Applied Medical Informatics
Research Letter Vol. 48, No. 1/2026, pp: 9-15

Enhancing the Accuracy of Large Language Models in Medical Coding through Retrieval-Based Approaches

Keith KWAN [1, 2], Hao CHEN [3], Billy Ho Hung CHEUNG [*,2]

[1] AI Native Health, Unit D, 1/F, Sunshine Plaza 17 Sung On Street, Hung Hom Kowloon, Hong Kong, China
[2] LKS Faculty of Medicine, the University of Hong Kong, 6/F, William MW Mong Block, 21 Sassoon Road, Pokfulam, Hong Kong SAR, China
[3] Department of Computer Science and Engineering, Faculty of Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
E-mail: thisiskeith@gmail.com; jhc@ust.hk; billychh@hku.hk

* Author to whom correspondence should be addressed; Tel.: +852 2255 4773; Fax: +852 2817 2291

Received: 28 March 2025/Accepted: 17 March 2026 / Published online: 31 March 2026

Abstract

Purpose: Medical coding, essential for healthcare administration and research, requires significant expertise and resources. While Large Language Models (LLMs) showed promise in automating this task, recent studies highlighted their limitations, with even advanced models like GPT-4 achieving only moderate accuracy. This study presents a novel Retrieve-Rank system combining ColBERT-V2 retriever with GPT-3.5-turbo for medical coding automation. *Methods:* We evaluated the performance of our Retrieve-Rank system against a Vanilla LLM approach using a dataset of 100 single-term medical conditions with corresponding International Classification of Diseases, 10th edition, Clinical Modification (ICD-10-CM) codes, which is the latest version of the standardized system used to code diseases and medical conditions used in the United States. The system employed a two-step process: first, retrieving the top-15 most relevant codes using ColBERT-V2, then applying GPT-3.5-turbo for reranking to select the most appropriate code. The experiment was conducted on 1st June 2024. Performance was measured using top-one accuracy with normalized ICD-10-CM codes. *Results*: Our Retrieve-Rank system achieved 100% accuracy in code identification, significantly outperforming the Vanilla LLM approach's 6% accuracy. This improvement is particularly noteworthy as it was achieved using GPT-3.5, a more accessible model than GPT-4, demonstrating that LLMs, when equipped with appropriate retrieval mechanisms, can effectively overcome their inherent limitations in medical coding tasks. *Conclusions*: While our study was limited to single-term conditions, the results suggest significant potential for broader applications in healthcare administration. This research contributes to bridging the gap between AI capabilities and clinical implementation, offering a promising approach to automating medical coding while maintaining high accuracy. Future research should focus on validating these findings with more complex, real-world medical cases and unstructured clinical notes.

Keywords: ICD-10-CM coding; Medical informatics; Natural Language Processing; Retrieve-Rank system; Automated diagnosis coding; Machine learning in healthcare

Introduction

Medical coding, the process of translating medical diagnoses, procedures, and services into standardized codes, is crucial for healthcare administration, billing, epidemiological studies, and quality assessment [1,2]. The International Classification of Diseases (ICD), published by the World Health Organization (WHO), provides the global standard for coding medical diagnoses [3]. In the United States, the ICD-10 clinical modification of this

system (ICD-10-CM) is the official coding system for patient records, billing, public health tracking, and research [4]. However, the work involves the coder to be competent in understanding and match the codes, and the workload can be huge given the amount of codes that can be involved. Recently, the application of large language models (LLMs), a type of artificial intelligence (AI) focusing on text-based task, in medical coding has gained attention, with the aim of automating and streamlining this complex task [5,6]. One of the most prominent models that gained a lot of attention from the research community as well as the general public is the GPT series developed by OpenAI [7].

However, a recent study by Soroush et al. [8] highlighted significant limitations in the ability of LLMs to accurately generate medical codes. Their evaluation of several prominent LLMs on a comprehensive dataset of ICD-9-CM, ICD-10-CM, and CPT codes from the Mount Sinai Health System electronic health record revealed that even the best-performing model, GPT-4, achieved exact-match rates of only 45.9%, 33.9%, and 49.8% across the respective code systems. These findings led the authors to conclude that LLMs are currently not suitable for direct use in medical coding tasks.

However, the potential utility of LLM in medicine is believed to be substantial [9], and an additional tool may help provide a significant boost to its performance [10]. Hence, we hypothesized that equipping LLMs with appropriate tools, including some retrieval mechanisms, could significantly improve their performance in medical coding. This approach aligns with recent advancements in Retrieval-Augmented Generation (RAG) [11] and the use of external knowledge bases to enhance LLM performance [12].

To test this hypothesis, we designed an experiment using a combination of the ColBERT-V2 retriever 13 and GPT-3.5-turbo for reranking, drawing inspiration from successful applications of similar techniques in other domains [14]. Our study aimed to evaluate the performance of this novel Retrieve-Rank system against a Vanilla LLM approach, using a dataset of a single-term medical conditions and their corresponding ICD-10-CM codes [15].

Materials and Methods

2.1 Dataset

A dataset of 100 single-term medical conditions and their corresponding ICD-10-CM codes was used in this study (Supplementary File 1). The dataset was generated via a simple random sampling [16] in Excel from the full list of ICD-10-CM codes (https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2022/icd10cm_codes_2022.txt). Simple random sampling was used to minimize the risk of bias while ensuring a wide spectrum of conditions can be represented, where all conditions are assumed to be of equal importance [16].

2.2 *Retrieve-Rank System*

The system comprises two main components, with a retrieve-rank architecture for automated ICD-10-CM coding within our Retrieval-Augmented Generation (RAG) pipeline (Figure 1).

The material used to conduct and analyze the experiments is publicly available on GitHub at https://github.com/ainativehealth/GoodMedicalCoder. The system first performed semantic retrieval using a pre-trained ColBERT (Contextualized Late Interaction over BERT) model [17] as a retriever. This model conducts a deep, context-aware similarity search against a pre-indexed knowledge base of ICD-10 codes and their clinical descriptions. Given a clinical query, this retrieval stage efficiently narrows the vast code space—comprising over 70,000 potential labels [15]—to a focused, semantically relevant candidate set. The focused candidate set was designed to comprise 15 codes (k=15), based on our prior experiment, to improve the performance of LLMs.

Subsequently, the architecture transitions to a context-aware ranking phase. The initial retrieval results, comprising codes paired with their descriptive passages, are formatted as context and presented to a specialized reranking systemic, where GPT-3.5-turbo [18] was employed to re-rank the retrieved codes and select the most likely ICD-10-CM code for the given condition.

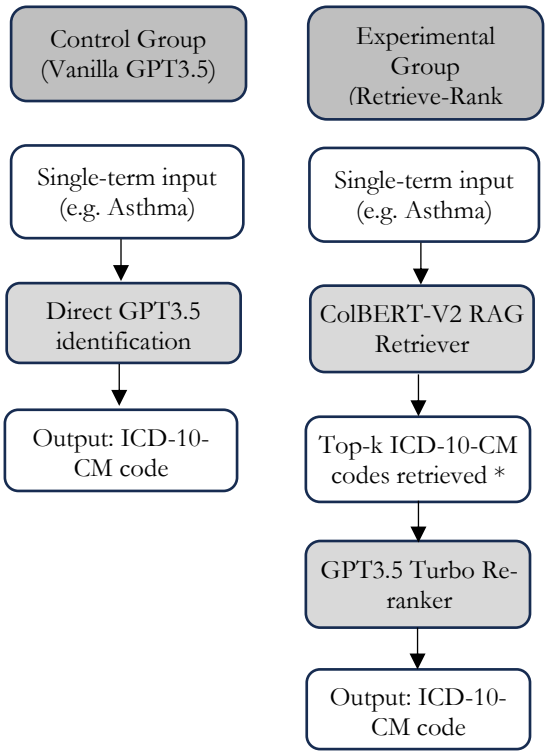

Figure 1. Comparison of Control Group and Experiment Group methodologies and results. *k = 15

2.3 Vanilla (Original) Large Language Model

As a control group, we implemented an original (as known as "vanilla") LLM using GPT-3.5-turbo. The model was prompted (instructed) with "You are a medical coding expert who can suggest an ICD-10-CM code for a given query," followed by the single-term medical condition.

2.4 *Evaluation Metrics*

The experiment consists of two arms – code identification by our Rank-Retrieval system and by the Vanilla LLM. The experiment was performed on June 1, 2024. Our analysis focused on top-one accuracy, comparing the code retrieved by each approach with the reference code. A match was considered successful if the main part of the identified code (before any subdivisions, Figure 2) matched the reference code.

Figure 2. Example of an ICD-10-CM code: puncture wound without foreign body of the right little finger without damage to nail, sequela S61.236S

To ensure a consistent format for fair comparison, ICD-10-CM codes were standardized by removing periods (e.g., from S61.236S to S61236S) and converting to uppercase (e.g., from s61236s to S61236S).

Results

3.1 *Accuracy Comparison*

The Retrieve-Rank system achieved 100% accuracy in identifying ICD-10-CM codes for the 100 sampled medical conditions. In contrast, the Vanilla LLM using GPT-3.5-turbo achieved only 6% accuracy.

3.2 *Performance* Analysis

The Retrieve-Rank system consistently identified the correct ICD-10-CM code across various medical specialties, demonstrating high precision in capturing specific anatomical locations, encounter types, and complex conditions (Table 1). The Vanilla LLM's approach's errors often involved identifying codes in the same general category but missing crucial details.

Table 1. Comparison of examples in various domains.

Diagnosis Description	Reference Code	Retrieve-Rank	Vanilla **GPT-3.5-turbo**	Correct System
Salter-Harris Type II physeal fracture of lower end of humerus, unspecified arm, subsequent encounter for fracture with malunion	S49129P	S49I29P	S59102P	Retrieve-Rank
Nondisplaced fracture of proximal third of navicular [scaphoid] bone of unspecified wrist, initial encounter for closed fracture	S62036A	S62036A	S62002A	Retrieve-Rank
Glaucoma secondary to eye inflammation, right eye, indeterminate stage	H404IX4	H404IX4	H4060X4	Retrieve-Rank
Poisoning by aspirin, accidental (unintentional), initial encounter	T39011A	T390l1A	T390l1A	Both
Other specified injury of right renal vein, subsequent encounter	S35494D	S35494D	S35602D	Retrieve-Rank
Other specified fracture of right acetabulum, initial encounter for open fracture	S32491B	S32491B	S32431B	Retrieve-Rank
Displacement of biological heart valve graft, sequela	T82222S	T82222S	T82590S	Retrieve-Rank
Open bite of unspecified thumb with damage to nail, sequela	S6ll59S	S6ll59S	S6I049S	Retrieve-Rank
Burn of unspecified degree of trunk, unspecified site, sequela	T2100XS	T2I00XS	T3l0	Retrieve-Rank
Other specified injury of peroneal artery, unspecified leg, subsequent encounter	S85299D	S85299D	S95IXXA	Retrieve-Rank
Other complications of anesthesia, subsequent encounter	T8859XD	T8859XD	T8859XD	Both
Follicular lymphoma, unspecified, lymph nodes of axilla and upper limb	C8294	C8294	C82l 1	Retrieve-Rank
Other injury of flexor muscle, fascia and tendon of other finger at wrist and hand level, sequela	S66198S	S66I98S	S66299S	Retrieve-Rank
Contusion and laceration of cerebrum, unspecified, with loss of consciousness greater than 24 hours with return to pre-existing conscious level, initial encounter	S06335A	S06335A	S069X0A	Retrieve-Rank

Discussion

With a combination of the ColBERT-V2 retriever [13] and GPT-3.5-turbo for reranking, there is a significant performance boost of accuracy (100% correct from our tool vs. 6% accuracy from Vanilla GPT-3.5-turbo, Table 1). The retrieve-rank system identified the correct ICD-10-CM code with different specialties, while Vanilla LLM usually only identified codes in the same broader category, but missed the actual code.

In the landmark study by Soroush et al. [8] found that the high-performance Vanilla LLM (GPT-4) achieved exact match rates of only at 49.8% for ICD-10-CM codes, which is significantly lower than the current study using retrieval-based approach, which demonstrated a 100% accurate match. It is encouraging that despite our use of the less capable GPT-3.5-turbo model, compared to a Vanilla GPT-4, our system could perform better when a retrieval step is introduced. As GPT-3.5-turbo is more available and cost-effective, by incorporating a retrieval step using ColBERT-V2 and a reranking step using GPT-3.5-turbo, we have significantly improved the accuracy of medical coding, demonstrating that LLMs, even with a less superior model, when equipped with appropriate tools, can indeed be effective in this task. A higher accuracy reduces required manpower for cross-checking and decreases

error risk. With 100% accuracy demonstrated, there is potential that with more rigorous testing, a fully automatic system can be built while minimizing the risk of hallucinations and false information, which has been a documented concern with LLMs [19].

Another study focused on ICD codes identification focusing on retina diseases found that using Vanilla ChatGPT yielded a 59% accuracy rate [20]. It is higher than our 6% accuracy with Vanilla GPT-3.5-Turbo but still significantly lower than our 100% using the RAG pipeline. Moreover, their higher accuracy could be partially contributed by the incorporation of long and detailed clinical notes, narrow spectrum of retinal diseases, and a different version of ChatGPT (research demo in their study vs 3.-turbo in our study) [20].

Like our study, using RAG pipeline to dynamically retrieve candidate codes had been utilized in other studies. A 2025 study demonstrated an LLM-RAG pipeline for generating preoperative assessments based on 58 distinct surgical guidelines across 14 clinical scenarios [21]. The GPT-4-based RAG system achieved a 96.4% accuracy in determining surgical fitness and generating instructions, significantly outperforming human experts (86.6%) on the same task. Crucially, this system demonstrated an absence of hallucinations, which is different to prevent despite advances in LLM if they are used alone (for example, the latest model of Gemini-2.5-flash-lite by Google still exhibit a grounded hallucination rate of 3.3% [22]). The retrieval component dynamically fetched the relevant institutional protocols, ensuring each response was current and compliant.

The limitations of LLMs in medical coding tasks when using simple prompts can be attributed to several fundamental factors. First of all, medical coding requires the ability to filter and prioritize contextually relevant information while discarding irrelevant details [23]. Large language models, operating on statistical correlations across their training data, lack this rationality and may incorporate spurious patterns or irrelevant features into their decisions [23]. Secondly, unlike human coders who develop expertise through cue utilization and pattern recognition within a specific clinical context [24], LLMs operate in a disorganized manner where decisions are dataset-driven rather than clinically-driven. This often results in the model missing crucial diagnostic nuances or failing to maintain coding consistency across similar cases. LLMs lack this meta-cognitive capability, potentially leading to overconfident but incorrect code assignments. Finally, the lack of explicit retrieval and reasoning mechanisms means LLMs cannot systematically apply coding rules or verify their decisions against authoritative sources, resulting in inconsistent adherence to coding standards.

Our Retrieve-Rank approach addresses these limitations by leveraging the strengths of both retrieval and language models. The ColBERT-V2 retriever allows for efficient access to relevant ICD-10-CM codes based on the input medical condition, narrowing down the search space and providing the language model with a focused set of candidates [11]. The GPT-3.5-turbo model, in turn, can use its language understanding capabilities to re-rank the retrieved codes and select the most appropriate one based on the context. This combination of retrieval and ranking allows our system to capture the necessary specificity and nuance for accurate coding.

The implications of our research extend beyond the scope of this study. Focusing on medical coding area, an accurate, automated coding system could substantially reduce the workload on medical coders, minimize coding errors, and improve the overall quality of medical records. This could lead to more efficient billing processes, better resource allocation, and improved patient care through enhanced data quality for medical research and decision-making. The Retrieve-Rank approach can be continuously adapted for use with ever-improving LLMs and newer coding systems, such as ICD-11 [25]. Additionally, this approach could potentially be extended to other tasks involving the retrieval and application of medical knowledge, such as clinical decision support or medical literature search.

While the results are promising, the research harbors certain limitations. Our evaluation was conducted on a relatively small dataset of 100 samples with simplified single-term inputs. Further testing on larger, more diverse datasets with more complex, realistic medical cases is necessary to confirm the system's generalizability and robustness. Additionally, evaluating the system's performance on larger, more diverse datasets with complex, real-world medical cases will be crucial for assessing its generalizability and robustness. The next step for research should involve putting the system into a real world data set, and continue development should proceed with coding from unstructured clinical notes from electronic health system. This will be the ultimate goal to achieve that can make the whole process even more automatic and minimize clerical work for human especially medical professionals.

Conclusion

Our study presents a promising step towards more efficient and accurate automated medical coding by combining the Retrieve-Rank system with the existing LLM technology. The system's perfect accuracy across a diverse set of medical conditions, compared to the low accuracy of a Vanilla LLM, demonstrates the significant potential of retrieval-based approaches in enhancing the performance of LLMs approach in complex coding tasks. Our findings contribute to the ongoing effort to improve the efficiency and accuracy of medical coding through AI-based approaches. The implications of this research are substantial for healthcare administration, potentially reducing the workload on medical coders, minimizing errors, and improving the overall quality of medical records.

List of Abbreviations: LLMs-Large language models; AI-Artificial intelligence; GPT-General pretrained transformer; RAG-Retrieval-Augmented Generation.

Author Contributions: KK performed the roles of conceptualization, data curation, formal analysis, methodology, validation, and writing the original draft. HC supervised and reviewed the original draft and revisions. BC performed the role of supervision, review, writing of the manuscript, editing, and resources. All authors read and approved the final manuscript.

Funding: This research received no funding.

Ethics Statement: No patient data is involved and hence no ethical application is required.

Data Availability: The complete dataset of 100 medical cases, which includes identifications from both systems, is available as a supplementary file with this submission.
The code used to conduct and analyse the experiments is publicly available on GitHub at https://github.com/ainativehealth/GoodMedicalCoder. This repository contains:
• Python scripts for running the ICD-10 code retrieval experiment (experiment.py)
• Code for creating the index using the RAG model (index.py)
• ICD-10 code datasets (ICD-10.csv and ICD-10_formatted.csv)
• Requirements file listing all necessary Python dependencies (requirements.txt)
• Detailed instructions for reproducing the experiments
Researchers interested in replicating or building upon this work can access all necessary code and data through this GitHub repository. The repository is open-source and licensed under the Apache-2.0 license, allowing for broad use and adaptation of the materials.

Conflict of Interest: The authors declare no conflict of interest.

References

1. Lucyk K, Lu M, Sajobi T, Quan H. Administrative health data in Canada: lessons from history. BMC Medical Informatics and Decision Making. 2015;15(1):69. https://doi.org/10.1186/s12911-015-0196-9.
2. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring Diagnoses: ICD Code Accuracy. Health Services Research. 2005;40(5p2):1620–1639. https://doi.org/10.1111/j.1475-6773.2005.00444.x.
3. Organization WH. Importance of ICD. World Health Organization (Accessed 22 December 2025) Available from: https://www.who.int/standards/classifications/frequently-asked-questions/importance-of-icd
4. Statistics NCfH. ICD-10-CM. (Accessed 9 December 2025). Available from: https://www.cdc.gov/nchs/icd/icd-10-cm/index.html#:~:text=ICD%2D10%2DCM%20(the,CM%20codes%20when%20diagnosing%20patients.
5. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural Natural Language Processing for unstructured data in electronic health records: A review. Computer Science Review. 2022;46:100511. https://doi.org/10.1016/j.cosrev.2022.100511.
6. Liu S, Wang X, Hou Y, Li G, Wang H, Xu H, et al. Multimodal Data Matters: Language Model Pre-Training Over Structured and Unstructured Electronic Health Records. IEEE Journal of Biomedical and Health Informatics. 2023;27(1):504–514. https://doi.org/10.1109/JBHI.2022.3217810
7. OpenAI. OpenAI. OpenAI. (Accessed 9 December 2025). Available from: https://openai.com.

8. Soroush A, Glicksberg BS, Zimlichman E, Barash Y, Freeman R, Charney AW, et al. Large Language Models Are Poor Medical Coders — Benchmarking of Medical Code Querying. NEJM AI. 2024;1(5):AIdbp2300040. https://doi.org/10.1056/AIdbp2300040.
9. Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chunget HW, al. Large language models encode clinical knowledge. Nature. 2023;620(7972):172–180. https://doi.org/10.1038/s41586-023-06291-2.
10. Finkelstein J, Cui W, Morgan K, Kawamoto K. Reducing Diagnostic Uncertainty Using Large Language Models. 2024 IEEE First International Conference on Artificial Intelligence for Medicine, Health and Care (AIMHC), Laguna Hills, CA, USA, 2024, pp. 236-242. https://doi.org/10.1109/AIMHC59811.2024.00049.
11. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. 2020;33:9459–9474.
12. Petroni F, Piktus A, Fan A, Lewis P, Yazdani M, De Cao N, et al. KILT: a Benchmark for Knowledge Intensive Language Tasks. Association for Computational Linguistics; 2021:2523–2544. https://doi.org/10.18653/v1/2021.naacl-main.200.
13. Khattab O, Zaharia M. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. presented at: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020; Virtual Event, China. https://doi.org/10.1145/3397271.3401075
14. Qu C, Zamani H, Yang L, Croft WB, Learned-Miller E. Passage retrieval for outside-knowledge visual question answering. SIGIR '21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021, pp. 1753–1757. https://doi.org/10.1145/3404835.3462987
15. Services CfMM. ICD code lists. (Accessed 9 December 2025). Available from: https://www.cdc.gov/nchs/icd/icd-10-cm/files.html
16. Noor S, Tajik O, Golzar J. Simple Random Sampling. International Journal of Education & Language Studies. 2022;1(2):78–82. https://doi.org/10.22034/ijels.2022.162982.
17. Khattab O, Zaharia M. Colbert: Efficient and effective passage search via contextualized late interaction over BERT. SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2020, pp. 39–48. https://doi.org/10.1145/3397271.3401075.
18. OpenAI. GPT-3.5 Turbo. OpenAI. (Accessed 9 December 2024). Available from: https://chatgpt.com/g/g-F00faAwkE-open-a-i-gpt-3-5.
19. Azamfirei R, Kudchadkar SR, Fackler J. Large language models and the perils of their hallucinations. Crit Care. 2023;27(1):120. https://doi.org/10.1186/s13054-023-04393-x.
20. Ong J, Kedia N, Harihar S, Vupparaboina SC, Singh SR, Venkatesh R, et al. Applying large language model artificial intelligence for retina International Classification of Diseases (ICD) coding. Journal of Medical Artificial Intelligence. 2023;6:21. https://doi.org/10.21037/jmai-23-106.
21. Ke YH, Jin L, Elangovan K, Abdullah HR, Liu N, Sia ATH, et al. Retrieval augmented generation for 10 large language models and its generalizability in assessing medical fitness. NPJ Digit Med. 2025;8(1):187. https://doi.org/10.1038/s41746-025-01519-z.
22. Hallucination Leaderboard. (Accessed 9 January 2026). Available from: https://github.com/vectara/hallucination-leaderboard?tab=readme-ov-file
23. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. J Am Med Inform Assoc. 2010;17(6):646–51. https://doi.org/10.1136/jamia.2009.001024
24. Norman GR, Monteiro SD, Sherbino J, Ilgen JS, Schmidt HG, Mamede S. The Causes of Errors in Clinical Reasoning: Cognitive Biases, Knowledge Deficits, and Dual Process Thinking. Acad Med. 2017;92(1):23–30. https://doi.org/10.1097/acm.0000000000001421.
25. Harrison JE, Weber S, Jakob R, Chute CG. ICD-11: an international classification of diseases for the twenty-first century. BMC Med Inform Decis Mak. 2021;21(Suppl 6):206. https://doi.org/10.1186/s12911-021-01534-6.