Ascle: A Python Natural Language Processing Toolkit for Medical Text Generation

Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Amisha Dave, Tiarnan Keenan, Yu He Ke, Chuan Hong, Nan Liu, Emily Chew, Dragomir Radev, Zhiyong Lu, Hua Xu, Qingyu Chen, Irene Li

Submitted to: Journal of Medical Internet Research
on: May 16, 2024

Table of Contents

Ascle: A Python Natural Language Processing Toolkit for Medical Text Generation

Rui Yang[1*]; Qingcheng Zeng[2*]; Keen You[3*]; Yujie Qiao[4*]; Lucas Huang[3]; Chia-Chun Hsieh[3]; Benjamin Rosand[3]; Jeremy Goldwasser[3]; Amisha Dave[5]; Tiarnan Keenan[6]; Yu He Ke[7]; Chuan Hong[8]; Nan Liu[1, 9, 10]; Emily Chew[6]; Dragomir Radev[3]; Zhiyong Lu[11]; Hua Xu[12]; Qingyu Chen[12]; Irene Li[13, 14]

[1]Centre for Quantitative Medicine, Duke-NUS Medical School Singapore SG
[2]Department of Linguistics, Northwestern University Evanston US
[3]Department of Computer Science, Yale University New Haven US
[4]Yale School of Public Health, Yale University New Haven US
[5]Yale New Haven Hospital, Yale School of Medicine Yale University New Haven US
[6]Division of Epidemiology and Clinical Applications National Eye Institute, National Institutes of Health Bethesda US
[7]Department of Anesthesiology, Singapore General Hospital Singapore SG
[8]Department of Biostatistics and Bioinformatics Duke University Durham US
[9]Program in Health Services and Systems Research Duke-NUS Medical School Singapore SG
[10]Institute of Data Science, National University of Singapore Singapore SG
[11]National Center for Biotechnology Information, National Library of Medicine National Institutes of Health Bethesda US
[12]Department of Biomedical Informatics and Data Science, Yale School of Medicine Yale University New Haven US
[13]Information Technology Center, University of Tokyo Tokyo Kashiwa JP
[14]Smartor LLC. Tokyo JP
[*]these authors contributed equally

Corresponding Author:
Irene Li
Information Technology Center, University of Tokyo
Tokyo
6-2-3 Kashiwanoha, Chiba
Kashiwa
JP

Abstract

Background: Medical texts present significant domain-specific challenges, and manually curating these texts is a time-consuming and labor-intensive process. Therefore, natural language processing (NLP) algorithms have been developed to automate text processing. In the biomedical field, there are various toolkits for text processing, which have greatly improved the efficiency of handling unstructured text. However, these existing toolkits tend to emphasize different perspectives, and the lack of generation capabilities in any of them leaves a significant void.

Objective: This study introduces Ascle, a pioneering NLP toolkit designed for medical text generation. Ascle is tailored for biomedical researchers and clinical staff with an easy-to-use, all-in-one solution that requires minimal programming expertise. For the first time, Ascle provides four advanced and challenging generative functions: question-answering, text summarization, text simplification, and machine translation. Additionally, Ascle integrates 12 essential NLP functions, along with query and search capabilities for clinical databases.

Methods: We fine-tuned 32 domain-specific language models and evaluated them thoroughly on 27 established benchmarks. Additionally, for the question-answering task, we develop a retrieval-augmented generation (RAG) framework for LLMs that incorporates a medical knowledge graph with ranking techniques to enhance the reliability of generated answers.

Results: The fine-tuned models and RAG framework consistently enhanced text generation tasks. For example, the fine-tuned models improved the machine translation task by 20.27 in terms of BLEU score. In the question-answering task, the RAG framework raised the ROUGE-L score by 18% over the vanilla models.

Conclusions: This study introduces the development and evaluation of Ascle, a user-friendly NLP toolkit designed for medical text generation. All code is publicly available via https://github.com/Yale-LILY/Ascle. All fine-tuned language models can be

accessed via Hugging Face.

(JMIR Preprints 16/05/2024:60601)
DOI: https://doi.org/10.2196/preprints.60601

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Ascle: A Python Natural Language Processing Toolkit for Medical Text Generation

Abstract

Background: Medical texts present significant domain-specific challenges, and manually curating these texts is a time-consuming and labor-intensive process. Therefore, natural language processing (NLP) algorithms have been developed to automate text processing. In the biomedical field, there are various toolkits for text processing, which have greatly improved the efficiency of handling unstructured text. However, these existing toolkits tend to emphasize different perspectives, and the lack of generation capabilities in any of them leaves a significant void.

Objective: This study introduces Ascle, a pioneering NLP toolkit designed for medical text generation. Ascle is tailored for biomedical researchers and clinical staff with an easy-to-use, all-in-one solution that requires minimal programming expertise. For the first time, Ascle provides four advanced and challenging generative functions: question-answering, text summarization, text simplification, and machine translation. Additionally, Ascle integrates 12 essential NLP functions, along with query and search capabilities for clinical databases.

Methods: We fine-tuned 32 domain-specific language models and evaluated them thoroughly on 27 established benchmarks. Additionally, for the question-answering task, we develop a retrieval-augmented generation (RAG) framework for LLMs that incorporates a medical knowledge graph with ranking techniques to enhance the reliability of generated answers.

Results: The fine-tuned models and RAG framework consistently enhanced text generation tasks. For example, the fine-tuned models improved the machine translation task by 20.27 in terms of BLEU score. In the question-answering task, the RAG framework raised the ROUGE-L score by 18% over the vanilla models.

Conclusions: This study introduces the development and evaluation of Ascle, a user-friendly NLP toolkit designed for medical text generation. All code is publicly available via https://github.com/Yale-LILY/Ascle. All fine-tuned language models can be accessed via Hugging Face.

Keywords: natural language processing; machine learning; deep learning; generative artificial intelligence; large language models; retrieval-augmented generation; healthcare

Introduction

Medical texts pose considerable challenges due to their domain-specific nature, including issues such as ambiguities, frequent abbreviations, and specialized terminology [1,2]. The manual curation of these texts is both time-consuming and labor-intensive [2]. Therefore, natural language processing (NLP) algorithms have been developed to automate text processing [2–4]. Recent years have seen a notable shift towards the use of domain-specific pre-trained language models, transitioning from shallow embeddings like BioWordVec [5] and BioSentVec [6] to advanced architectures like Bidirectional Encoder Representations from

Transformers (BERT) [7] such as BioBERT [8], ClinicalBERT [9], and PubMedBERT [10]. Furthermore, large language models (LLMs) such as Med-PaLM [11] and Med-Gemini [12] have demonstrated powerful generative capabilities, possessing exceptional zero- and few-shot performance. These domain-specific language models have substantially enhanced the effectiveness of NLP tasks in the biomedical and clinical domains [13–15].

Despite the success of these advanced methods, their complexity remains a significant barrier to practical application for healthcare professionals lacking basic programming skills. Consequently, there is an increasing demand for user-friendly and accessible toolkits designed to simplify medical text processing. Multiple toolkits for text processing are available in the biomedical domain. Table 1 summarizes representative tools. While there are many other useful tools, here we mainly limit our comparison with Python-based open-source toolkits.

Table 1. A comparison of Ascle with existing Python-based toolkits. The ⋆ indicates that for the question-answering task, we specifically propose a retrieval-augmented generation (RAG) framework for LLMs that incorporates a medical knowledge graph with ranking techniques. Basic NLP Functions include abbreviation extraction, sentence tokenization, word tokenization, negation detection, hyponym detection, UMLS concept extraction, named entity recognition, document clustering, POS tagging, entity linking, text summarization (extractive methods) and multi-choice QA. It is worth noting that not every toolkit includes these 12 basic NLP functions, but Ascle includes them all.

	⋆ Question Answering	Text Summarization	Text Simplification	Machine Translation	Basic NLP Functions	Query Search
MIMIC-Extract [16]						✓
ScispaCy [17]					✓	
MedspaCy [18]					✓	
Transformers-sklearn [19]					✓	
Stanza Biomed [20]					✓	
Ascle (this work)	✓	✓	✓	✓	✓	✓

These existing toolkits tend to emphasize different perspectives, and the absence of generation capabilities in any of them leaves a significant void. In response, we present Ascle, a pioneering NLP toolkit for medical text generation, which for the first time, includes four advanced generative functions. We fine-tuned 32 domain-specific language models and evaluated them thoroughly on 27 established benchmarks. Additionally, for the question-answering task, we develop a retrieval-augmented generation (RAG) framework [21] that combines a medical knowledge graph (The Unified Medical Language System (UMLS)) [22] with ranking techniques, aimed at improving the reliability of long-form answers [15]. We uploaded all fine-tuned language models to Hugging Face and listed 32 fine-tuned language models along with 27 benchmarks in Multimedia Appendix 1 for clearer explanation.

In conclusion, Ascle empowers a diverse spectrum of users, from novices to experienced professionals, enabling them to effortlessly address their NLP tasks, even with limited technical expertise in handling textual data. We believe that Ascle not only democratizes access to cutting-edge methods but also expedites their integration into healthcare.

Methods

Ascle consists of three modules, with the core module being the Generative Functions, including four challenging generation tasks: question-answering, text summarization, text simplification, and machine translation, covering a variety of application scenarios in

healthcare. Additionally, Ascle integrates 12 basic NLP functions, as well as query and search capabilities for clinical databases. The overall architecture of Ascle is shown in Figure 1. This section will focus on introducing the core module of Ascle - Generative Functions. For more information on basic NLP functions and query and search functions within Ascle, please refer to Multimedia Appendix 2 and Multimedia Appendix 3, respectively.

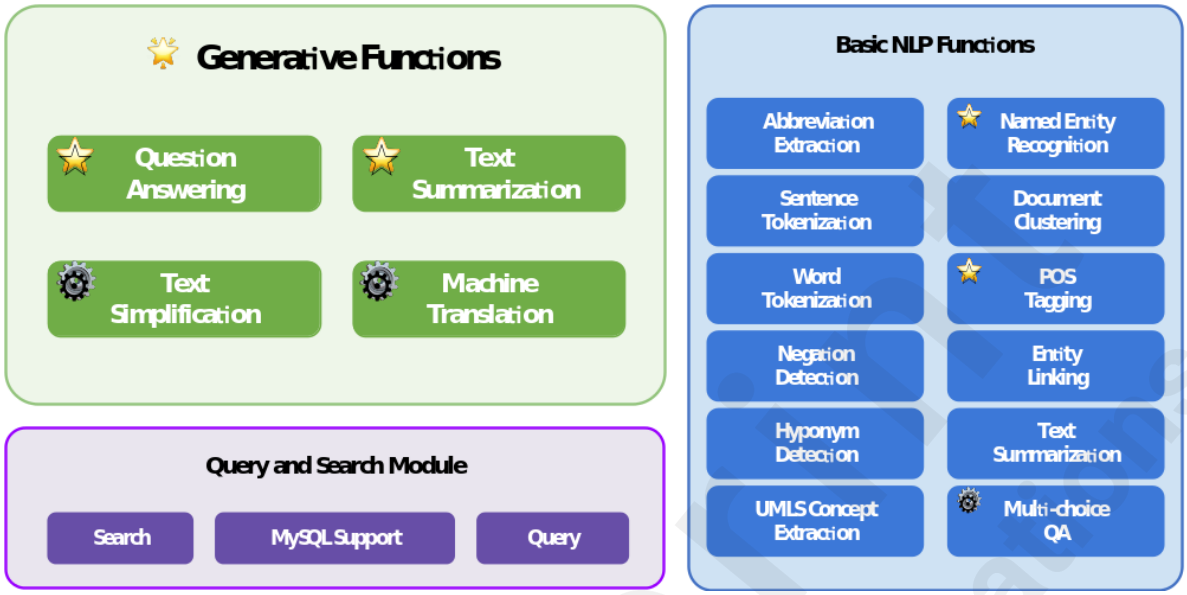


Figure 1. The overall architecture of Ascle. ⚙indicates that we have our fine-tuned language models for this task. ☆indicates that we conducted evaluations for this task.

Generative Functions

Ascle offers a range of generative functions through pre-trained and fine-tuned language models, all of which are publicly available for user access. In the following sections, we will introduce these powerful generative functions separately.

Question-answering

Question-answering is particularly crucial in healthcare [13]. When integrated into healthcare systems, it assumes roles such as pre-consultation and remote consultation, effectively coping with the exponential increase in patient load. Moreover, specialized question-answering systems hold the potential to contribute to medical education [13,21].

In Ascle, we first provide an interface for medical LLMs, such as Baize-healthcare [23], allowing users to utilize them directly. Moreover, we develop a RAG framework that utilizes UMLS with ranking techniques to enhance LLMs in generating long-form answers [21]. Specifically, when receiving a query, the RAG framework first extracts medical entities within the query automatically, and then retrieves related triplets from UMLS for each extracted entity. A triplet consists of two medical concepts and the relation between them, i.e., (Myopia, clinically_associated_with, HYPERGLYCEMIA). Subsequently, the RAG framework employs ranking and re-ranking techniques to refine the ordering of these triples. Finally, the query and the retrieved triplets would be prompted to LLM for inference. For more details about the RAG framework, please refer to Multimedia Appendix 4. We apply this framework to the GPT [24] and LLaMA [25] series of LLMs.

We conducted evaluations on four medical QA datasets, including LiveQA [26], ExpertQA (Med & Bio) [27], and MedicationQA [28]. LiveQA consists of health questions submitted by

consumers to the National Library of Medicine. It includes a training set with 634 QA pairs and a test set comprising 104 QA pairs, and the test set was used for evaluation. ExpertQA is a high-quality long-form QA dataset covering multiple fields, along with answers verified by domain experts. Among them, we utilized 504 medical questions (Med) and 96 biology (Bio) questions for evaluation. MedicationQA includes 690 drug-related consumer questions along with information retrieved from reliable websites and scientific papers.

Text Summarization

In healthcare, clinicians and researchers are confronted with an increasing amount of information, including literature, clinical notes, and more [29,30]. Text summarization is an important generation task, aiming to distill essential information from the overwhelming complexity of texts and compress it into a more concise format [31]. Through automatic text summarization, clinicians and researchers can efficiently acquire information, thereby avoiding information overload.

We provide an abstractive text summarization function and compare general pre-trained summarization models, including Pegasus [32], BigBird [33], BART [34], PRIMERA [35], as well as domain-specific models such as SciFive [36] and BioBART [37], which make use of biomedical corpora. Furthermore, we chose PubMed [38], MIMIC-CXR [39], and MEDIQA-AnS [40] datasets for evaluation. The PubMed dataset consists of biomedical scientific publications from the PubMed database, where each input document is a scientific article with its abstract serving as the ground truth. We reported the evaluation results for the test set, which contains 1.66k examples. MIMIC-CXR is a de-identified dataset of chest radiographs with free-text radiology reports, and we used a subset of MIMIC-CXR that includes 2,000 instances for evaluation. MEDIQA-AnS is a collection of 156 consumer health questions along with passages that contain relevant information. It supports both single-document and multiple-document summarization evaluation.

Text Simplification

Biomedical texts are typically laden with intricate terminologies, which can hinder the understanding of individuals without a clinical background [41]. In Ascle, the function of text simplification is to translate complex and technical biomedical texts into understandable content. This will enhance the comprehension and involvement of non-clinical individuals, including patients, enabling them to better engage with the information and participate in clinical decisions more effectively.

We fine-tuned and evaluated widely-used pre-trained language models on three datasets: eLife, PLOS [42], and MedLane[43]. This included two general models, BigBirdPegasus [32] and BART, as well as a biomedical-specific model, BioBART. The eLife and PLOS are shared task data released from the BioLaySumm 2023 Task 1, which contains biomedical journal articles alongside expert-written lay summaries. We evaluated the validation sets for eLife and PLOS, which contain 241 and 1,376 examples, respectively. MedLane is a large-scale human-annotated dataset containing professional-to-customer sentences selected from MIMIC-III. For MedLane, we utilized the test set for evaluation, which includes 1,016 instances.

Machine Translation

Language barriers pose difficulties for patients to access timely information and communicate effectively with healthcare providers, resulting in low-quality healthcare services [44]. Our machine translation function aims to translate the text from a source language into a target

language in a clinical scenario. By fine-tuning pre-trained language models on the medical corpus, Ascle supports machine translation from English (en) to 8 target languages: Spanish (es), French (fr), Romanian (ro), Czech (cs), German (de), Hungarian (hu), Polish (pl), and Swedish (sv). Here, we only emphasize the 8 languages fine-tuned on medical data, while other languages, such as from English to Chinese, are supported by the pre-trained language models.

We fine-tuned the existing MarianMT [45] and multilingual T5 [46] using UFAL Medical Corpus [47] which includes various medical text sources, such as titles of medical wikipedia articles, medical term-pairs, patents, and documents from the European Medicines Agency. During the pre-processing phase, we excluded general domain data from UFAL, such as parliamentary proceedings, and randomly shuffled the medical-domain corpora, splitting them into two parts at a ratio of 85% and 15% for training and testing, respectively. We reported the results on the test set, the size of which varies from 111,779 to 407,388 depending on the different language pairs. Moreover, for each language pair, we utilized all available parallel data to maximize the breadth and accuracy of our machine translation function.

Results

Overall Performance of Generation Tasks

In the question-answering task, we utilized ROUGE-L [48], BERTScore [49], MoverScore [50] and BLEURT [51] for comprehensive evaluation, and employed GPT-4 and LLaMA2-13b as the vanilla LLMs. As shown in Table 2(A), our RAG framework surpasses the zero-shot setting on all evaluation metrics for the LiveQA, ExpertQA-Bio, ExpertQA-Med, and MedicationQA datasets. Among them, the ROUGE-L score has increased by more than 18% on the ExpertQA-Bio dataset.

For the text summarization task, we evaluated five pre-trained language models on single-document summarization, as shown in Table 2(B). To ensure a fair comparison, we excluded the results of BioBART and SciFive on PubMed, as they were fine-tuned on this dataset. It is worth noting that BART consistently demonstrated strong performance across three benchmarks, while BioBART only outperformed BART in one of the benchmarks. Additionally, we evaluated the multi-document summarization task, discussed the differences between abstractive and extractive methods, as well as the limitations of evaluation metrics, which can be found in the Discussion section.

Regarding the text simplification task, we compared the performance of fine-tuned models and conducted an analysis of readability using the Flesch-Kincaid Grade Level (FKGL) score [52], as indicated in Table 2(C). For the eLife and PLOS datasets, the ground truth exhibits FKGL scores of 12 and 15, respectively. Interestingly, the BioBART model performs competitively in terms of ROUGE metrics but fails to significantly reduce the difficulty of understanding, as evidenced by its FKGL score of 17 in both datasets. On the other hand, the BART model manages to slightly lower the FKGL score to 14 and 16 for eLife and PLOS, respectively. However, in the case of the MedLane dataset, all methods appear to reach a similar level of complexity as the ground truth. This can be attributed to the dataset's shorter examples and potentially smaller vocabulary size, which limits the observed differences.

In the machine translation task, we fine-tuned the models across eight languages, as illustrated in Table 2(D). After fine-tuning, the BLEU scores significantly improved, with the most substantial improvement observed in the "en-fr" language pair, increasing by over 61%. This

enhancement can be attributed to the larger amount of training data available for "en-fr" (2,812,305 samples).

Table 2. Evaluation for the generative tasks. (A): Evaluation for the question-answering task: we compared ROUGE-L, BERTScore, MoverScore, BLEURT on zero-shot (Z.S) and RAG framework (RAG). The superior scores among the same models are highlighted in bold. (B): Evaluation for the single-document summarization task: we compared ROUGE-1, ROUGE-2, ROUGE-L, and some results are derived from other papers [53]. (C): Evaluation for the text simplification task: we compared ROUGE-1, ROUGE-2, ROUGE-L, and FKGL score. (D): Evaluation for the machine translation task: we evaluated BLEU score on eight language pairs. F.T refers to the results after fine-tuning.

Table 2(A)

	LLaMa2-13b				GPT-4			
	ROUGE-L	BERTScore	MoverScore	BLEURT	ROUGE -L	BERTScore	MoverScore	BLEURT
LiveQA								
Z.S	17.73	81.93	53.37	40.45	18.89	82.50	54.02	39.84
RAG	**18.83**	**82.79**	**53.79**	**40.59**	**19.44**	**83.01**	**54.11**	**40.55**
ExpertQA-Bio								
Z.S	23.26	84.38	55.58	44.65	23.00	84.50	56.15	44.53
RAG	**25.79**	**85.18**	**56.17**	**45.20**	**27.20**	**85.83**	**57.11**	**45.91**
ExpertQA-Med								
Z.S	24.86	84.89	55.74	46.32	25.45	85.11	56.50	45.98
RAG	**27.49**	**85.80**	**56.58**	**46.47**	**28.08**	**86.30**	**57.32**	**47.00**
MedicationQA								
Z.S	13.30	81.81	51.96	38.30	14.41	82.55	52.62	37.41
RAG	**14.71**	**82.79**	**52.59**	**38.42**	**16.19**	**83.59**	**53.30**	**37.91**

Table 2(B)

	PubMed			MIMIC-CXR			MEDIQA-AnS (p)			MEDIQA-AnS (s)		
	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Pegasus	45.97	20.15	28.25	22.49	11.57	20.35	18.29	4.82	13.87	22.21	8.23	16.76
BigBird	46.32	20.65	**42.33**	38.99	29.52	38.59	13.18	2.14	10.04	14.89	3.13	11.15
BART	**48.35**	**21.43**	36.90	**41.70**	**32.93**	**41.16**	**24.02**	7.20	**17.09**	38.19	22.20	30.58
SciFive	-	-	-	35.41	26.48	35.07	13.08	2.15	10.10	16.88	6.47	14.42
BioBART	-	-	-	41.61	32.90	41.00	22.58	**7.49**	16.69	**39.40**	**24.64**	**32.07**

Table 2(C)

	eLife				PLOS				MedLane			
	R-1	R-2	R-L	FKGL	R-1	R-2	R-L	FKGL	R-1	R-2	R-L	FKGL
Ground Truth	-	-	-	12	-	-	-	15	-	-	-	13
BigBirdPegasus	14.00	3.42	9.16	**13**	18.92	4.79	12.54	17	74.96	65.37	74.56	**13**
BART	**16.16**	**4.31**	**10.19**	14	21.09	7.20	14.17	**16**	**83.25**	**74.50**	**82.99**	**13**
BioBART	14.31	3.70	9.36	17	**23.80**	**7.83**	**15.65**	17	82.89	74.26	82.65	**13**

Table 2(D)

	BLEU Score							
	en-es	**en-fr**	**en-ro**	**en-cs**	**en-de**	**en-hu**	**en-pl**	**en-sv**
MarianMT	38.02	33.02	40.45	-	-	-	-	-
F.T-MarianMT	41.64	43.72	43.88	-	-	-	-	-
F.T-mT5	**45.88**	**53.29**	**47.28**	43.30	50.73	32.25	40.24	44.17

Physician Validation

Since the automated metrics cannot effectively assess the quality of generated content, especially in terms of factuality, we performed physician validation. 50 question-answer pairs from LiveQA were randomly selected, with answers generated by Baize-healthcare. Subsequently, two healthcare professionals (one resident and one attending specialist) rated these generated answers on the criteria of Readability, Relevancy, Accuracy, and Completeness, using a 5-point Likert scale, as shown in Figure 2(A). Detailed evaluation criteria can be found in Multimedia Appendix 5.

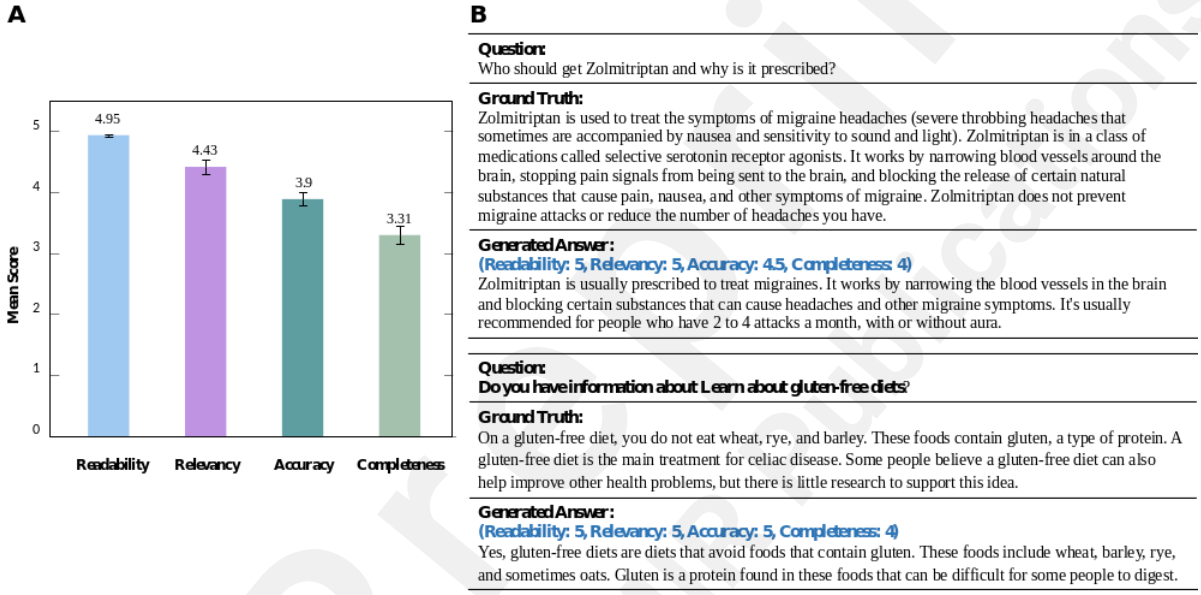


Figure 2. (A) Physician validation (Readability, Relevancy, Accuracy, Completeness) for 50 question-answer pairs. (B) Two examples of generated answers with ground truth.

The generated answers have good readability and relatively good relevancy, with scores of 4.95 and 4.43, respectively. In contrast, the completeness score is relatively lower (3.31). Figure 2(B) shows two cases. In the first case, compared to the ground truth, the generated answer does not point out that Zolmitriptan is used for treating acute migraines, nor does it indicate that it cannot be used to prevent migraine attacks or to reduce the frequency of headaches. And in the second case, the generated answer does not mention that a gluten-free diet is the main treatment for celiac disease. We provide two additional cases in Multimedia Appendix 6.

Additionally, we calculated the Inter-evaluator Agreement using percentage agreement for each criterion. Two healthcare professionals demonstrated a high level of consistency across all criteria, with the percentage agreement consistently exceeding 0.65.

Discussion

In-depth Analysis of the Text Summarization Task

In the multi-document summarization task, we included models based on traditional methods such as TextRank [54], as well as pre-trained language models such as BART, Pegasus, PRIMERA, and BioBART. We evaluated their performance using ROUGE scores on the MEDIQA-AnS dataset, which consists of 156 examples, and the results are shown in Table 3. However, it is noteworthy that although TextRank outperforms almost all generative models in ROUGE scores, this does not necessarily indicate superior performance. Since ROUGE scores are calculated based on the overlap between the generated content and reference summaries, and TextRank is an extractive summarization model, it tends to score higher by this measure.

Table 3. Evaluation for the multi-document summarization task.

	MEDIQA-AnS (p)			**MEDIQA-AnS (s)**		
	ROUGE-1	ROUGE -2	ROUGE -L	ROUGE -1	ROUGE -2	ROUGE -L
TextRank	29.88	10.23	17.01	43.77	26.80	30.52
BART	**24.56**	**7.56**	**17.18**	**32.32**	15.42	**24.03**
Pegasus	17.44	5.36	13.44	19.54	7.46	14.93
PRIMERA	16.66	4.89	12.68	21.78	9.77	16.85
BioBART	23.16	7.47	16.47	30.87	**15.91**	23.66

While generative models possess semantic comprehension abilities, enabling them to distill complex information into an easy-to-understand format. As shown in Table 4, the summarizations generated by BART display well-structured patient information, with a brief description of events and corresponding conditions of the current patient (highlighted in blue), exhibiting high readability. In contrast, the summarizations produced by TextRank are less readable and include noise (highlighted in orange); the generated content is often a literal collage of text fragments. Despite TextRank achieving higher ROUGE scores, it lacks the ability to discern information and integrate it into coherent and readable content, showing significant limitations for practical use.

Table 4. Two MIMIC-III (parts) examples of the text summarization task, generated by BART and TextRank, respectively. (We eliminated sensitive information).

BART	**TextRank**
The patient is an XXX-year-old man with a history of a question of coronary artery disease, borderline diabetes mellitus. **He was in his usual state of health** until 11 p.m. last night when he experienced chest pain with radiation to his back, positive shortness of breath, positive diaphoresis, no vomiting, no lightheadedness. The patient had had a similar episode of chest pain and was taken to an XXX. **He had successful angioplasty and stent of LAD and CX. He is a middle aged XXX man** in no acute hypertensive distress. He has had anginal chest pain, which is similar to his presenting complaint, but without radiations to hisBack. His blood pressure was 105/73, pulse 84, respiratory 21, O2 saturation 92% on 2	Admission Date: XXX Discharge Date: XXXDate of Birth: XXX Sex: MService: CCU-6HISTORY OF PRESENT ILLNESS: The patient is a XXX-year-old manwith a history of a question of coronary artery disease,status post myocardial infarction in [**December 2175**],hypertension, borderline diabetes mellitus who was in hisusual state of health until 11 p.m. last night when, while[**4-12**] midsternal pressure like chest pain with radiation toback, positive shortness of breath, positive diaphoresis,positive nausea, no vomiting, no lightheadedness. Mucous membranes moist.Oropharynx clear.NECK: No jugular venous distention, no carotid bruits.CARDIOVASCULAR: Regular rate,

liters. His CPK was 594, The index was 7.7, and he was admitted to the hospital with a high blood pressure. His condition was described as "stable" and "normal" by the doctor. The doctor referred the patient to a cardiologist for further treatment. The cardiologist said the patient was in good condition and should be discharged in a few days.	S1, S2, artificial S1 gallopand balloon pump, no murmurs or rubs.LUNGS: Bibasilar rales, left greater than right.ABDOMEN: Normoactive bowel sounds, nontender, nondistended.EXTREMITIES: No cyanosis, clubbing or edema.NEUROLOGIC: Alert and oriented x3.LABS AT OUTSIDE HOSPITAL: CPK was 304, troponin 1.75.Electrocardiogram at 1:23 a.m. was normal sinus rhythm at101, normal axis deviation, **2 to [**Street Address(2) 1755**]** elevation V1 to V5,Q V3, AVF.LABS AT [**Hospital6 **] AT 8 A.M.: CBC-white blood cells 11.2, hemoglobin 13.0, hematocrit 36.7,platelets 232. CARDIOVASCULAR: Coronary artery disease: Three vesseldisease with successful intervention on LAD and leftcircumflex, but RCA not done secondary to good collateral.The patient was continued on aspirin 325 qd.
Patient has CABG complicated by postop bleed and pleural effusion with discharge to [**Hospital1 **] Rehabilitation presents with abdominal pain. Zosyn was given in the ED. **Patient was otherwise doing well and was to go back to** rehab to finish his course of Cipro and Flagyl on [**5-17**]. Patient was last seen normal sometime last evening. **He woke up and noticed that the left side of his body felt "numb". He was not aware of any other** neurological weakness, and mostly complained of being very tired. He denied any new vision problems, did not have a headache. He sounded somewhat slurred but did not feel as if his speech was changed significantly. **He felt sleepy** but able to sustain attention, currently apparently in no distress. He was on standing. Plavix and [**State **] which had been held for the last few days (at least since the 14), since he had the percutaneous drainage. He did not. feel that the weakness had progress and reported that he felt the sense of numbness was starting to improve and had some difficulty squeezing an examiners hand. He is a retired postal worker. He lives with wife. and son who is a chiropractor.	Of note he was on standingPlavix and [**State **] which had been held for the last few days (atleast since the 14), since he had the percutaneous drainage.The patient was otherwise doing well and was to go back to rehabto finish his course of Cipro and Flagyl on [**5-17**].Past Medical History:coronary artery disease s/p right coronary artery stent x2([**10-3**], [**3-4**]), hypertension, hyperlipidemia, chronic obstructivepulmonary disease, asbestos exposure, chronic back pain,insomnia and obstructive sleep apnea (untreated)PSH: [**2144-4-21**]Endoscopic, minimally invasive, off pump coronary artery bypassgraft x1 with left internal mammary artery to left anteriordescending artery.[**2144-4-21**]Re-exploration for bleeding, post coronary artery bypassgrafting.Social History:Lives with wife.Exposure to asbestos.**Defers all medical decisions** to son who is a chiropractor.Occupation: retired postal worker.Tobacco: 3 PPD x 30 years, quit 45 years agoETOH: NoneFamily History:Non-contributory to cholecystitis.Physical Exam:Physical Exam:Vitals: T: 97.9 P:75 R: 16 BP:128/73 SaO2:96**General: Awake, felt sleepy but able to sustain attention, poorhistorian currently.**

System Usage

Ascle provides an easy-to-use approach for biomedical researchers and clinical staff. Users can efficiently utilize it by merely inputting text and calling the required functions. Figure 3 illustrates two use cases.

```
# create Ascle
from Ascle import Ascle
med = Ascle()

# Text Simplification
main_record = """
              The patient presents with symptoms of acute bronchitis,
              including cough, chest congestion, and mild fever.
              Auscultation reveals coarse breath sounds and occasional
              wheezing. Based on the clinical examination, a diagnosis
              of acute bronchitis is made, and the patient is prescribed
              a short course of bronchodilators and advised to rest and
              stay hydrated.
              """

# choose the model
layman_model = "ireneli1024/bart-large-elife-finetuned"

med.update_and_delete_main_record(content)

# call the text simplification function and print the output
print(med.get_layman_text(layman_model, min_length=20, max_length=70))

>> The patient presents with symptoms of acute bronchitis including
   cough, chest congestion and mild fever. Auscultation reveals coarse
   breath sounds and occasional wheezing. Based on these symptoms and
   the patient's history of previous infections with the same condition,
   the doctor decides that the patient is likely to have a cold or bronch.

# Machine Translation
main_record = """
              Myeloid derived suppressor cells (MDSC) are immature myeloid
              cells with immunosuppressive activity. They accumulate in
              tumor-bearing mice and humans with different types of cancer,
              including hepatocellular carcinoma (HCC).
              """

med.update_and_delete_main_record(record)

# call the machine translation function and print the output
print(med.get_translation_mt5("French"))

>> Les cellules suppressives dérivées de myéloïdes (MDSC) sont des
   cellules myéloïdes immatures ayant une activité immunosuppressive,
   accumulées chez des souris et des humains ayant différents types de
   cancer, y compris le carcinome hépatocellulaire (HCC).
```

Figure 3. Demonstration of system usage. We show two use cases: text simplification and machine translation.

Estimated Inference Time and Computational Resources

As shown in Table 5, we list the estimated inference time and computational resources required for the four generative tasks in Ascle. It is worth noting that the inference time is specific to our experimental settings, and the actual inference time for users may vary depending on the length of the input text and the computational resources used. For the question-answering task, GPT's response time is faster compared to LLaMA2-13b. However, it is important to mention that LLaMA2-13b was not deployed with quantization, and with quantization, the required inference time and computational resource requirements would be reduced.

Table 5. Estimated inference time and computational resources required for the generative tasks in Ascle.

Tasks	Estimated inference time (seconds/item)	Computational resource
Question Answering	LLaMA2-13b: < 60 s/item GPT4: < 15 s/item	LLaMA2-13b: 4 * NVIDIA A100 GPU GPT4: OpenAI API
Text Summarization	< 2 s/item	1 * NVIDIA V100 GPU
Text Simplification	< 2 s/item	1 * NVIDIA V100 GPU
Machine Translation	< 2 s/item	1 * NVIDIA V100 GPU

Clinicians' Utilization of Ascle

To evaluate the ease of usability of Ascle for clinicians, we report the time required for two clinicians with different backgrounds to use the package after receiving guidance. The backgrounds of the clinicians are as follows: (1) Physician 1: Singapore General Hospital, Senior Resident, 7 years of working experience. Physician 1 has a basic level of programming knowledge and is able to perform basic statistical analyses. (2) Physician 2: SengKang General Hospital, Senior Consultant, 15 years of working experience. Physician 2 has no programming knowledge.

Both clinicians received guidance on using Ascle, including setting up a virtual environment and accessing models from Hugging Face. The entire guidance process took about 10 minutes, after which both clinicians could independently and easily use Ascle and experiment with various generative functions without any issues. The main difficulty for the clinicians was setting up the virtual environment, as they lacked AI-specific knowledge. In response, Ascle provided a very simple virtual environment setup guideline. The clinicians' experience further confirms the user-friendliness of Ascle.

Limitations

In the case of generation tasks, we primarily chose automatic metrics for evaluation, such as ROUGE and BLEU scores. However, these metrics cannot effectively assess factual correctness [55] and may not align with human preference [56]. While human evaluation serves as an invaluable aspect in assessing the performance of the model, its incorporation may pose certain challenges due to various factors, including budget constraints.

Future Work

Recent LLMs have shown great potential in generative applications especially its superior zero- and few-shot performance [13,57,58]. Despite this, the generated content can be unfaithful, inconsistent, and biased [21,55,59,60]. We plan to thoroughly evaluate LLMs and extend to Ascle in the future. Meanwhile, we will strengthen the ethical review of these generative AI techniques to ensure their application truly and responsibly benefits biomedical researchers and healthcare professionals [61].

Conclusions

We introduce Ascle, a comprehensive NLP toolkit designed specifically for medical text generation. For the first time, it integrates four challenging generative functions, including question-answering, text summarization, text simplification and machine translation. Our

research fills the gap of existing toolkits for generative tasks, which holds significant implications for the entire medical domain. Ascle boasts remarkable flexibility, allowing users to access a variety of cutting-edge pre-trained language models. Meanwhile, it stands as a user-friendly toolkit, ensuring ease of use even for clinical staff without a technical background. We will continue to maintain and extend Ascle.

Acknowledgements

Tiarnan D.L. Keenan and Emily Y Chew were supported by the NIH Intramural Research Program (IRP), National Eye Institute. Zhiyong Lu, and Qingyu Chen were supported by the NIH IRP, National Library of Medicine. Qingyu Chen was also supported by the National Library of Medicine of the National Institutes of Health under award number 1K99LM014024.

Authors' Contribution

Rui Yang, Qingcheng Zeng, Keen You, Yujie Qiao, Lucas Huang, Chia-Chun Hsieh, Benjamin Rosand, Jeremy Goldwasser, Yu He Ke and Irene Li performed the data collection, data processing, and experiments. Amisha D Dave, Tiarnan D.L. Keenan conducted manual reviews. Rui Yang, Qingyu Chen, and Irene Li created the figures and tables and drafted the manuscript. Chuan Hong, Nan Liu, Emily Y Chew, Dragomir Radev, Zhiyong Lu, Hua Xu, Qingyu Chen, and Irene Li were responsible for project administration. All authors conceived of the idea for the article.

Conflicts of Interest

None declared.

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
FKGL: Flesch-Kincaid Grade Level
LLMs: Large Language Models
NLP: Natural Language Processing
RAG: Retrieval-Augmented Generation
UMLS: The Unified Medical Language System

References

1. Li I, Yasunaga M, Nuzumlalı MY, Caraballo C, Mahajan S, Krumholz H, Radev D. A neural topic-attention model for medical term abbreviation disambiguation. arXiv; 2019; doi: 10.48550/ARXIV.1910.14076

2. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, Rosand B, Li Y, Zhang M, Chang D, Taylor RA, Krumholz HM, Radev D. Neural Natural Language Processing for unstructured data in electronic health records: A review. Comput Sci Rev 2022 Nov;46(100511):100511. doi: 10.1016/j.cosrev.2022.100511

3. Shickel B, Tighe PJ, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE J Biomed Health Inform 2018 Sep;22(5):1589–1604. PMID:29989977

4. al-Aiad A, Duwairi R, Fraihat M. Survey: Deep learning concepts and techniques for electronic health record. 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA) IEEE; 2018. doi: 10.1109/aiccsa.2018.8612827

5. Zhang Y, Chen Q, Yang Z, Lin H, Lu Z. BioWordVec, improving biomedical word embeddings with subword information and MeSH. Sci Data 2019 May 10;6(1):52. PMID:31076572

6. Chen Q, Peng Y, Lu Z. BioSentVec: creating sentence embeddings for biomedical texts. 2019 IEEE International Conference on Healthcare Informatics (ICHI) IEEE; 2019. doi: 10.1109/ichi.2019.8904728

7. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. arXiv; 2018; doi: 10.48550/ARXIV.1810.04805

8. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020 Feb 15;36(4):1234–1240. PMID:31501885

9. Alsentzer E, Murphy JR, Boag W, Weng W-H, Jin D, Naumann T, McDermott MBA. Publicly available clinical BERT embeddings. arXiv; 2019; doi: 10.48550/ARXIV.1904.03323

10. Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J, Poon H. Domain-specific language model pretraining for biomedical natural language processing. ACM Trans Comput Healthc Association for Computing Machinery (ACM); 2022 Jan 31;3(1):1–23. doi: 10.1145/3458754

11. Singhal K, Tu T, Gottweis J, Sayres R, Wulczyn E, Hou L, Clark K, Pfohl S, Cole-Lewis H, Neal D, Schaekermann M, Wang A, Amin M, Lachgar S, Mansfield P, Prakash S, Green B, Dominowska E, Arcas BA y., Tomasev N, Liu Y, Wong R, Semturs C, Mahdavi SS, Barral J, Webster D, Corrado GS, Matias Y, Azizi S, Karthikesalingam A, Natarajan V. Towards expert-level medical question answering with large language models. arXiv; 2023; doi: 10.48550/ARXIV.2305.09617

12. Saab K, Tu T, Weng W-H, Tanno R, Stutz D, Wulczyn E, Zhang F, Strother T, Park C, Vedadi E, Chaves JZ, Hu S-Y, Schaekermann M, Kamath A, Cheng Y, Barrett DGT, Cheung C, Mustafa B, Palepu A, McDuff D, Hou L, Golany T, Liu L, Alayrac J-B, Houlsby N, Tomasev N, Freyberg J, Lau C, Kemp J, Lai J, Azizi S, Kanada K, Man S, Kulkarni K, Sun R, Shakeri S, He L, Caine B, Webson A, Latysheva N, Johnson M, Mansfield P, Lu J, Rivlin E, Anderson J, Green B, Wong R, Krause J, Shlens J, Dominowska E, Eslami SMA, Chou K, Cui C, Vinyals O, Kavukcuoglu K, Manyika J, Dean J, Hassabis D, Matias Y, Webster D, Barral J, Corrado G, Semturs C, Mahdavi SS, Gottweis J, Karthikesalingam A, Natarajan V. Capabilities of Gemini models in medicine. arXiv; 2024; doi: 10.48550/ARXIV.2404.18416

13. Yang R, Tan TF, Lu W, Thirunavukarasu AJ, Ting DSW, Liu N. Large language models in health care: Development, applications, and challenges. Health Care Sci 2023 Aug;2(4):255–263. PMID:38939520

14. Li C, Mowery DL, Ma X, Yang R, Vurgun U, Hwang S, Donnelly HK, Bandhey H, Akhtar Z, Senathirajah Y, Sadhu EM, Getzen E, Freda PJ, Long Q, Becich MJ. Realizing the Potential of Social Determinants Data: A Scoping Review of Approaches for Screening, Linkage, Extraction, Analysis and Interventions. medRxiv 2024 Feb 6; PMID:38370703

15. Yang R, Ning Y, Keppo E, Liu M, Hong C, Bitterman DS, Ong JCL, Ting DSW, Liu N. Retrieval-augmented generation for generative artificial intelligence in medicine. arXiv; 2024; doi: 10.48550/ARXIV.2406.12449

16. Wang S, McDermott MBA, Chauhan G, Ghassemi M, Hughes MC, Naumann T. MIMIC-Extract. Proceedings of the ACM Conference on Health, Inference, and Learning New York, NY, USA: ACM; 2020. doi: 10.1145/3368555.3384469

17. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and robust models for biomedical natural

language processing. Proceedings of the 18th BioNLP Workshop and Shared Task Stroudsburg, PA, USA: Association for Computational Linguistics; 2019. doi: 10.18653/v1/w19-5034

18. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, Box TL, DuVall SL, Patterson OV. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. AMIA Annu Symp Proc 2021;2021:438–447. PMID:35308962

19. Yang F, Wang X, Ma H, Li J. Transformers-sklearn: a toolkit for medical language understanding with transformer-based models. BMC Med Inform Decis Mak 2021 Jul 30;21(Suppl 2):90. PMID:34330244

20. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. Biomedical and clinical English model packages for the Stanza Python NLP library. J Am Med Inform Assoc 2021 Aug 13;28(9):1892–1899. PMID:34157094

21. Yang R, Liu H, Marrese-Taylor E, Zeng Q, Ke YH, Li W, Cheng L, Chen Q, Caverlee J, Matsuo Y, Li I. KG-Rank: Enhancing large Language Models for medical QA with knowledge graphs and ranking techniques. arXiv; 2024; doi: 10.48550/ARXIV.2403.05881

22. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 1;32(Database issue):D267–70. PMID:14681409

23. Xu C, Guo D, Duan N, McAuley J. Baize: An open-source chat model with parameter-efficient tuning on Self-chat data. arXiv; 2023; doi: 10.48550/ARXIV.2304.01196

24. OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser Ł, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner JH, Kiros J, Knight M, Kokotajlo D, Kondraciuk Ł, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres F de AB, Petrov M, Pinto HP de O, Michael, Pokorny, Pokrass M, Pong VH, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, Rae J, Ramesh A, Raymond C, Real F, Rimbach K, Ross C, Rotsted B, Roussez H, Ryder N, Saltarelli M, Sanders T, Santurkar S, Sastry G, Schmidt H, Schnurr D, Schulman J, Selsam D, Sheppard K, Sherbakov T, Shieh J, Shoker S, Shyam P, Sidor S, Sigler E, Simens M, Sitkin J, Slama K, Sohl I, Sokolowsky B, Song Y, Staudacher N, Such FP, Summers N, Sutskever I, Tang J, Tezak N, Thompson MB, Tillet P, Tootoonchian A, Tseng E, Tuggle P, Turley N, Tworek J, Uribe JFC, Vallone A, Vijayvergiya A, Voss C, Wainwright C, Wang JJ, Wang A, Wang B, Ward J, Wei J, Weinmann CJ, Welihinda A, Welinder P, Weng J, Weng L, Wiethoff M, Willner D, Winter C, Wolrich S, Wong H, Workman L, Wu S, Wu J, Wu M, Xiao K, Xu T, Yoo S, Yu K, Yuan Q, Zaremba W, Zellers R, Zhang C, Zhang M, Zhao S, Zheng T, Zhuang J, Zhuk W, Zoph B. GPT-4 Technical Report. arXiv; 2023; doi:

10.48550/ARXIV.2303.08774

25. Touvron H, Martin L, Stone K, Albert P, Almahairi A, Babaei Y, Bashlykov N, Batra S, Bhargava P, Bhosale S, Bikel D, Blecher L, Ferrer CC, Chen M, Cucurull G, Esiobu D, Fernandes J, Fu J, Fu W, Fuller B, Gao C, Goswami V, Goyal N, Hartshorn A, Hosseini S, Hou R, Inan H, Kardas M, Kerkez V, Khabsa M, Kloumann I, Korenev A, Koura PS, Lachaux M-A, Lavril T, Lee J, Liskovich D, Lu Y, Mao Y, Martinet X, Mihaylov T, Mishra P, Molybog I, Nie Y, Poulton A, Reizenstein J, Rungta R, Saladi K, Schelten A, Silva R, Smith EM, Subramanian R, Tan XE, Tang B, Taylor R, Williams A, Kuan JX, Xu P, Yan Z, Zarov I, Zhang Y, Fan A, Kambadur M, Narang S, Rodriguez A, Stojnic R, Edunov S, Scialom T. Llama 2: Open foundation and fine-tuned chat models. arXiv; 2023; doi: 10.48550/ARXIV.2307.09288

26. Abacha AB, Agichtein E, Pinter Y, Demner-Fushman D, editors. Overview of the Medical Question Answering Task at TREC 2017 LiveQA. Text Retrieval Conference; 2017.

27. Malaviya C, Lee S, Chen S, Sieber E, Yatskar M, Roth D. ExpertQA: Expert-curated questions and attributed answers. arXiv; 2023; doi: 10.48550/ARXIV.2309.07852

28. Abacha AB, Mrabet Y, Sharp M, Goodwin TR, Shooshan SE, Demner-Fushman D. Bridging the Gap Between Consumers' Medication Questions and Trusted Answers. Stud Health Technol Inform 2019 Aug 21;264:25–29. PMID:31437878

29. Mishra R, Bian J, Fiszman M, Weir CR, Jonnalagadda S, Mostafa J, Del Fiol G. Text summarization in the biomedical domain: a systematic review of recent research. J Biomed Inform 2014 Dec;52:457–467. PMID:25016293

30. Ke Y, Yang R, Liu N. Comparing Open-Access Database and Traditional Intensive Care Studies Using Machine Learning: Bibliometric Analysis Study. J Med Internet Res 2024 Apr 17;26:e48330. PMID:38630522

31. Xie Q, Luo Z, Wang B, Ananiadou S. A survey for biomedical text summarization: From pre-trained to large language models. arXiv; 2023; doi: 10.48550/ARXIV.2304.08763

32. Zhang J, Zhao Y, Saleh M, Liu PJ. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. arXiv; 2019; doi: 10.48550/ARXIV.1912.08777

33. Zaheer M, Guruganesh G, Dubey A, Ainslie J, Alberti C, Ontanon S, Pham P, Ravula A, Wang Q, Yang L, Ahmed A. Big bird: Transformers for longer sequences. arXiv; 2020; doi: 10.48550/ARXIV.2007.14062

34. Lewis M, Liu Y, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. arXiv; 2019; doi: 10.48550/ARXIV.1910.13461

35. Xiao W, Beltagy I, Carenini G, Cohan A. PRIMERA: Pyramid-based Masked Sentence Pre-training for Multi-document Summarization. arXiv; 2021; doi: 10.48550/ARXIV.2110.08499

36. Phan LN, Anibal JT, Tran H, Chanana S, Bahadroglu E, Peltekian A, Altan-Bonnet G. SciFive: a text-to-text transformer model for biomedical literature. arXiv; 2021; doi: 10.48550/ARXIV.2106.03598

37. Yuan H, Yuan Z, Gan R, Zhang J, Xie Y, Yu S. BioBART: Pretraining and evaluation of A biomedical generative language model. arXiv; 2022; doi: 10.48550/ARXIV.2204.03905

38. Cohan A, Dernoncourt F, Kim DS, Bui T, Kim S, Chang W, Goharian N. A discourse-aware attention model for abstractive summarization of long documents. arXiv; 2018; doi: 10.48550/ARXIV.1804.05685

39. Johnson AEW, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-Y, Mark RG, Horng S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data 2019 Dec 12;6(1):317. PMID:31831740

40. Savery M, Abacha AB, Gayen S, Demner-Fushman D. Question-driven summarization of answers to consumer health questions. Sci Data 2020 Oct 2;7(1):322. PMID:33009402

41. Devaraj A, Wallace BC, Marshall IJ, Li JJ. Paragraph-level Simplification of Medical Texts. Proc Conf 2021 Jun;2021:4972–4984. PMID:35663507

42. Goldsack T, Zhang Z, Lin C, Scarton C. Making science simple: Corpora for the lay summarisation of scientific literature. arXiv; 2022; doi: 10.48550/ARXIV.2210.09932

43. Luo J, Zheng Z, Ye H, Ye M, Wang Y, You Q, Xiao C, Ma F. Benchmarking automated clinical language simplification: Dataset, algorithm, and evaluation. arXiv; 2020; doi: 10.48550/ARXIV.2012.02420

44. Khoong EC, Rodriguez JA. A Research Agenda for Using Machine Translation in Clinical Medicine. J Gen Intern Med 2022 Apr;37(5):1275–1277. PMID:35132559

45. Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, Seide F, Germann U, Aji AF, Bogoychev N, Martins AFT, Birch A. Marian: Fast Neural Machine Translation in C++. arXiv; 2018; doi: 10.48550/ARXIV.1804.00344

46. Xue L, Constant N, Roberts A, Kale M, Al-Rfou R, Siddhant A, Barua A, Raffel C. mT5: A massively multilingual pre-trained text-to-text transformer. arXiv; 2020; doi: 10.48550/ARXIV.2010.11934

47. UFAL Medical Corpus. Available from: https://ufal.mff.cuni.cz/ufal_medical_corpus

48. Lin C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. Text Summarization Branches Out 2004. p. 74–81. Available from: https://aclanthology.org/W04-1013.pdf

49. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: Evaluating Text Generation with BERT. arXiv; 2019; doi: 10.48550/ARXIV.1904.09675

50. Zhao W, Peyrard M, Liu F, Gao Y, Meyer CM, Eger S. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv; 2019; doi: 10.48550/ARXIV.1909.02622

51. Sellam T, Das D, Parikh AP. BLEURT: Learning robust metrics for text generation. arXiv; 2020; doi: 10.48550/ARXIV.2004.04696

52. Peter Kincaid J, Fishburne RP Jr, Rogers RL, Chissom BS. Derivation Of New Readability Formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) For Navy Enlisted Personnel. Institute for Simulation and Training, University of Central Florida; 1975; Available from: https://stars.library.ucf.edu/istlibrary/56

53. Rohde T, Wu X, Liu Y. Hierarchical learning for generation with long source sequences. arXiv; 2021; doi: 10.48550/ARXIV.2104.07545

54. Mihalcea R, Tarau P. TextRank: Bringing Order into Text. Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing 2004. p. 404–411. Available from: https://aclanthology.org/W04-3252.pdf

55. Xie Q, Schenck EJ, Yang HS, Chen Y, Peng Y, Wang F. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. medRxiv 2023 Jul 1; PMID:37398329

56. Fleming SL, Lozano A, Haberkorn WJ, Jindal JA, Reis EP, Thapa R, Blankemeier L, Genkins JZ, Steinberg E, Nayak A, Patel BS, Chiang C-C, Callahan A, Huo Z, Gatidis S, Adams SJ, Fayanju O, Shah SJ, Savage T, Goh E, Chaudhari AS, Aghaeepour N, Sharp C, Pfeffer MA, Liang P, Chen JH, Morse KE, Brunskill EP, Fries JA, Shah NH. MedAlign: A clinician-generated dataset for instruction following with electronic medical records. arXiv; 2023; doi: 10.48550/ARXIV.2308.14089

57. Gao F, Jiang H, Yang R, Zeng Q, Lu J, Blum M, Liu D, She T, Jiang Y, Li I. Large Language Models on Wikipedia-style survey generation: An evaluation in NLP concepts. arXiv; 2023; doi: 10.48550/ARXIV.2308.10410

58. Yang R, Yang B, Ouyang S, She T, Feng A, Jiang Y, Lecue F, Lu J, Li I. Leveraging Large Language Models for concept graph recovery and question answering in NLP education. arXiv; 2024; doi: 10.48550/ARXIV.2402.14293

59. Ke YH, Yang R, Lie SA, Lim TXY, Abdullah HR, Ting DSW, Liu N. Enhancing diagnostic accuracy through multi-agent conversations: Using large language models to mitigate cognitive bias. arXiv; 2024; doi: 10.48550/ARXIV.2401.14589

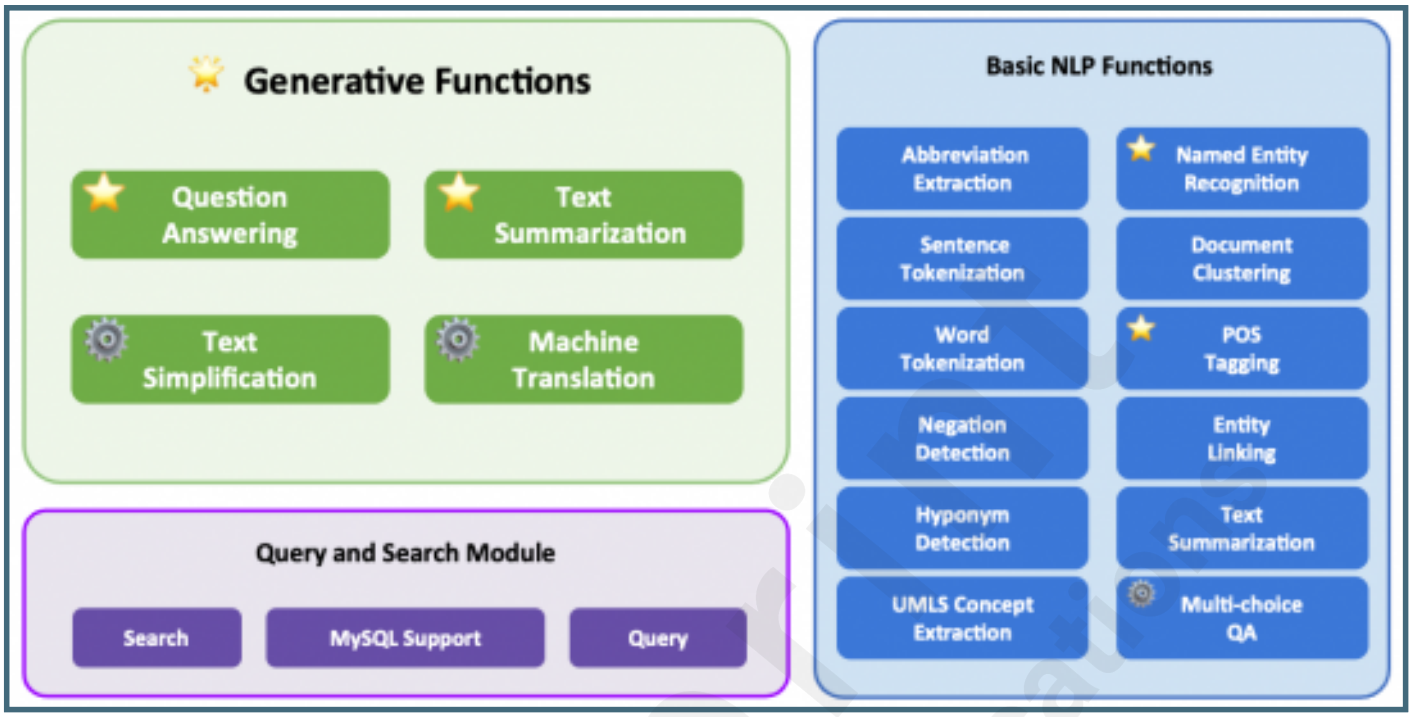
60. Tian S, Jin Q, Yeganova L, Lai P-T, Zhu Q, Chen X, Yang Y, Chen Q, Kim W, Comeau DC, Islamaj R, Kapoor A, Gao X, Lu Z. Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Brief Bioinform 2023 Nov 22;25(1). PMID:38168838

61. Ning Y, Teixayavong S, Shang Y, Savulescu J, Nagaraj V, Miao D, Mertens M, Ting DSW, Ong JCL, Liu M, Cao J, Dunn M, Vaughan R, Ong MEH, Sung JJ-Y, Topol EJ, Liu N. Generative artificial intelligence in healthcare: Ethical considerations and assessment checklist. arXiv; 2023; doi: 10.48550/ARXIV.2311.02107

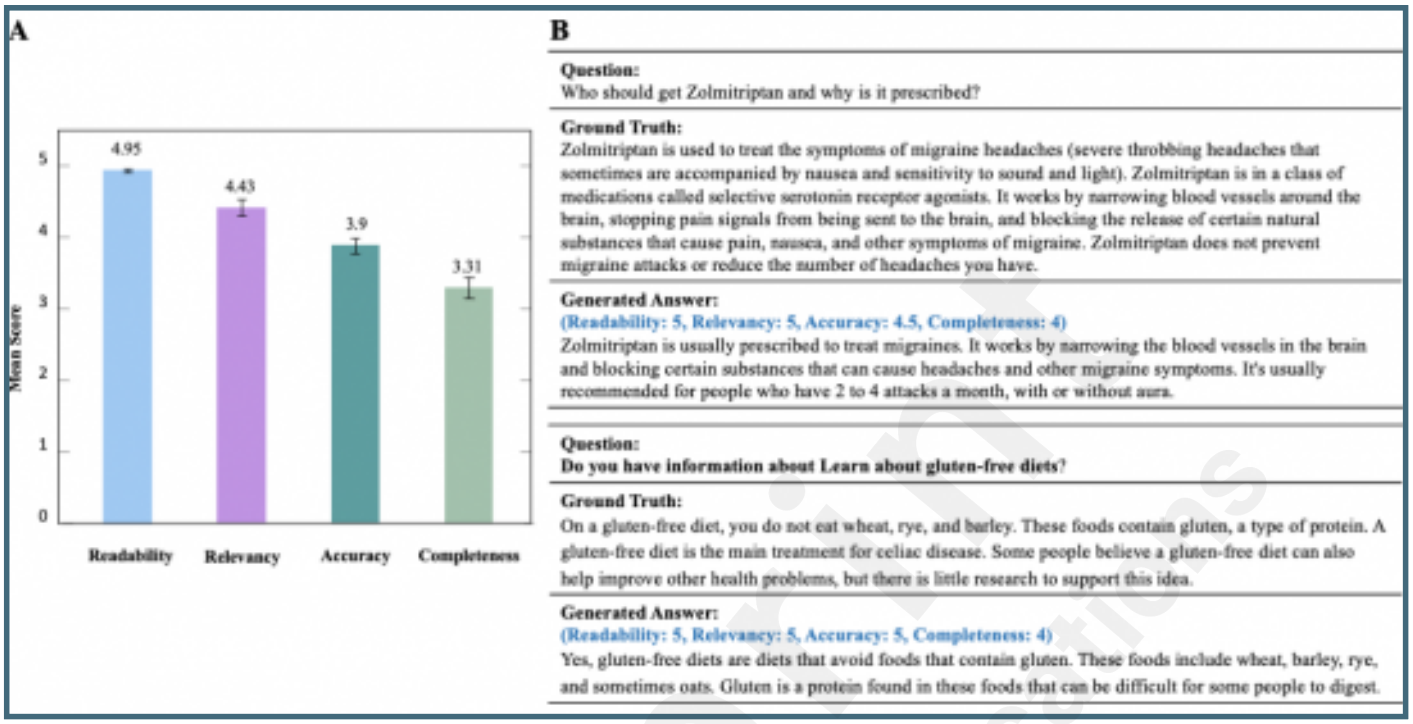
Supplementary Files

Figures

The overall architecture of Ascle. ??indicates that we have our fine-tuned language models for this task. ??indicates that we conducted evaluations for this task.



(A) Physician validation (Readability, Relevancy, Accuracy, Completeness) for 50 question-answer pairs. (B) Two examples of generated answers with ground truth.



Demonstration of system usage. We show two use cases: text simplification and machine translation.

```
# create Ascle
from Ascle import Ascle
med = Ascle()

# Text Simplification
main_record = """
              The patient presents with symptoms of acute bronchitis,
              including cough, chest congestion, and mild fever.
              Auscultation reveals coarse breath sounds and occasional
              wheezing. Based on the clinical examination, a diagnosis
              of acute bronchitis is made, and the patient is prescribed
              a short course of bronchodilators and advised to rest and
              stay hydrated.
              """

# choose the model
layman_model = "ireneli1024/bart-large-elife-finetuned"

med.update_and_delete_main_record(content)

# call the text simplification function and print the output
print(med.get_layman_text(layman_model, min_length=20, max_length=70))

>> The patient presents with symptoms of acute bronchitis including
   cough, chest congestion and mild fever. Auscultation reveals coarse
   breath sounds and occasional wheezing. Based on these symptoms and
   the patient's history of previous infections with the same condition,
   the doctor decides that the patient is likely to have a cold or bronch.

# Machine Translation
main_record = """
              Myeloid derived suppressor cells (MDSC) are immature myeloid
              cells with immunosuppressive activity. They accumulate in
              tumor-bearing mice and humans with different types of cancer,
              including hepatocellular carcinoma (HCC).
              """

med.update_and_delete_main_record(record)

# call the machine translation function and print the output
print(med.get_translation_mt5("French"))

>> Les cellules suppressives dérivées de myéloïdes (MDSC) sont des
   cellules myéloïdes immatures ayant une activité immunosuppressive,
   accumulées chez des souris et des humains ayant différents types de
   cancer, y compris le carcinome hépatocellulaire (HCC).
```

Multimedia Appendixes

32 Fine-Tuned Language Models and 27 Benchmarks in Ascle.
URL: http://asset.jmir.pub/assets/ef7cdcd0ef56db1d91e8e406a06a8cca.docx

Basic NLP Functions in Ascle.
URL: http://asset.jmir.pub/assets/40f9cfc361600deb90a5dba307e7613a.docx

Query and Search Module in Ascle.
URL: http://asset.jmir.pub/assets/70edfe43adda388a1c341a68136542f2.docx

The RAG Framework in Ascle – KG-Rank.
URL: http://asset.jmir.pub/assets/ba856e5cc28f2369b94274ed44ac403a.docx

Evaluation Criteria for Physician Validation.
URL: http://asset.jmir.pub/assets/222878f2094e9d66abfe10617e364a71.docx

Case Study.
URL: http://asset.jmir.pub/assets/a9ba8907f9866e00367bd707e71cbc31.docx