Journal of Dental Sciences 20 (2025) 368—372

Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.e-jds.com

Original Article

Evaluating the image recognition capabilities of GPT-4V and Gemini Pro in the Japanese national dental examination

Hikaru Fukuda [a], Masaki Morishita [b,c*], Kosuke Muraoka [c], Shino Yamaguchi [d], Taiji Nakamura [e], Izumi Yoshioka [f], Shuji Awano [c], Kentaro Ono [g]

[a] *Division of Maxillofacial Surgery, Department of Science of Physical Functions, Kyushu Dental University, Kitakyushu, Japan*
[b] *Health Information Management Office, Kyushu Dental University Hospital, Kitakyushu, Japan*
[c] *Division of Clinical Education Development and Research, Department of Oral Function, Kyushu Dental University, Kitakyushu, Japan*
[d] *School of Oral Health Sciences, Kyushu Dental University, Kitakyushu, Japan*
[e] *Division of Periodontology, Department of Oral Function, Kyushu Dental University, Kitakyushu, Japan*
[f] *Division of Oral Medicine, Department of Science of Physical Functions, Kyushu Dental University, Kitakyushu, Japan*
[g] *Division of Physiology, Department of Health Promotion, Kyushu Dental University, Kitakyushu, Japan*

Received 9 May 2024; Final revision received 21 June 2024
Available online 2 July 2024

KEYWORDS
ChatGPT-4V;
Gemini Pro;
Japanese national dental examination;
Large language models

Abstract *Background/purpose:* OpenAI's GPT-4V and Google's Gemini Pro, being Large Language Models (LLMs) equipped with image recognition capabilities, have the potential to be utilized in future medical diagnosis and treatment, ands serve as valuable educational support tools for students. This study compared and evaluated the image recognition capabilities of GPT-4V and Gemini Pro using questions from the Japanese National Dental Examination (JNDE) to investigate their potential as educational support tools.

Materials and methods: We analyzed 160 questions from the 116th JNDE, administered in March 2023, using ChatGPT-4V, and Gemini Pro, which have image recognition functions. Standardized prompts were used for all LLMs, and statistical analysis was conducted using Fisher's exact test and the Mann—Whitney U test.

Results: For the 160 JNDE questions, the accuracy rates of GPT-4V and Gemini Pro were 35.0% and 28.1%, respectively, with GPT-4V being the highest, although not statistically significant. Across dental specialties, the accuracy rates of the GPT-4V were generally higher than those

[*] Corresponding author. Kyushu Dental University, Division of Clinical Education Development and Research, Department of Oral Function, 2-6-1 Manazuru, Kokurakita, Kitakyushu 803-8580, Japan.
E-mail address: r08morishita@fa.kyu-dent.ac.jp (M. Morishita).

https://doi.org/10.1016/j.jds.2024.06.015

of the Gemini Pro, with some areas showing equal accuracy. Accuracy rates tended to decrease with an increased number of images within a question, suggesting that the number of images influenced the correctness of the responses.

Conclusion: The overall superior performance of GPT-4V compared to Gemini Pro may be attributed to the continuous updates in OpenAI's model. This research demonstrates the potential of LLMs as educational support tools in dentistry, while also highlighting areas that require further technological development.

Introduction

Large Language Models (LLMs) are anticipated to complement traditional learning methods such as textbooks, lectures, and practical exercises through AI-driven learning experiences, thereby supporting students' learning processes.[1—3] ChatGPT-4V developed by OpenAI (September 2023 model) and Gemini Pro by Google (December 13, 2023 model) are classified as multimodal AIs, equipped with image recognition capabilities as LLMs.[4,5] OpenAI has explicitly stated that the current image recognition features are not suited for medical imaging. However, LLMs with image recognition capabilities not only serve as learning support tools but also have the potential of future applications in patient diagnosis and treatment within the medical field.[6,7] Therefore, evaluating the current scope of image recognition abilities of GPT-4V and Gemini Pro is beneficial.

The Japanese National Dental Examination (JNDE) includes a variety of images such as intraoral and extraoral photographs, CT and MRI scans, ultrasound, panoramic X-ray images, dental X-ray images, dental technical work, charts, illustrations, tables, and orthodontic polygonal charts.[8] This study was conducted to reveal the accuracy rates of GPT-4V and Gemini Pro when using image recognition capabilities to answer questions with various types of image information in the JNDE conducted in January 2023. We aimed to investigate the potentials of GPT-4V and Gemini Pro as educational support tools.

Materials and methods

Obtaining and processing data from the Japanese national dental examination

We collected questions and their corresponding correct answers from the 116th Japanese National Dental Examination, administered in March 2023, from the Ministry of Health, Labour and Welfare (MHLW) website of Japan. Specifically, we selected 160 questions from the JNDE, including intraoral and extraoral photographs, CT, and MRI scans, echo, panoramic X-ray images, dental X-ray images, dental laboratory work, diagrams, and charts like orthodontic polygon tables, illustrations, and tables. To input questions with images, charts, and tables, we used ChatGPT-4V and Gemini Pro, both equipped with image recognition functions. We then analyzed the data, including the percentage of correct answers and other relevant information. The JNDE question criteria indicate that the JNDE will have three areas: compulsory, general, and clinical practical. Of the 160 questions used in this study, seven fell under the compulsory area, 55 under the general area, and 98 under clinical practice.

Large language models

The LLMs used were ChatGPT-4V from OpenAI (model: September 2023) and Gemini Pro from Google (model: December 13, 2023). These LLMs are classified as multimodal AI, capable of analyzing and outputting data based on two or more types of data, such as text, audio, images, videos, and numbers. Standardized prompt input and question text, along with charts and images, were individually copied and pasted into the corresponding web interface of each LLM, and all responses were recorded. The inputs were submitted on December 20, 2023.

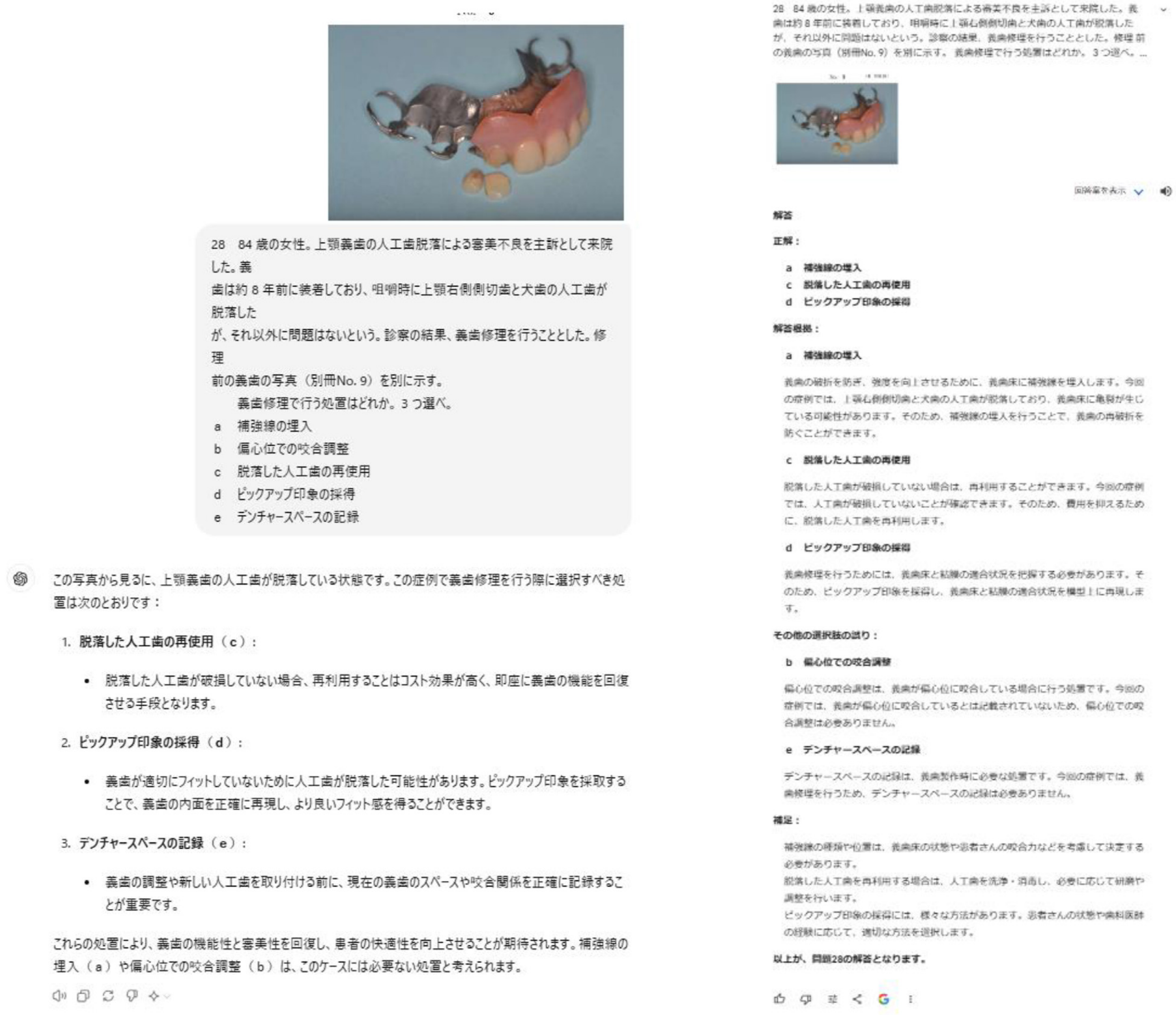
A prompt is "an instruction given to an LLM to enforce a specific rule, automate a process, or guarantee a certain quality and quantity of the output produced." The format of the prompts for entering question text and images was standardized as follows: "You are a student taking the National Dental Examination. Please first provide the correct answers based on the text of the questions. Next, please explain your choice and why the other choices are inappropriate."

Data and statistical analysis

For data analysis, we used Qlik Sense® Enterprise August 2022 Patch 2 (Qlik Technologies Inc., King of Prussia, PA, USA). Statistical analysis was conducted using GraphPad Prism 9.5.1 (GraphPad Software, Boston, MA, USA) employing Fisher's exact test and Mann—Whitney U test.

Results

Figure 1 shows the interface of GPT-4V and Gemini Pro as it processes inputs from JNDE and illustrates a pivotal moment in the present study. The inputs include both textual questions and photographs. Upon submission of these inputs, the system generated corresponding responses.



28　84 歳の女性。上顎義歯の人工歯脱落による審美不良を主訴として来院した。義
歯は約 8 年前に装着しており、咀嚼時に上顎右側側切歯と犬歯の人工歯が脱落した
が、それ以外に問題はないという。診察の結果、義歯修理を行うこととした。修
理
前の義歯の写真（別冊No. 9）を別に示す。
　義歯修理で行う処置はどれか。3 つ選べ。
a　補強線の埋入
b　偏心位での咬合調整
c　脱落した人工歯の再使用
d　ピックアップ印象の採得
e　デンチャースペースの記録

この写真から見るに、上顎義歯の人工歯が脱落している状態です。この症例で義歯修理を行う際に選択すべき処置は次のとおりです：

1. **脱落した人工歯の再使用（c）：**
   - 脱落した人工歯が破損していない場合、再利用することはコスト効果が高く、即座に義歯の機能を回復させる手段となります。
2. **ピックアップ印象の採得（d）：**
   - 義歯が適切にフィットしていないために人工歯が脱落した可能性があります。ピックアップ印象を採取することで、義歯の内面を正確に再現し、より良いフィット感を得ることができます。
3. **デンチャースペースの記録（e）：**
   - 義歯の調整や新しい人工歯を取り付ける前に、現在の義歯のスペースや咬合関係を正確に記録することが重要です。

これらの処置により、義歯の機能性と審美性を回復し、患者の快適性を向上させることが期待されます。補強線の埋入（a）や偏心位での咬合調整（b）は、このケースには必要ない処置と考えられます。

28　84 歳の女性。上顎義歯の人工歯脱落による審美不良を主訴として来院した。義歯は約 8 年前に装着しており、咀嚼時に上顎右側側切歯と犬歯の人工歯が脱落したが、それ以外に問題はないという。診察の結果、義歯修理を行うこととした。修理 前の義歯の写真（別冊No. 9）を別に示す。義歯修理で行う処置はどれか。3 つ選べ。...

解答

正解：

a　補強線の埋入
c　脱落した人工歯の再使用
d　ピックアップ印象の採得

解答根拠：

a　補強線の埋入

義歯の破折を防ぎ、強度を向上させるために、義歯床に補強線を埋入します。今回の症例では、上顎右側側切歯と犬歯の人工歯が脱落しており、義歯床に亀裂が生じている可能性があります。そのため、補強線の埋入を行うことで、義歯の再破折を防ぐことができます。

c　脱落した人工歯の再使用

脱落した人工歯が破損していない場合は、再利用することができます。今回の症例では、人工歯が破損していないことが確認できます。そのため、費用を抑えるために、脱落した人工歯を再利用します。

d　ピックアップ印象の採得

義歯修理を行うためには、義歯床と粘膜の適合状況を把握する必要があります。そのため、ピックアップ印象を採得し、義歯床と粘膜の適合状況を模型上に再現します。

その他の選択肢の誤り：

b　偏心位での咬合調整

偏心位での咬合調整は、義歯が偏心位に咬合している場合に行う処置です。今回の症例では、義歯が偏心位に咬合しているとは記載されていないため、偏心位での咬合調整は必要ありません。

e　デンチャースペースの記録

デンチャースペースの記録は、義歯製作時に必要な処置です。今回の症例では、義歯修理を行うため、デンチャースペースの記録は必要ありません。

補足：

補強線の種類や位置は、義歯床の状態や患者さんの咬合力などを考慮して決定する必要があります。
脱落した人工歯を再利用する場合は、人工歯を洗浄・消毒し、必要に応じて研磨や調整を行います。
ピックアップ印象の採得には、様々な方法があります。患者さんの状態や歯科医師の経験に応じて、適切な方法を選択します。

以上が、問題28の解答となります。

Figure 1 A screenshot depicting the process where question text and intraoral photographs from the Japanese national dental examination were entered into GPT-4V and Gemini Pro, resulting in the generation of responses.

Table 1 shows the number of correct answers, incorrect answers, and percentage of correct answers obtained for each of the 160 questions using GPT-4V and Gemini Pro. The correct response rates for GPT-4V and Gemini Pro were 35.0% and 28.1%, respectively, indicating that the correct response rate for GPT-4V was higher but not statistically significant.

Table 2 shows the percentage of correct answers for GPT-4V and Gemini Pro across the compulsory, general, and clinical practical areas of the JNDE. The correct response rates for GPT-4V and Gemini Pro were 57.1% and 28.6% in compulsory area, 43.6% and, 29.1% in general area, and 28.6% and 27.6% in clinical practice, respectively. Fisher's exact test was performed on the number of correct and incorrect answers from GPT-4V and Gemini Pro for the compulsory, general, and clinical practice questions;

Table 1 Performance of large language models (LLMs) in the Japanese national dental examination.

	Questions (n)	Correct (n)	Incorrect (n)	Correct answer (%)
GPT-4V	160	56	104	35
Gemini Pro	160	45	115	28.1

Table 2 The percentage of correct answers for GPT-4V and Gemini Pro for the compulsory, general, and clinical practical areas.

Area	Questions (n)	Correct answer (%)	
		GPT-4V	Gemini Pro
Compulsory	7	57.1	28.6
General	55	43.6	29.1
Clinical practical	98	28.6	27.6

however, no statistically significant differences were found. GPT-4V had a higher percentage of correct answers in all three areas than Gemini Pro.

Table 3 presents the percentage of positive matches between GPT-4V and Gemini Pro in the three areas. Compliance rates were 57.1%, 41.8 %, and 28.6% for compulsory, general, and clinical practice, respectively. The overall compliance rate for the 160 questions was 34.4%.

Table 4 outlines the number of questions by dental subject and the percentage of correct answers for GPT-4V and Gemini Pro, ordered by the percentage of correct answers for GPT-4V. Dental anesthesiology and endodontics had more than 70% correct responses on GPT-4V, whereas Gemini Pro had fewer than 50% correct responses. The correct response rates were generally higher for GPT-4V, with some subjects having identical correct response rates. Only for Partial Dentures did, Gemini Pro exhibit a higher percentage of correct answers than GPT-4V. Overall, GPT-4V showed more correct answers than Gemini Pro.

Table 5 demonstrates the differences in the number of images inputted into GPT-4V and Gemini Pro for correct and incorrect responses. GPT-4V showed that, statistically, fewer images were inputted for correct responses. Gemini Pro tended to answer correctly, using fewer images. However, it showed more images than GPT-4V, and the number of inputted images for correct and incorrect responses was not statistically significant.

Lastly, Table 6 shows the number of images inputted into GPT-4V and Gemini Pro for illustrations, highlighting the significant differences between the images and charts in this study. Both models showed significantly more correct responses when fewer images were inputted.

Discussion

In the Japanese national dental examination, the accuracy rates of GPT-4V and Gemini Pro in response to questions were higher for GPT-4V across the mandatory, general, and clinical practice areas, although the difference was not statistically significant. This suggests that while LLMs have evolved to understand and respond to complex question formats, their level of understanding in specialized fields, such as dentistry, remains a challenge.[9–11] According to OpenAI 4, the interpretation of medical images by GPT-4V is inconsistent, and the model sometimes gives accurate and incorrect answers to the same questions. The lack of statistically significant differences between LLMs across different dental areas in the JNDE highlights the need for further enhancement of LLM features, particularly to improve our understanding of domain-specific knowledge.

Table 3 The percentage of positive matches between GPT-4V and Gemini Pro in the three areas.

	Questions (n)	Match (%)
Compulsory	7	57.1
General	55	41.8
Clinical practical	98	28.6
Total	160	34.4

Table 4 The number of questions by dental subject and the percentage of correct answers for GPT-4V and Gemini Pro.

Subjects	Questions (n)	Correct answer (%)	
		GPT-4V	Gemini Pro
Dental anesthesiology	4	75	50
Endodontics	7	71.4	42.9
Oral public health	8	50	50
Dental radiology	6	50	33.3
Dental materials	2	50	0
Pediatric dentistry	16	43.8	25
Oral surgery	34	38.2	38.2
Full denture	12	33.3	25
Restorative dentistry	13	30.8	30.8
Partial denture	11	27.3	36.4
Orthodontics	20	25	25
Periodontics	13	23.1	0
Crown and bridge	11	9.1	9.1
Dental anatomy	1	0	0
Oral physiology	1	0	0
Oral pathology	1	0	0

Table 5 The difference in the number of images inputted into GPT-4V and Gemini Pro for correct and incorrect responses.

Number of images inputted	Mean ± Standard deviation		*P*-value
	Correct	Incorrect	
GPT-4V	2.48 ± 1.83	3.64 ± 2.52	0.003
Gemini Pro	3.11 ± 2.29	3.28 ± 2.40	0.73

Table 6 The number of images inputted into GPT-4V and Gemini Pro for illustrations.

Illustration	Mean ± Standard deviation		*P*-value
	Correct	Incorrect	
GPT-4V	0.16 ± 0.42	0.30 ± 0.46	0.04
Gemini Pro	0.11 ± 0.32	0.30 ± 0.48	0.02

This study demonstrated the importance of image recognition capabilities in answering JNDE questions. The introduction of LLMs with image recognition features represents a significant advancement in LLM technology, enabling the analysis of both textual and visual information.[12–14] However, our data indicate that the current image recognition capabilities of LLMs are insufficient for interpreting images in specialized medical fields. Interestingly, our research data showed that the number of images included in a question affects the accuracy rate of the LLM, with a tendency toward higher accuracy with fewer images. This suggests that the quantity and complexity of the visual information affect the LLM's image recognition capabilities, thereby clearly identifying potential improvements for future model updates.

One limitation of our study is that although we assessed the accuracy rate of the LLMs' responses, we should have evaluated the quality of the explanations provided by the LLMs. In future research, it would be beneficial to include an evaluation of these models' explanatory capabilities to provide a more comprehensive assessment of their performance. Moreover, we acknowledge the limitation that only one question was given to some dental subjects, and only one assessment of correct responses was made. However, the results of this study offer valuable insights into the image recognition deficits of the two LLM models. Future studies should consider assessing the correct response rate multiple times on the same dataset to enhance our understanding. This approach could lead to different results, offer a more comprehensive understanding of model performance, and help identify consistent patterns and areas for improvement in image recognition performance. Therefore, it is essential to interpret the results cautiously, as some subjects were given only one question, and the correct response rate results were assessed only once.

The overall superior performance of GPT-4V compared to Gemini Pro in our study may be attributed to continuous updates in OpenAI's model.[15] Furthermore, the fact that GPT-4V showed superior accuracy rates in specific areas, such as dental anesthesia and endodontics, suggesting that there are certain fields where the current LLM's image recognition capabilities are more suitable.

In conclusion, this study demonstrates the potential of LLMs as educational support tools in dentistry, while also highlighting areas that require further technological development. In addition to the dental medical questions used in this study, it is necessary to evaluate the capabilities of LLMs for questions in other specialized fields. Advancements in LLMs, such as improved image recognition capabilities and the acquisition of specialized knowledge, can significantly transform dental education by complementing textbooks, lectures, and practical exercises with AI-driven learning experiences, thereby supporting students in their learning process.

Declaration of competing interest

The authors have no conflicts of interest relevant to this article.

Acknowledgments

This work was not supported by any organizations.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015;521:436—44.
2. Chui M, Manyika J, Miremadi M, et al. *Notes from the AI frontier: insights from hundreds of use cases*. Available at: https://www.mckinsey.com/&#x223C;/media/mckinsey/featured%20insights/artificial%20intelligence/notes%20from%20the%20ai%20frontier%20applications%20and%20value%20of%20deep%20learning/notes-from-the-ai-frontier-insights-from-hundreds-of-use-cases-discussion-paper.ashx. McKinsey Global Institute. [Date accessed: March 2024].
3. Jeon J, Lee S. Large language models in education: a focus on the complementary relationship between human teachers and ChatGPT. *Educ Inf Technol* 2023;28:15873—92.
4. *GPT-4V(ision) system card*. Available at: https://cdn.openai.com/papers/GPTV_System_Card.pdf. [Accessed 15 March 2024].
5. Bard's latest updates. *Access Gemini Pro globally and generate images*. Available at: https://blog.google/products/gemini/google-bard-gemini-pro-image-generation/. [Accessed 15 March 2024].
6. Deng L, Liu Y. *Deep learning in natural language processing*, 329. Springer, 2018:1—22.
7. Huang G, Liu Z, Van DM, et al. Densely connected convolutional networks. *arXiv* 2018;1608:06993.
8. The Ministry of Health, Labour and Welfare of Japan. *The Japanese national dental examination 2023*. Available at: https://www.mhlw.go.jp/seisakunitsuite/bunya/kenkou_iryou/iryou/topics/tp230524-02.html. [Accessed 15 March 2024].
9. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? *JMIR Med Educ* 2023;9:e45312.
10. Takagi S, Watari T, Erabi A, Sakaguchi K. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: comparison Study. *JMIR Med Educ* 2023;9:e48002.
11. Taira K, Itaya T, Hanada A. Performance of the large language model ChatGPT on the national nurse examinations in Japan: evaluation Study. *JMIR Nurs* 2023;6:e47305.
12. American Foundation for the Blind. *GPT-4 image recognition: an absolute game changer in accessibility*. Available at: https://www.afb.org/blog/entry/gpt-4-image-recognition-accessibility. [Accessed 15 March 2024].
13. AI-Scholar. *[mPLUG-Owl] Developing an LLM that can understand images and text*. Available at: https://ai-scholar.tech/articles/computation-and-language%2FmPLUG-Owl. [Accessed 15 March 2024] [Date accessed.
14. Yang S, Shang Z, Wang Y, et al. Data-free multi-label image recognition via LLM-powered prompt tuning. *arXiv* 2024;2403:01209.
15. Microsoft Research AI. 4Science. The impact of large language models on scientific discovery: a preliminary study using GPT-4. *arXiv* 2024;2311:07361.