko. Adibidez, esaldia "38 años" (38 urte) bada, ez daude era asko berdina esateko hitzak errepikatu gabe.

  Edozein kasutan, haien kalitatea bermatu behar da. Horregatik, datuen fidagarritasuna egiaztatzeko, gaztelaniazko bi hiztunek sortutako adibideak aztertu dituzte. Ebaluatzaile bakoitzak esaldi-pareak aztertu ditu eta etiketatze prozesuan, ebaluatzaileei esaldi bakoitzeko hiru etiketatik bat esleitzea eskatu zaie: "OK" etiketak adierazten du esaldia zehatza dela; "KO" etiketak adierazten du esaldiak koherentziarik ez duela; eta "-" etiketak, sorkuntzak koherentzia baduela, baina testuinguru kliniko edo elkarrizketa batean ez duela zentzurik. Adibidez, txosten klinikoko testu zatian, "38 urte" esaldiaren ordez, ereduak "38 udaberri" sortzen badu, esaldia gramatikala da, baina txosten klinikoaren testuinguruan ez du zentzurik.

  Ebaluatzaileen arteko adostasuna kalkulatzeko, ITA-ko (*Inter Tagger Agreement*) *Cohen's kappa* [13] erabili da, non 0 balioak adostasun eza adierazten duen eta 1 balioak adostasun osoa. Lortutako puntuazioa 0,441 izan da, beraz, neurrizko adostasuna lortu dela esan

[4] *Token* bat testuko unitate linguistiko bat da, normalean hitz bat edo puntuazio marka bat, testua banatu eta prozesatzeko erabiltzen dena.

[5] Aplikazioen Programaziorako Interfazea (*Application Programming Interface*).

dezakegu. Emaitza hori azaltzeko, ebaluatzaile bakoitzaren pertzepzio subjektiboak izandako eragina kontuan izan behar da, batzuetan testu baten naturaltasuna ebaluatzea hasieran esperotakoa baino zailagoa izan baita. Hala ere, bi anotatzaileek instantzien bi heren baino gehiago sailkatu dituzte guztiz egokitzat, eta horrek adierazten du datuak, oro har, kalitate onekoak direla.

- Azkenik, aurreikusita 2.2.1. atalean azaldutako "erantzun behar ez diren" galdera motek LNS modulua nahastu dezaketela, erabaki da eredu onenekin testak instantzia horiek kenduta ere egitea, haien eragina neurtzeko.

3. Ebaluazioa

Aurretik aipatu bezala, corpusa eta bere aldaerak hiru multzotan banatu dira: entrenamendua, garapena eta test. Entrenamenduko multzoa ereduak entrenatzeko erabiliko da; garapen multzoa esperimentuen aldaketak gidatzeko; eta test multzoa, aldiz, behin sistemak garapen multzoan emaitza onak lortzen dituenean erabiltzen da, inoiz ikusi ez diren datuen igarpenak egiteko. Ereduen errendimendua konparatzeko, gure helburua itzulpen automatikoarekin erlazionatuta dagoenez, esparru horretarako erabiltzen diren metrika desberdinak erabili dira.

Hasteko, garapeneko esperimentuetarako BLEU eta ROUGE-L metrikak erabiliko dira, eta azken hau izango da amaierako erabakia hartzen lagunduko duen nagusia. Horrez gain, bi modelo onenekin esperimentu osagarriak egingo dira test multzoan, eta kasu horietan BERTScore metrika ere erabiliko da, azken epaile modura. Amaitzeko, giza-ebaluazio bat egingo da, ikusteko gure azken bi modelo onenetatik benetan zein den mundu errealean eraginkorrena, metrika automatikoak alde batera utzita.

3.1. BLEU

BLEU (*Bilingual Evaluation Understudy*) metrika testu-sorkuntza automatikoko sistemak ebaluatzeko erabiltzen da, itzulpen automatikoan batez ere [14]. Metrika honek sistemak sortutako testua eta erreferentziazko testua alderatzen ditu, n-gramen[6] antzekotasunean oinarrituta. BLEUren balioak 0 eta 1 artean daude, non 1 balioak bat-etortze osoa adierazten duen eta 0 balioak, berriz, desberdintasun osoa. BLEU kalkulatzeko, (1) ekuazioa erabiltzen da.

$$BLEU = \ BP \, . \, exp \, (\sum_{n=1}^{N} w_n \, log \, p_n) \tag{1}$$

BP (*Brevity Penalty*) laburtasun-zigorra da, eta sistemak emandako testua erreferentziazko testua baino laburragoa denean, emaitza jaisten du. Gainera, formulak n-grama bakoitzaren puntuazioaren batez besteko ponderatua erabiltzen du azken puntuazioa kalkulatzeko.

Bestalde, w_n pisua da, eta n-grama bakoitzaren garrantzia (ponderazioa) adierazten du azken puntuazioan. Hala ere, BLEUren ohiko inplementazioetan n-grama bakoitzari pisu berdina ematea da metodo erabiliena; hau da, pisu estandarra $w_n = 1/N$ da.

Azkenik, p_n balioa sistemaren eta erreferentziazko testuaren artean partekatutako n-gramen proportzioa da, eta erreferentziazko testuaren n-grama guztien kopuruarekiko erlazioan kalkulatzen da; hau da, erreferentziazko testuaren zein proportzio partekatzen den sistemaren testuarekin.

Metrika honen abantaila nagusia da ebaluazio ataza asko errazten duela, guztiz automatikoa delako. Hala ere, BLEU metrika bakartzat hartzea ez da komenigarria; izan ere, n-grametan gehiegi zentratzen denez, horrek ekar dezake esanahi berdina duten, baina modu ezberdinean idatzita

[6]N-grama terminoak N luzerako hitz segida bati egiten dio erreferentzia. Gehien erabiltzen diren n-gramak unigramak, bigramak, trigramak eta 4-gramak dira.

dauden bi esaldik BLEU baxua izatea, sinonimoen edo hitzen ordenaren ondorioz. Esaterako, 4. taulan ikus dezakegu bi esaldik esanahi bera dutela, nahiz eta BLEUk 0 balioa eman.

Sortutako Testua	Erreferentzia	BLEU	ROUGE-L	BERTScore
Atal hauetan, pazientearen datuen emaitzak aztertuko dira.	Ondorengo ataletan, emaitzen balorazioa egingo da pazientearen datuetan.	0,0	0,13	0,82

4. taula: Adibidea hiru metrika desberdinetan ebaluatua.

3.2. ROUGE-L: Azpisekuentzia komunetan luzeena

ROUGE metrikak, BLEUk bezala, bi esaldiren arteko antzekotasuna neurtzen du [15]. Metrika hau gure lanerako interesgarria da, oso garrantzitsua delako transformazio-prozesuan iturburuaren informazioa ez galtzea. ROUGE-L metrikak Azpisekuentzia Komunetan Luzeena (AKL) identifikatzen du, hau da, erreferentziazko esaldiaren eta sistemak lortutako testuaren artean agertzen den sekuentzia luzeena, hitz ordena mantenduz. AKL kontzeptua funtsezkoa da, bai testuen antzekotasuna neurtzeko zein testuen egitura eta koherentzia aintzat hartzeko. ROUGE-L metrikak doitasuna, estaldura, eta F1-puntuazioa metrikak erabiltzen ditu. BLEUrekiko desberdintasun handiena da BLEUk, doitasuna kontuan hartzen duen bitartean, ez duela estaldura aintzat hartzen.

Sortutako testuaren doitasuna eta estaldura kalkulatzeko, sistemak emandako esaldia eta erreferentziazkoa beharko dira. (2) eta (3) ekuazioetan, X aldagaiak erreferentzia adierazten du eta Y aldagaiak sistemak emandako testua. Zenbakitzailean, bi esaldien arteko n-grama luzeenaren hitz-kopurua jarriko da. Izendatzailean, n aldagaiak Y-ren luzera adierazten du, eta m aldagaiak X-ren luzera.

$$Doitasuna_{AKL} = \frac{AKL(X,Y)}{n} \tag{2}$$

$$Estaldura_{AKL} = \frac{AKL(X,Y)}{m} \tag{3}$$

F1-puntuazioa lortzeko (ikusi (5) eta (4) ekuazioak), doitasuna eta estalduraren balioen batez besteko harmonikoa kalkulatzen da. Alabaina, kasu honetan parametro gehigarri bat gehitzen da: β parametroa. β-k $Doitasuna_{AKL}$ eta $Estaldura_{AKL}$ balioen artean dagoen ratioa adierazten du.

$$\beta = \frac{Doitasuna_{akl}}{Estaldura_{akl}} \tag{4}$$

$$F1_{akl} = \frac{(1+\beta^2)Estaldura_{akl}Doitasuna_{akl}}{Estaldura_{akl} + \beta^2 Doitasuna_{akl}} \tag{5}$$

BLEUk bezala, ROUGE-L metrikak 0 eta 1 bitarteko balioak hartzen ditu. Horrela, 0 balioak transformazio bat erreferentziekin alderatuta guztiz okerra dela adierazten du eta 1 balioak, berriz, sortutakoa eta erreferentzia berdinak direla. Metrika hau ere n-grametan zentratzen denez, BLEUrekin gertatzen denaren kasu berdina eman daiteke: emaitza txarra lortzea bi esaldik esanahi berdina izan arren. Hala ere, BLEU baino metrika leunagoa denez, bi esaldiek zerikusirik badute lortutako puntuazioa handiagoa izango da (ikusi 4. taula).

3.3. BERTScore

BERTScore [16] metrikak, bi testu izanda, haien arteko antzekotasun semantikoa neurtzen du, BERT eredu [17] batek sortutako esaldi edo hitzen errepresentazio bektoriala erabiliz. Metrika

hau sortutako testuaren kalitatea ebaluatzeko erabiltzen da, erreferentziazko testu batekin alderatuta. Garrantzitsua da azpimarratzea metrika honen emaitzak erabilitako ereduaren arabera alda daitezkeela, eta bere puntuazioa ez dela fidagarria izango baliabide gutxiko hizkuntzekin. Gure kasuan, datu asko duen hizkuntza batekin ari gara lanean (gaztelaniaz), beraz, BERT ereduak ez du arazorik izango datuak prozesatzeko.

Esaldien arteko bektore antzekotasunak zehazteko, (6) ekuazioko kosinuko antzekotasuna (KA) erabiltzen da non, erreferentzia esaldia, esaldiko hitz bakoitza, eta hitzen bektore zerrendak $x = [x_1, ..., x_k]$, eta $X = [X_1, ..., X_k]$ diren, eta sortutako esaldikoak $\hat{x} = [\hat{x}_1, ..., \hat{x}_l]$, $\hat{X} = [\hat{X}_1, ..., \hat{X}_l]$ hurrenez hurren.

$$KA(X, \hat{X}) = \frac{X_i^T \hat{X}_j}{||X_i||\ ||\hat{X}_j||} \tag{6}$$

BERTScore-k aurretik normalizatutako bektoreak erabiltzen ditu, hortaz, izendatzailea kendu daiteke, konputazio kostua txikitzeko. (7), (8) eta (9) ekuazioek azaltzen dute nola kalkulatzen diren estaldura (E), doitasuna (D) eta F1-puntuazioa, hurrenez hurren. Estaldura kalkulatzeko, erreferentziako bektore bakoitzeko, hautagaiaren bektoreen artean antzekotasun handiena duen bektorea bilatzen da, eta batez bestekoa egiten da erreferentziako bektorearen luzerarekin. Doitasunaren kalkulurako, berriz, hautagaiaren luzera erabiltzen da. Azkenik, F1-puntuazioa doitasuna eta estalduraren batez besteko harmonikoaren bidez kalkulatzen da.

$$E_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} KA(X, \hat{X}) \tag{7}$$

$$D_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} KA(X, \hat{X}) \tag{8}$$

$$F1_{BERT} = 2\frac{D_{BERT} \cdot E_{BERT}}{D_{BERT} + E_{BERT}} \tag{9}$$

F1-puntuazioa 0 eta 1 bitartean dago, non 0k korrelazio eza, eta 1ek korrelazio perfektua adierazten duten. Kasu honetan, metrika honek ez dituenez n-gramak kontuan hartzen, baizik eta esaldiaren esanahia, esanahi bera duten bi esaldi desberdin ebaluatzekotan, aurreko metrikekin baino emaitza hobeak lor daitezke (ikusi 4. taula).

3.4. Giza-ebaluazioa

Azkenik, LNS eredu onenaren emaitzak paziente baten erantzunak simulatzeko nahiko onak direla egiaztatzeko, osasun-profesional baten eta adimen artifizialean aditu baten iritziak eskatu ditugu, ereduaren erantzunen ezaugarri desberdinak baloratzeko [18]. Ebaluatzaile bakoitzari medikuaren galdera, elkarrizketako pazientearen erantzuna eta ereduak sortutako erantzuna parekatuta eman zaizkio. Horren arabera, hiru ezaugarri ezberdin ebaluatu behar izan dituzte, ezaugarri bakoitza 1 (txarra) eta 5 (bikaina) bitartean puntuatu delarik:

- **Jariakortasuna:** Esaldiaren egitura morfosintaktikoa ebaluatzeko.
- **Zuzentasuna:** Zuzentasuna edo egiazkotasuna ebaluatzeko, hau da, sortutako erantzunak txosten klinikotik erauzitakoak ematen duen informazio guztia aipatzen duen ebaluatzeko. Emaitza paziente errealak emandakoaren desberdina bada, puntuazio baxua izango du.
- **Naturaltasuna:** Sortutako esaldiak paziente erreal batek esango lukeen neurtzeko. Adibidez, “30 minutu” esan beharrean, “1800 segundo” esatea, esanahi berdina izan arren, ez da naturaltzat hartzen.

Ebaluatzaile bakoitzari hainbat instantzia emango zaizkio, eta ezaugarri guztien eta ebaluatzaileen arteko batez bestekoak kalkulatuko dira amaierako puntuazioa lortzeko. Gainera, ebaluatzaileen arteko adostasun maila ere ebaluatu egingo da.

4. Emaitzak

Hurrengo ataletan, garatutako ereduen emaitzak aztertuko dira, bai garapen-multzoan, zein test-multzoan (4.1. eta 4.2. azpiatalak). Emaitza horien interpretazioa ere egingo da, metrika bakoitzak ereduen errendimendua nola islatu dezakeen azpimarratuz. Halaber, esperimentuen arteko alderaketak egingo dira, egindako hautapenek eta doikuntzek izandako eragina ulertzeko. Amaitzeko, eredu guztien artean onena erabakiko da eta horren giza ebaluazioa egingo da (ikusi 4.3. azpiatala).

4.1. Garapen-multzoko emaitzak

5. taulan, eredu bakoitzak garapen-multzoan lortutako puntuazioak daude. Zalantza kasuan, eredu onena aukeratzeko ROUGE-L metrika erabili da, esanahi semantikoa BLEUk baino gehiago kontuan hartzen duelako.

Eredua	Tamaina	Domeinua	Intentzioa	Galdera	Banatzailea	LoRa	ROUGE-L	BLEU
mT5-L	L	Orokorra	-	-	-	-	0,3622	0,0273
mT5-L-Gal	L	Orokorra	-	✓	-	-	0,3821	0,0325
mT5-L-Int-Gal-Sep	L	Orokorra	✓	✓	✓	-	0,3921	0,0423
mT5-L-Int-Gal-Sep-LoRa	L	Orokorra	✓	✓	✓	✓	0,3912	0,0386
mT5-L-Med	L	Medikuntza	-	-	-	-	0,3667	0,0318
mT5-L-Med-Gal	L	Medikuntza	-	✓	-	-	0,3842	0,0411
mT5-L-Med-Int-Gal-Sep	L	Medikuntza	✓	✓	✓	-	0,3907	0,0318
mT5-L-Med-Int-Gal-Sep-LoRa	L	Medikuntza	✓	✓	✓	✓	0,3924	0,0359
mT5-XL	XL	Orokorra	-	-	-	-	0,3668	0,0313
mT5-XL-Gal	XL	Orokorra	-	✓	-	-	0,3843	0,0248
mT5-XL-Int-Gal-LoRa	XL	Orokorra	✓	✓	-	✓	0,3903	0,0385
mT5-XL-Int-Gal-Sep	XL	Orokorra	✓	✓	✓	-	0,3773	0,0433
mT5-XL-Int-Gal-Sep-LoRa	XL	Orokorra	✓	✓	✓	✓	**0,3972**	0,0335
mT5-XL-Med	XL	Medikuntza	-	-	-	-	0,3650	0,0295
mT5-XL-Med-Gal	XL	Medikuntza	-	✓	-	-	0,3856	0,0393
mT5-XL-Med-Int-Gal-LoRa	XL	Medikuntza	✓	✓	-	✓	0,3878	0,0378
mT5-XL-Med-Int-Gal-Sep	XL	Medikuntza	✓	✓	✓	-	0,3726	**0,0451**
mT5-XL-Med-Int-Gal-Sep-LoRa	XL	Medikuntza	✓	✓	✓	✓	0,3910	0,0412

5. taula: mT5 ereduek garapen-multzoaren gainean lortutako emaitzak (L = *Large*, XL = *Extra Large*, Medikuntza = domeinu medikoa, Int = Intentzioa sartuta, Gal = galdera sartuta, Sep = bereizlea sartuta, LoRa = eredu optimizatua, Ins = instrukzioen bidezko eredua).

Eredua	Entrenamendua Instrukzioekin	Hurbilketa	ROUGE-L	BLEU
Flor-Instructed-OneShot	✓	Zero-shot	0.0954	**0.0033**
Flor-Instructed-ZeroShot	✓	One-shot	**0.1017**	0.0022
Flor-OneShot	-	Zero-shot	0.0627	0.0010
Flor-ZeroShot	-	One-shot	0.0575	0.0027

6. taula: FLOR guztiek garapen multzoaren gainean lortutako emaitzak.

4.1.1. mT5

5. taula aztertuz gero, ikusten da ereduaren tamainak eta domeinu medikoan egindako egokitzapenak (Med) ez dutela eragin handirik izan ROUGE-L eta BLEU metriken emaitzetan. Corpusari galderak gehitzerakoan, aldiz, (Gal) eredu guztiak hobetzen dira bi metriketan. Hau espero zitekeen, ikusi dugulako erantzun egokiak lortzeko kasu batzuetan ez dela pazientearen erantzuna bakarrik behar, eta medikuak egindako galderak testuingurua ulertzeko informazio garrantzitsua

ematen du. Corpusari intentzioak (Int) gehituz gero, bi metrikak hobetzen dira. Horrez gain bereizlea (Sep) erabiltzerakoan emaitzak hobetu egiten dira neurri txikiago batean.

LoRa erabili duten ereduen emaitzak aztertzen baditugu, ikusten da optimizazio-teknika honek ez duela inolako eragin negatiborik izan; are gehiago, kasu gehienetan teknika hau erabiltzeak emaitzak hobetu egin ditu, izan ere, emaitza onenen parekoak lortzen dira. Horrez gain, doiketa-prozesua nabarmen arindu da, eta eredu handietan eragin handiena du, doiketa-denbora ia erdira murriztuz.

4.1.2. FLOR

Deskodetzaile soileko ereduen emaitzek beherakada nabarmena erakusten dute mT5 familiakoekin alderatuta. Gainera, instrukziorik gabekoek instrukzioekin (Ins) egokitutakoek baino emaitza baxuagoak lortu dituzte. Beste hurbilpenari dagokionez, adibiderik gabeko (*Zero-Shot*) edo bakarreko (*One-Shot*) tekniken artean ez dago desberdintasun esanguratsurik. Esan beharra dago eredu honek gehiegizko informazioa emateko joera duela eta, batzuetan, oso erantzun naturalak sortu arren, ez daude erreferentziatik hurbil. Hori dela eta, metrikek jokabide hori zigortzen dute, erreferentziatik aldentzeak eta informazioa gehitzeak bat-etortze esplizitua murrizten dutelako. Honek eragin dezake deskodetzaile soileko ereduek puntuazio baxuagoak lortzea.

4.2. Test-multzoko emaitzak

Garapen-multzoko esperimentuen emaitzak aztertu ostean, eredurik onena *mT5-XL-Int-Gal-Sep-LoRa* da, ROUGE-L altuena duen eredua delako. *XL* tamainako eredu bat denez, *large* tamainako eredu onena ere aukeratzea erabaki da (*mT5-L-Med-Int-Gal-Sep-LoRa*). Bi eredu horiek test multzoaren gainean ebaluatu dira haien errendimendua neurtzeko. 7. taulan ikusten den bezala, *mT5-L-Med-Int-Gal-Sep-LoRa* ereduak emaitza hobeak lortu ditu, nahiz eta bi ereduen emaitzen arteko diferentzia txikia den ROUGE-L metrikaren bidez.

Eredua	Tamaina	Domeinua	Intentzioa	Galdera	Banatzailea	LoRa	ROUGE-L	BLEU
mT5-XL-Int-Gal-Sep-LoRa	XL	Orokorra	✓	✓	✓	✓	0,3646	0,0339
mT5-L-Med-Int-Gal-Sep-LoRa	L	Medikuntza	✓	✓	✓	✓	**0,3682**	**0,0385**

7. taula: Eredu onenen emaitzak test multzoaren gainean

Amaitzeko, bi sistema onenekin azken esperimentuak egin dira, datu gehikuntzak (DG) eta "erantzuna behar ez duten" galdera motak kentzearen (EG) eragina neurtzeko (ikusi 8. taula). Esperimentu honetan, BERTScore metrika ere erabili da BLEU eta ROUGE-L baino semantikoagoa delako. Metrika hau ez dugu aurreko esperimentuetan erabili, kostu konputazional altua duelako.

8. taulan ikusten denez, datu gehikuntza izan duten ereduek puntuazio txarragoak lortu dituzte ROUGE-L metrikan, nahiz eta BERTScore-n ez dagoen eragin nabaririk. Gure hipotesia da ereduek lexikoa ikasi dutela eta, ondorioz, erantzuna ez dela erreferentziaren hain antzekoa.

EG ereduek hobekuntza txikia lortzen dute ROUGE-L eta BLEU metrikekin alderatuz gero, baina BERTScore-aren puntuazioa apur bat txikitzen da. Hori, neurri batean, EG dituzten ereduen erantzunak motzagoak direlako gerta daiteke. Izan ere, ROUGE-L eta BLEU metrikek n-gramen antzekotasuna neurtzen dutenez, esaldi laburragoek bat-etortze handiagoa izan dezakete. Aldiz, BERTScore-k semantika osoa aztertzen duenez, erantzunak motzagoak izateak puntuazioa apur bat jaistea ekar dezake. Taulan ikusten da berriro eredurik onenak *mT5-L-Med-Int-Gal-Sep-LoRa* eta bere EG aldaera direla (ikusi 8. taula), ROUGE-L eta BERTScore altuenak lortzen dituztelako.

4.3. Giza-ebaluazioa

Giza-ebaluazioa egiteko, eredu onena eta haren datu gehikuntzaren bidez sortutako bertsioa ebaluatu dira: *mT5-L-Med-Int-Gal-Sep-LoRa* eta *mT5-L-Med-Int-Gal-Sep-LoRa-DG*. Bi anota-

Eredua	EG	DG	ROUGE-L	BLEU	BERTScore
mT5-L-Med-Int-Gal-Sep-LoRa	-	-	0,3682	0,0385	**0,7253**
mT5-L-Med-Int-Gal-Sep-LoRa-EG	✓	-	**0,3725**	0,0442	0,7228
mT5-L-Med-Int-Gal-Sep-LoRa-DG	-	✓	0,3265	0,0425	0,7184
mT5-L-Med-Int-Gal-Sep-LoRa-DG-EG	✓	✓	0,3313	0,0434	0,7154
mT5-XL-Int-Gal-Sep-LoRa	-	-	0,3646	0,0339	0,7202
mT5-XL-Int-Gal-Sep-LoRa-EG	✓	-	0,3705	0,0365	0,7170
mT5-XL-Int-Gal-Sep-LoRa-DG	-	✓	0,3158	0,0450	0,7190
mT5-XL-Int-Gal-Sep-LoRa-DG-EG	✓	✓	0,3271	**0,0451**	0,7183

8. taula: Azken esperimentuen emaitzak test multzoan, BERTScore metrika gehituta (DG = *Datu Gehikuntza*, EG = erantzuna behar ez duten galderak kontuan hartu gabe).

tzailek ebaluatu dute, esparru medikoko profesionala eta adimen artifizialean aditu bat. Anotatzaile bakoitzak test multzoko 20 instantzia ebaluatu ditu eredu bakoitzerako, hau da, guztira 40 instantzia anotatzaile bakoitzeko.

Anotatzaileen arteko adostasuna *Cohen's kappa* neurriaren bidez ebaluatu da, emaitza hauekin (0tik 1erako eskalan): jariakortasuna 0,88, zuzentasuna 0,54 eta naturaltasuna 0,78. Puntuazio hauek aztertuta, ondoriozta daiteke zuzentasuna dela adostasun txikiena duen ezaugarria, osasunarloan aditua den ebaluatzaileak puntuazio hobeak eman dituelako adimen artifizialeko adituarekin alderatuta. Desadostasun hori eman da ereduak benetako pazienteak emandako informazio osoa ez duenean eman. Izan ere, baliteke osasun-profesionalarentzat ereduak eman ez duen informazioa medikuaren galderarako hain garrantzitsua ez izatea, baina, adimen artifizialean adituak ez dakienez benetan informazio hori garrantzitsua den edo ez, gehiago zigortu izatea.

Nahiz eta antzeko puntuazioak izan (ikusi 9. taula), ikusi daiteke datu gehikuntzaren bidez sortutako bertsioak jariakortasunean eta naturaltasunean emaitza hobeak, baina zuzentasunean puntuazio baxuagoak lortzen dituela beste ereduarekin alderatuta. Hori izan daiteke datuak gehitzean ereduak egitura gehiago ikusi dituelako, eta hortaz, jariakortasun eta naturaltasun handiagoa lortu duelako. Hala ere, datu gehiago izateak haluzinazioak gertatzeko arriskua handitzen du, eta horrek zuzentasunaren puntuazioa era negatiboan eragin dezake.

Eredua	Jariakortasuna	Zuzentasuna	Naturaltasuna
mT5-L-Med-Int-gal-Sep-LoRa	4,50	**3,48**	4,33
mT5-L-Med-Int-Gal-Sep-LoRa-DG	**4,85**	3,35	**4,63**

9. taula: Giza-ebaluazioaren puntuazioak.

4.3.1. Erroreen Analisia

LNS ereduek egindako erroreak aztertzeak metriken puntuazioa ulertarazten lagunduko digu eta, aldi berean, ereduen alde onak eta ahuleziak ere erakutsiko dizkigu. Analisia eginda, ikusi dugu erroreak ereduaren araberakoak direla, T5 eta FLOR ereduek errore desberdinak egiten baitituzte. Horrez gain, esan beharra dago ereduen arteko beste desberdintasunek ez dutela eragin nabaririk erroreetan:

- **mT5 ereduen erroreak.**
  - Propagazio erroreak: Erantzun Erauzketa (EE) modulutik jasotzen ditugun erantzunak sarreratzat erabiltzen ditugunez, EE moduluak huts egiten badu bere iragarpenetan, errore hori LNS eredura igaroko da. Adibidez, EE moduluak "ez" erantzuten badu, baina erantzun zuzena baiezkoa izango balitz, LNS ereduak "ez" hori erabiliko luke benetako erantzuna erabili beharrean errorearen propagazioa eginez.

- Kontraesan erroreak: Batzuetan, LNS ereduak "bai" edo "ez" gehitzen ditu erantzunen hasieran, elkarrizketetan ohikoak diren partikulak direlako eta giza-itxura handiagoa ematen dutelako. Hala ere, zenbaitetan partikulak gaizki erabiltzen ditu hasieran, eta jarraian datorren erantzunaren zatia kontraesanean dago ezezko partikularekin. Adibidez:

  (1) Galdera: *¿Zure mukiak berdeak al dira?*
  Erauzketa: *muki berdeak ditu*
  Erantzuna: *Ez, nire mukiak berdeak dira.*

- Metriken mugetatik sortutako erroreak: Metrikek erantzun batzuk gehiegi zigortzen dituzte. Kasurik nabarmenena da ereduak erantzun labur bat iragartzen duenean eta helburua luzea denean edo alderantziz. 2 adibidean, helburuko eta iragarritako erantzunak bideragarriak dira eta zuzenak izan daitezke, baina metrikak ez du gaitasunik hori neurtzeko. Horren ordez, iragarpena erreferentziarekin oso urrun dagoela jotzen du, ia guztiz okerra dela baloratuz.

  (2) Elkarrizketako erantzuna: *Ez*
  Pazientearen erantzuna: *Ez, ez dut sukarra.*

- **FLOR modeloen erroreak:**
  - Zentzurik gabeko erroreak: Instrukzioekin entrenatu ez diren ereduen iragarpenek, maiz, galderarekin edo medikuntzaren domeinuko edukiarekin zerikusirik ez duten erantzunak sortzen dituzte, eta beste kasu batzuetan sortutako erantzunek ez dute egitura sintaktiko zuzenik.
  - Haluzinazioak: Mota honetako ereduekin ohikoa den moduan, haluzinazioak agertzen dira, hau da, benetakoak diruditen baina egiazkoak ez diren erantzunak. Gure modeloen erantzunetan ere halako kasuak ikusi ditugu. Batzuetan, pazientearen erantzuna simulatu beharrean, ereduek medikuaren edo erizainaren rola hartzen dute. Beste batzuetan, ereduek paziente gisa ondo egiten dute, baina txosteneko erantzunean ez zegoen informazio gehigarria sartzen dute.
  - Metriken mugak sortutako erroreak: Errore mota hau T5 ereduetan azaldutakoen berdina da. FLOR ereduek erantzun luzeagoak sortzeko joera dute, eta hori batzuetan zigortua izaten da, egokia izan arren.

4.4. Metriken inguruko eztabaida

Ereduek lortutako emaitzak hobeto ulertzeko, 10. taula erabiliko da azalpenaren oinarri gisa. Lehenik eta behin, esan beharra dago BLEU metrikak, orokorrean, puntuazio baxuak ematen dituela gizakien ebaluazioarekin alderatuta. Izan ere, esaldiak ia esanahi bera izan arren, BLEU metrikak 0tik hurbileko puntuazioa ematen du. Adibidez, "Zigarroak ez, inoiz ez" eta "Ez, ez dut erretzen" esaldiek antzeko esanahia dute, baina BLEU metrikak ez du 0 baino puntuazio handiagorik ematen. Horrek erakusten du metrika honek testuaren egitura edo hitzen arteko antzekotasuna zorrotz neurtzen duela; hau da, n-grametan guztiz oinarrituta dago, baina ez du beti esanahiaren baliokidetasuna kontuan hartzen, gizakiek egiten duten moduan.

Bestalde, ROUGE-L metrika ere, n-grametan oinarritzen den arren, bere kalkuluak BLEUk baino hurbilpen semantikoagoa ahalbidetzen du. Hala ere, ez dator guztiz bat giza ebaluazioarekin,

hori dela eta, esanahiaren antzekotasuna neurtzeko modu zehatzagoak behar dira, testuinguruaren eta hitzen antzekotasuna kontuan hartzeko.

Horregatik, erabaki da BERTScore ere erabiltzea, testuaren errepresentazio semantiko sakonagoa eskaintzen duelako. Emaitzak alderatuta, ikusi da BERTScore-aren balorazioak giza ebaluazioaren antzekoagoak direla, testuingurua eta antzekotasuna kontuan hartzen dituelako. Adibidez, 10. taulako lehenengo eta bigarren adibideetan ikus daiteke nola bi esaldik, nahiz eta egitura desberdina izan, BERTScore altua lortzen duten, esanahi oso antzekoa dutelako.

Adibideetan ikusten da ereduak gehienetan zigortuak izan direla pazientearen informazio guztia ez dutelako islatu bere esaldietan. Bi arrazoi nagusirengatik gerta daiteke hau. Lehenengoa, txostenean ez delako gorde pazienteak esandako informazio guztia, adibidez 10. taulako bigarren adibidean txostenean ez da idatzi marihuana lagunekin kontsumitzen duela, seguruenik medikuaren ikuspuntutik ez delako garrantzitsua, eta giza ebaluazioarenetik ezta, ondo ebaluatua izan delako. Bigarren arrazoia, ereduak erabakiak hartu behar dituenean gertatzen da, hau da, eredua ez da beti gai informazio guztia behar bezala islatzeko, eta hori bereziki nabariagoa da galdera luzeetan. Hala ere, galdera laburretan ere gerta daiteke. Adibidez, azken adibidean, ereduak ondoezari buruzko aipamenik egin ez duela ikusten da, nekean bakarrik zentratu delako.

Galdera	Txosteneko Esaldia	Pazientearen Esaldia	Ereduaren Esaldia	ROUGE-L	BLEU	BERTScore	Giza Ebaluazioa (1-5)
Ok, ¿y usa alguna medicación para el asma? Ondo, eta asmarako medikamendurik erabiltzen duzu?	Usa el inhalador azul ocasionalmente Noizean behin inhalagailu urdina erabiltzen du	Sí, uso ese inhalador azul a veces. Supongo que sólo cuando lo necesito, pero no muy a menudo. Bai, batzuetan inhalagailu urdin hori erabiltzen dut. Behar dudanean bakarrik, baina ez oso maiz.	Sí, utilizo el inhalador azul como único tratamiento. Bai, inhalagailu urdina erabiltzen dut tratamendu bakar gisa.	0,22	0,00	0,76	4,67
Vale, ¿y consume alguna droga como la marihuana? Bale, eta drogarik kontsumitzen duzu, marihuana adibidez?	Consume marihuana 1-2 veces por semana Astean 1-2 alditan kontsumitzen du marihuana	Sí, a veces sí, fumo como una o dos veces por semana, con eso con un amigo o algo así. Bai, batzuetan bai, astean behin edo bitan erretzen dut, lagun batekin, edo halako zerbait	Um, tomo como una o dos veces por semana. Um, astean behin edo bitan hartzen dut.	0,48	0,16	0,80	3,83
Ok, ¿y fuma? Ondo, eta erretzen al duzu?	no fuma ez du erretzen	Cigarrillos no, nunca. Zigarroak ez, inoiz ez.	No, no fumo. Ez, ez dut erretzen	0,33	0,00	0,76	4,33
Ok y que tipo de dolor diría que es, ¿es como un dolor sordo? ¿Es un dolor agudo y punzante? ¿Cómo lo describiría? Ondo eta zein min mota dela esango zenuke, minsorra bezalakoa da? Min zorrotz eta ziztagarria da? Nola deskribatuko zenuke?	sordo sorra	Es un dolor sordo. Min sorra da	Es como un dolor sordo. Min sor modukoa da	0,88	0,54	0,94	4,50
¿Algún cansancio o malestar? Nekerik edo ondoezik?	dice tener fatiga y malestar nekea eta ondoeza dituela esaten du	Sí, he estado un poco más fatigado de lo habitual, y sí, he tenido algo de malestar también. Bai, ohi baino apur bat nekatuago egon naiz, eta bai, ondoez pixka bat ere izan dut	Sí, me he sentido bastante cansada. Bai, nahiko nekatuta sentitu naiz.	0,15	0,00	0,78	4,83

10. taula: Erantzun desberdinen alderaketa metrika bakoitzean (giza ebaluazioa jariakortasuna, zuzentasuna eta naturaltasunaren batez bestekoa da)

5. Ondorioak eta etorkizuneko lana

Artikulu honetan, elkarrizketa-sistema bat osatzeko oinarrizko baliabideak eta tresnak aurkeztu dira, mediku eta pazientearen arteko elkarrizketa modelatzeko eta, bereziki, paziente birtualaren rola hartuko duen gaztelaniazko txatbot baten prototipo batean, lengoaia naturalaren sorkuntzan.

Helburu horrekin, alde batetik baliabideak garatu dira, tresna automatikoak inplementatzeko datuak izatea ezinbestekoa baita. Lehenengo, pazientearen deskribapena duen txosten mediko bateko esaldietatik paziente batek emango lituzkeen esaldien corpus bat osatu da. Hori erabilita, paziente birtualaren papera har dezakeen txatbot baten lengoaia naturalaren sorkuntzarako sistema entrenatu da adimen artifiziala eta sare neuronalak erabiliz. Azkenik, metrika desberdinak erabiliz ebaluatu egin da, emaitza positiboekin, horrela medikuntzako elkarrizketa sistema automatikoak eraikitzeko bidea irekiz.

Lan honek etorkizunerako bide berriak irekitzen ditu elkarrizketa sistema garatzeko. Alde batetik, euskarara egokitzea bideragarria da, orain eskuragarri dauden itzulpen teknikak eta euskarazko galdera-erantzun sistemak erabiliz. Bestalde, erronka handia izango da elkarrizketa-sistemei benetako pazienteen nortasun eta ezaugarri psikologikoak gehitzea, bakarrik pazientearen datu objektiboak lortzeak suposatzen duen zailtasunez gain, erabiltzailearekiko harremana hobea izan dadin.

Eskerrak

Lan hau Virtual Patient 1.0 proiektuaren baitan garatu da, EUSK22/19 kode ofizialarekin, Euskampus Fundazioaren Misiones Euskampus 2.0 Programa 2022 barruan. Horrez gain, lehen egileak UPV/EHUren doktorego-aurreko bekaren (PIF24/223) babesa jaso du lan honen garapenean. Lan hau HiTZ Zentroan egin da, Eusko Jaurlaritza (ikerketa-taldeen deialdia IT1570-22), eta Espainiako Unibertsitate, Zientzia eta Berrikuntzako Ministeritzaren proiektu hauen bidez: EDHIA PID2022-136522OB-C22 (FEDER, UE funtsekin batera), DeepR3 TED2021-130295B-C31 (MCIN/AEI/10.13039/5011 00011033, European Union NextGeneration EU/PRTR funtsekin lagunduta).

Erreferentziak

[1] J. G.-E. LÓPEZ, 2013, «Prueba nacional de evaluación de competencias clínicas de la conferencia nacional de decanos de facultades de medicina de españa», *FEM Rev Fund Educ Médica*, **16**, 59–70.

[2] A. VARGAS GUZMÁN eta M. F. RICO GAVIRIA, 2011, «Los riesgos del actor : cómo lo afectan las variables de su escuela, director y papel», .
URL `http://hdl.handle.net/10554/5809`

[3] J. ARANA, M. IDOYAGA, M. URRUELA, E. ESPINA, A. ATUTXA SALAZAR eta K. GOJENOLA, 2024, «A virtual patient dialogue system based on question-answering on clinical records», N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI eta N. XUE,,Editoreak, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Orrialdeak 2017–2027, ELRA and ICCL, Torino, Italia.
URL `https://aclanthology.org/2024.lrec-main.182/`

[4] A. ISAZA, M. TERESA, G. CIFUENTES ALVAREZ eta A. ARGUELLO, 2018, «The virtual patient as a learning tool: A mixed quantitative qualitative study», *BMC Medical Education*, **18**.

[5] F. FAREEZ, T. PARIKH, C. WAVELL, S. SHAHAB, M. CHEVALIER, S. GOOD, I. DE BLASI, R. RHOUMA, C. MCMAHON, J.-P. LAM, T. LO eta C. W. SMITH, 2022, «A dataset of simulated patient-physician medical interviews with a focus on respiratory cases», *Scientific Data*, **9**(1), 313.
URL `https://doi.org/10.1038/s41597-022-01423-1`

[6] A. VASWANI, N. SHAZEER, N. PARMAR, J. USZKOREIT, L. JONES, A.Ñ. GOMEZ, L. U. KAISER eta I. POLOSUKHIN, 2017, «Attention is all you need», I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN eta R. GARNETT,,Editoreak, *Advances in Neural Information Processing Systems*, Bolumena 30, Curran Associates, Inc.
URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`

[7] C. RAFFEL, N. SHAZEER, A. ROBERTS, K. LEE, S.ÑARANG, M. MATENA, Y. ZHOU, W. LI eta P. J. LIU, 2020, «Exploring the limits of transfer learning with a unified text-to-text transformer», *J Mach Learn Res*, **21**(1).

[8] L. XUE, N. CONSTANT, A. ROBERTS, M. KALE, R. AL-RFOU, A. SIDDHANT, A. BARUA eta C. RAFFEL, 2021, «mt5: A massively multilingual pre-trained text-to-text transformer», .

[9] I. GARCÍA-FERRERO, R. AGERRI, A. A. SALAZAR, E. CABRIO, I. DE LA IGLESIA, A. LAVELLI, B. MAGNINI, B. MOLINET, J. RAMIREZ-ROMERO, G. RIGAU, J. M. VILLA-GONZALEZ, S. VILLATA eta A. ZANINELLO, 2024, «Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain», , Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING).

[10] E. J. HU, Y. SHEN, P. WALLIS, Z. ALLEN-ZHU, Y. LI, S. WANG, L. WANG eta W. CHEN, 2021, «Lora: Low-rank adaptation of large language models», .

[11] B. ROMERA-PAREDES eta P. TORR, 2015, «An embarrassingly simple approach to zero-shot learning», F. BACH eta D. BLEI,,Editoreak, *Proceedings of the 32nd International Conference on Machine Learning*, Bolumena 37 of *Proceedings of Machine Learning Research*, Orrialdeak 2152–2161, PMLR, Lille, France.
URL `https://proceedings.mlr.press/v37/romera-paredes15.html`

[12] T. UCAR, A. GONZALEZ-MARTIN, M. LEE eta A. D. SZWARC, 2020, «One-shot learning for language modelling», .

[13] R. ARTSTEIN eta M. POESIO, 2008, «Inter-Coder Agreement for Computational Linguistics», *Computational Linguistics*, **34**(4), 555–596.
URL `https://doi.org/10.1162/coli.07-034-R2`

[14] K. PAPINENI, S. ROUKOS, T. WARD eta W.-J. ZHU, 2002, «Bleu: a method for automatic evaluation of machine translation», *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, Orria 311–318, Association for Computational Linguistics, USA.
URL `https://doi.org/10.3115/1073083.1073135`

[15] C.-Y. LIN eta F. J. OCH, 2004, «Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics», *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, Orrialdeak 605–612, Barcelona, Spain.
URL `https://aclanthology.org/P04-1077`

[16] T. ZHANG, V. KISHORE, F. WU, K. Q. WEINBERGER eta Y. ARTZI, 2020, «Bertscore: Evaluating text generation with bert», .

[17] J. DEVLIN, M.-W. CHANG, K. LEE eta K. TOUTANOVA, 2019, «Bert: Pre-training of deep bidirectional transformers for language understanding», .

[18] W. LIU, J. TANG, J. QIN, L. XU, Z. LI eta X. LIANG, 2020, «Meddg: A large-scale medical consultation dataset for building medical dialogue system», *CoRR*, **abs/2010.07497**.
URL `https://arxiv.org/abs/2010.07497`