

Recognizing Images and Extracting Useful Inferences by Asking Large Language Models Simple Questions


Nektarios Ioannis Kontolaimakis^{1,*} and Nicholas Panagiotis Sgouros²

ABSTRACT

Laboratory exercises are an essential component of engineering education while the increasing trend towards distance learning presents unique challenges in replicating hands-on experiences. A number of AI-driven solutions have been proposed to facilitate remote laboratory exercises, however the emergence of Multimodal Large Language Models offers novel possibilities for visual recognition in remote settings. Vision AI, a subfield of artificial intelligence, enhances LLM capabilities by allowing them to process visual data through tasks like image recognition and segmentation, making it particularly relevant for use in educational applications. This work evaluates the integration of Vision AI into LLMs like OpenAI's GPT-4 and Anthropic's Claude 3.5 Sonnet, examining their ability to recognize and understand images from laboratory devices such as displays, gauges, and control panels. Our study focuses on the estimation of the performance of GPT-4 and Claude 3.5 Sonnet in laboratory-related image recognition tasks, with results indicating similar high text recognition accuracy (92% for GPT-4 and 91% for Claude 3.5). Despite these successes, challenges persist in spatial awareness and object identification, which are critical for accurate interpretation of complex lab environments. These findings highlight the potential of Vision AI to support remote laboratory exercises, improve accessibility for students in geographically distributed settings, or students with disabilities, and enhance interactive learning tools in STEM education. Future work will focus on refining these capabilities through custom LLM development, advanced prompt engineering, and multimodal approaches, aiming to create more versatile and effective educational technologies for remote and hybrid learning environments.

Submitted: November 04, 2024

Published: December 30, 2024

 10.24018/lejeng.2024.1.CIE.3233

¹Laboratory Center of Neapolis Lasithiou, Greece.

²Department of Naval Engineering, University of West Attica, Greece.

*Corresponding Author:
e-mail: nekont@sch.gr

Keywords: AI, Image Recognition, Large Language Models, Vision AI.

1. INTRODUCTION

In recent years, Large Language Models (LLMs) have made significant strides in various applications, from natural language processing to multimodal tasks that include image recognition. LLMs, such as OpenAI's GPT-4 and Anthropic's Claude 3.5 Sonnet, have demonstrated remarkable abilities in visual comprehension, opening up new possibilities for integrating these capabilities in real-world scenarios. This paper explores the performance of the above models in recognizing and interpreting images, particularly focusing on their ability to accurately reporting the information shown in different types of experimental device displays, gauges, and control panels.

An LLM has been trained on a massive amount of text data to interpret natural language and generate human-like responses to text-based prompts or questions [1].

These systems models use a transformer architecture, which includes neural networks made up of an encoder and a decoder equipped with self-attention mechanisms. The encoder and decoder work together to derive meanings from a text sequence and comprehend the relationships between its words and phrases [2].

Moreover, Multimodal Large Language Models (MM-LLMs) represent a class of large-scale neural networks trained on a wide variety of data, including textual, visual, and occasionally auditory input [3]. One of the key areas of research into MM-LLM has been the exploration of their visual capabilities [4].



In the last two years there is strong competition between models developed by different vendors and organizations. Examples include OpenAI models such as GPT-3, GPT-3.5, GPT 4, GPT 4o, Google DeepMind's Gemini model, Anthropic's Claude 3 and Claude 3.5 Sonnet models and several custom-made models created by research groups and organizations [5]. Two of the most recent models released to the public are OpenAI's ChatGPT 4o (omni) and Anthropic's Claude 3.5 Sonnet model.

While the primary focus of this research is on the assessment of the technical capabilities of LLMs within a laboratory setting, there are additional significant educational implications. The integration of Vision AI into educational environments could revolutionize how students interact with complex visual materials, such as diagrams, control systems, and experimental device displays and gauges. By enhancing tools for STEM education and accessibility, Vision AI has the potential to improve educational outcomes, making distant learning more realistic, interactive and inclusive. This paper aims to bridge the gap between cutting-edge AI research and its potential applications in education, providing insights that could be applied in remote labs and accessibility tools, as well as remote student experiential learning.

2. EDUCATIONAL IMPLICATIONS

The findings from this research on image recognition and inference using LLMs have several key implications on educational practices, particularly in the field of STEM education. As the capabilities of LLMs evolve, their potential to improve educational environments and associated tools becomes more evident.

LLMs' ability to recognize and describe images with high accuracy has the potential to revolutionize remote participation in laboratory settings. As an example, Vision AI could assist educators by displaying real displays and gauges to students while superimposing correct readings to direct students to gain experience in display and gauge readings for different real world experimental devices. Underlying tasks such as reading technical diagrams, interpreting data from digital displays, or analyzing control panel configurations may eventually be streamlined through AI-based systems. By incorporating Vision AI, educational platforms could provide immediate feedback to students, enhancing their understanding and allowing for more efficient learning.

In engineering STEM education, hands-on experience with technical systems is a vital step in the learning process. However, the availability of such physical systems is often limited, or such systems are not reachable due to geographical or other constraints. LLMs integrated with Vision AI could simulate real-world environments, such as virtual labs or industrial settings. For instance, students studying engineering or mechanics could interact with virtual control panels, gauges, and machinery, while the AI provides real-time analysis and feedback on their inputs. This would not only improve engagement but also make it possible for students to practice complex tasks on actual devices remotely.

One of the most promising applications of Vision AI in education is improving accessibility for students with disabilities, particularly those with visual impairments. LLMs with image recognition capabilities can serve as assistive technologies by describing technical equipment and display or gauge readings in real-time. For example, students in science or engineering fields who rely on visual aids can use Vision AI to receive spoken or textual descriptions, making education more inclusive. This aligns with the broader goal of using AI to break down barriers and create inclusive learning environments.

Vision AI could offer improved assessment tools in order to evaluate students' ability to correctly interpret readings from control panels, identify system errors, or recognize values on display and gauges. By integrating LLMs into these assessments, educators could obtain more objective and detailed evaluations of students' competencies in real time, ultimately enhancing the learning process.

3. LARGE LANGUAGE MODELS

3.1. ChatGPT 4o

The main improvements and differences of GPT 4o compared to previous OpenAI models are its larger size and greater complexity. This gives it the ability to understand and produce more sophisticated text. It uses improved techniques and training methodologies with improved algorithms and better handling of training data. Also the data with which it is trained is more and more diverse. Thus it is possible to generate more accurate and appropriate answers to more topics and scenarios.

The model has been trained to produce safer content by following ethical rules so as to minimize the possibility of producing harmful and inappropriate content. Based on metrics GPT 4o has improved performance on various standard metrics such as question responses, generating summaries, translation and more.

3.2. Claude 3.5 Sonnet

Claude 3.5 developed by Anthropic is also an improved model compared to other previous models from the same company. Claude 3.5 has been trained on larger and more diverse datasets which allows it to generate more accurate and correct answers to various queries. The architecture of the model has been improved with individual optimization of the algorithm running in the background and the neural network structures. These have resulted in an increase in the speed of generating correct answers.

It also ensures to a greater extent the security of the answers and their alignment with human values and ethics. This minimizes the possibility of producing answers that are harmful to humans. Conversations with this model can be longer and thus can remember more data which leads to more relevant and accurate answers.

4. VISION

4.1. Vision AI

Vision AI, also referred to as computer vision, is a field within artificial intelligence that enables machines to interpret and understand visual information from the world. This technology involves several key tasks such as image recognition, image classification, object detection, image

segmentation, and video analysis. Vision AI leverages deep learning techniques and neural networks to analyze and process visual data, making it possible for machines to recognize objects, detect anomalies, and perform tasks that require visual comprehension [6].

Vision AI has significantly enhanced the capabilities of AI models like ChatGPT and Claude. Initially, AI models focused primarily on text-based tasks, but the integration

TABLE I: CATEGORY SCREENS

Figures	Answer from ChatGPT 4o	Answer from Claude 3.5
Fig. 1	<p>Here is the text visible in Fig. 1: Basic Refrigeration: V1 V2 V3 V4 V5 V6 V7 V8 CM OF S1 D1 S2 D2 SP PD E1 L1 E2 RT LP HP T1 T2 T3 T4 T5 T6 T7 T8 54 107 27 29 23 29 32 31</p>	<p>Fig. 1 shows a digital display or LCD screen with the following text: Basic Refrigeration: The top line of the display reads: U1 U2 U3 U4 U5 U6 U7 U8 CM OF The second line reads: S1 D1 S2 D2 SP PD E1 L1 E2 RT The third line contains: LP HP T1 T2 T3 T4 T5 T6 T7 T8 Note that "Refrigeration" appears to be a misspelling of "Refrigeration" on the display.</p>
Fig. 2	<p>Here is the text visible in Fig. 2: Basic Refrigeration: V1 V2 V3 V4 V5 V6 V7 V8 CM OF ON ON ON ON S1 D1 S2 D2 SP PD E1 L1 E2 RT S20 LO LP HP T1 T2 T3 T4 T5 T6 T7 T8 42 143 26 48 34 30 21 20</p>	<p>Fig. 2 shows an LCD display with the following text: "Basic Refrigeration" (note: this is misspelled, should be "Refrigeration") Below that, Fig. 2 shows: U1 U2 U3 U4 U5 U6 U7 U8 CM OF ON ON ON ON ON ON S1 D1 S2 D2 SP PD E1 L1 E2 RT 20 5 LO LP HP T1 T2 T3 T4 T5 T6 T7 T8 42 145 26 40 34 30 21 20 This appears to be a display for a refrigeration system, showing various sensor readings and system statuses.</p>
Fig. 3	<p>Here is the text visible Fig. 3: Basic Refrigeration: 20 V1 V2 V3 V4 V5 V6 V7 V8 CM OF ON ON ON ON S1 D1 S2 D2 SP PD E1 L1 E2 RT S20 LO LP HP T1 T2 T3 T4 T5 T6 T7 T8 4 17 25 27 25 26 30 29</p>	<p>Fig. 3 shows an LCD display panel with the following text: At the top: "Basic Refrigeration" (Note: This is misspelled, it should be "Refrigeration") Below that, in rows: 1. U1 U2 U3 U4 U5 U6 U7 U8 CM OF 2. ON ON ON ON ON ON 3. S1 D1 S2 D2 SP PD E1 L1 E2 RT 4. 20 5 LO 5. LP HP T1 T2 T3 T4 T5 T6 T7 T8 6. 4 17 25 27 25 26 30 29 At the bottom of the image, partially visible: "Suction pressure too low" The display is mounted on a white panel with four visible screws, one in each corner.</p>
Fig. 4	<p>Here is the text visible in Fig. 4: SALOON FIRE MASTER CABIN OK GALLEY OK ENGINE ROOM OK MAIN DECK OK UPPER DECK OK LAZZARETTE OK CYLINDER DOOR OPEN</p>	<p>Fig. 4 shows a display panel with the following text: SALOON MASTER CABIN OK GALLEY OK ENGINE ROOM OK MAIN DECK OK UPPER DECK OK LAZARETTE OK CYLINDER DOOR OPEN At the top right corner of the display, it shows "FIRE". Below the display screen, there are buttons labeled: Select Mute Dim ISOL Ent Reset The buttons are arranged in two rows of four buttons each.</p>

TABLE I: CONTINUED

Figures	Answer from ChatGPT 4o	Answer from Claude 3.5
Fig. 5	Here is the text visible in Fig. 5.: 120/80 93 98 60 36.0	Fig. 5 shows a medical monitoring device with a digital display. The text and numbers visible on the screen are: 120/80 93 98 60 36.0 There's also a waveform displayed, which appears to be an ECG or heart rate monitor. At the bottom of the device, there are some control buttons, but their labels are not clearly visible in this image.
Fig. 6	Here is the text visible in Fig. 6.: WEIGHING CONTROLLER 49.98 kg/49.99 kg 0.00/0.00 01 2 buckets/1 bag Auto Start Last Alarm Clear Print Lock Menu User	Fig. 6 shows a display screen for a weighing controller. The main text and numbers visible are: "WEIGHING CONTROLLER" at the top of the screen. Two large numerical displays: "49.98" on the left "49.99" on the right Below these, smaller "0.00" values are displayed under each of the larger numbers. There are several rows of buttons or indicators at the bottom of the screen with various labels, including: "Print", "Clear", "Stop", "Lock", "Menu" and others that are partially visible or too small to read clearly. The display appears to be showing weight measurements in some unit (likely kilograms or pounds), though the specific unit is not visible in this image.
Fig. 7	Time: Current Time: 13:00 Date and Day: Day: 50 Date: 15 (Day) 2 (Month) Weather and Temperature: Weather Icon: Sunny Indoor Temperature: 27°C Outdoor Temperature: 33°C Today's High Temperature: 34°C Today's Low Temperature: 27°C Humidity: Relative Humidity: 43% Humidity Indicator: Smiley face (indicating comfort level)	Here are the text labels and their corresponding values visible in the photo: Time: 13:03 DAY: 50 15 2 IN (Indoor temperature): 27°C OUT (Outdoor temperature): 33°C Hi (High temperature) Today: 34°C Lo (Low temperature) Today: 27°C Humidity: 43% "Life" (brand name at the bottom) Additional indicators: A sun icon (indicating sunny weather) A smiley face icon



Fig. 1. First LCD photo.

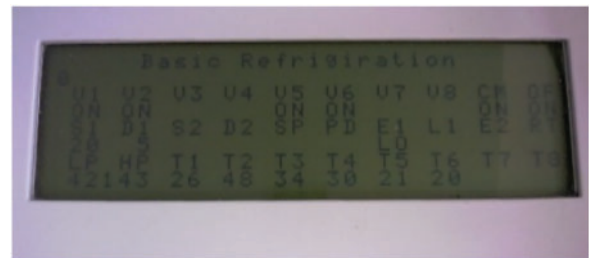


Fig. 2. Second LCD photo.



Fig. 3. Wider LCD photo.

of vision capabilities marked a pivotal evolution. OpenAI's GPT-4 incorporated vision AI, allowing it to interpret and describe images, thus expanding its applications in fields like content creation, customer service, and accessibility. This integration enables the model to analyze visual inputs and provide detailed contextual information. Similarly, Claude, developed by Anthropic, incorporated vision AI to enhance its conversational abilities, making interactions more contextually rich and visually informed. The combination of text and visual understanding in these models exemplifies the advancements in multimodal AI systems, offering more intuitive and comprehensive user experiences [7].

4.2. Visual Ability

Visual ability in the context of LLMs, generally refers to the model's ability to process, interpret, and generate



Fig. 4. Boat system photo.



Fig. 5. Patient monitor photo.



Fig. 6. Weighing controller photo.

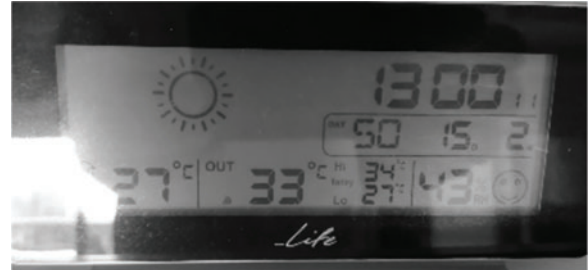


Fig. 7. Weather station photo.

- *Object Detection*: Identifying and locating objects within an image, often used in applications like autonomous driving or surveillance.
- *Visual Reasoning*: Understanding and reasoning about visual content to draw conclusions or make inferences. This can involve tasks like understanding the relationships between objects in an image or predicting future actions in a video.
- *Optical Character Recognition (OCR)*: Extracting and interpreting text from images, which is useful for digitizing printed documents or reading text in images.
- *Image Generation*: Creating images based on textual descriptions. This is seen in models like DALL-E, where the model generates images from detailed textual prompts.

visual information alongside text. This is often achieved by integrating LLMs with visual processing systems, enabling them to handle multimodal tasks. Here are some key aspects:

- *Image Recognition*: The ability of the model to identify and classify objects, scenes, or specific details within an image.
- *Image Captioning*: Generating descriptive text based on the content of an image. This involves understanding the visual elements and producing coherent and contextually appropriate descriptions.
- *Visual Question Answering (VQA)*: The ability to answer questions about the content of an image. This requires understanding the image, interpreting the question, and providing a relevant answer.
- *Text and Image Integration (Visual Language-VL)*: Combining text and visual data to perform tasks that require understanding both modalities. For example, interpreting a diagram with annotations or generating text based on a sequence of images.

5. RESEARCH PURPOSE

The purpose of this paper was to evaluate and compare the capabilities of the ChatGPT 4o and Claude 3.5 Sonnet LLMs in the area of recognition, reading and understanding of specific images.

The research focuses on specific images of clock-type instrument readings, images of displays of various readings and images of displays showing various readings and data; the images should also contain switches of all kinds (multiple switches, on/off switches, etc.) and lights of various colors and types.

In addition to image recognition, logic and judgement questions related to the images have been asked so that we can judge the degree of understanding of the model based on the information contained in the images.

5.1. Fields of Understanding of Models

There are many different fields of understanding of LLMs in which performance is different from one another.

For example one area is the performance of the model on various exams such as performance on United States Medical Licensing Examination (USMLE), performance on

TABLE II: CATEGORY GAUGES

Figures	Answer from ChatGPT 4o	Answer from Claude 3.5
Fig. 8	The needle on the gauge points to approximately 28 psi.	The needle on this pressure gauge is pointing to approximately 25 on the outer scale. This corresponds to about 25 psi (pounds per square inch) or around 1.75 kg/cm ² on the inner scale.
Fig. 9	The needle on the gauge points to approximately 225 psi.	The needle on this pressure gauge is pointing to approximately 270 on the outer scale. This corresponds to about 270 psi (pounds per square inch) on the main scale of the gauge.
Fig. 10	The needle on the gauge points to approximately 230 psi.	In this image, the needle of the pressure gauge is pointing to approximately 330 on the outer scale. This corresponds to about 330 psi (pounds per square inch) on the main scale of the gauge.
Fig. 11	The needle on the gauge points to approximately 25 psi.	In this image, the needle of the pressure gauge is pointing to approximately 25 on the outer scale. This corresponds to about 25 psi (pounds per square inch) on the main scale of the gauge.
Fig. 12	The needle on the gauge points to approximately 255 psi.	In this image, the needle is pointing to approximately: - 350 psi on the outer pressure scale - 24 bar on the inner pressure scale (between 20 and 25 bar) - 70°C on the temperature scale (the inner green and red circle) The gauge appears to be a combination pressure and temperature gauge for refrigerants, likely used in HVAC or refrigeration systems. The needle is indicating high pressure and temperature readings.
Fig. 13	The needle on the gauge points to approximately 255 psi.	In this image, the needle is pointing to approximately: - 350 psi on the outer pressure scale - 24 bar on the inner pressure scale (between 20 and 25 bar) - 70°C on the temperature scale (the inner green and red circle) The gauge appears to be a combination pressure and temperature gauge for refrigerants, likely used in HVAC or refrigeration systems. The needle is indicating high pressure and temperature readings.
Fig. 14	The needle on the gauge points to approximately 80 psi.	In this image, there are two needles on the gauge: 1. The longer black needle is pointing to approximately 27°C on the temperature scale (the green and red inner circle). 2. The shorter blue needle is pointing to about 8.5 on the outer blue scale, which corresponds to approximately 125 psi on the pressure scale. So to summarize: - Temperature: ~27°C - Pressure: ~8.5 bar or ~125 psi The gauge appears to be a combination pressure and temperature gauge for refrigerants, likely used in HVAC or refrigeration systems.
Fig. 15	The needle on the gauge points to approximately 160 psi.	In this image, the needle on the pressure gauge is pointing to approximately 175 psi (pounds per square inch) on the outer scale. This corresponds to about 1200 kPa (kilopascals) on the inner scale. The needle is positioned between the 150 and 200 psi markings, closer to the 200 psi mark, which indicates a reading of around 175 psi.
Fig. 16	The needle on the gauge points to approximately 60 psi.	The needle on this pressure gauge is pointing to approximately 50 psi (pounds per square inch) or 350 kPa (kilopascals). The gauge has two scales: an outer scale in psi and an inner scale in kPa. The needle is positioned about halfway between 0 and 100 on the psi scale, which corresponds to the 50 psi mark.

Chartered Financial Analyst (CFA), on Scholastic Assessment Test (SAT) and on Multistate Bar Examination (MBE).

Another area that characterizes human intelligence and is an area of challenge for LLM is reasoning. Reasoning is a key component of human intelligence, essential for tasks like problem-solving, decision-making, and critical thinking. Recently, LLMs have shown notable advancements in natural language processing, and it's been observed that they might display reasoning capabilities when they reach a certain size [4].

Translation is also a capability provided by LLMs. It is an important capability in a globalized society that eliminates the problems of understanding different languages. Models can nowadays produce very accurate translations since they have been trained on huge amounts of data [8].

The acoustic capability of models is also a new feature of models and is defined as or the ability to process and produce content based on audio data. This includes the ability to recognize emotion derived from speech and the ability to distinguish between different accents.

For the end we left a new relevant ability that the different LLMs have displayed in the last 2-3 years and it is referred to as visual ability.



Fig. 8. Low pressure manometer.

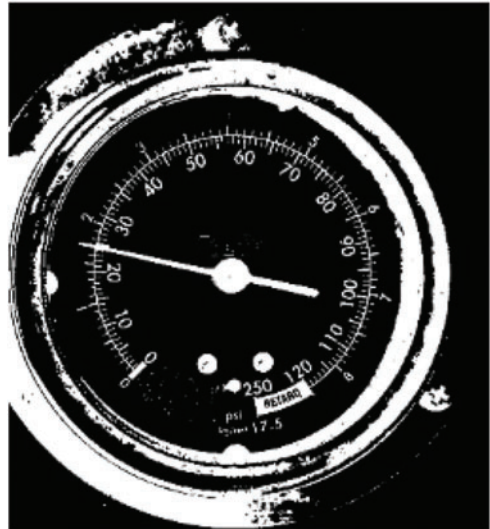


Fig. 11. Low pressure manometer edited photo.



Fig. 9. High pressure manometer.

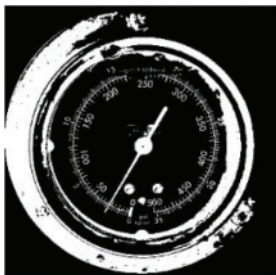


Fig. 10. High pressure manometer edited photo.

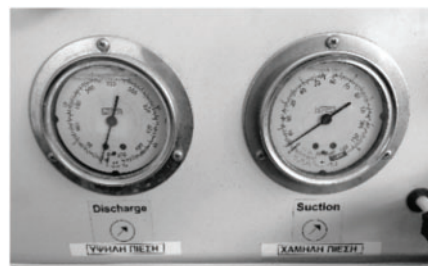


Fig. 12. Discharge and suction manometers together.



Fig. 13. Second high pressure manometer photo.

6. GOALS AND METHODOLOGY

6.1. Goals

This paper focus on the recognition by LLMs of images depicting instruments and display screens. The aim is to be able to investigate what is the success rate of recognition of different controls and displays using prompts alone from two of the most recent and successful LLMs, ChatGPT 4o and Claude 3.5 Sonnet without any pre-processing of the photos or pre-training of the model. Simple zero shot prompts were used without implementing any type of training [9].

6.2. Methodology

The control and handling panels of various machines and mechanisms have a wide variety of different components. The following groups of components are indicative:

- Displays of various types and sizes with display of text or other graphic elements
- Measuring clocks with the use of indicators (analogue manometers, thermometers, etc.)
- Selector switches
- Light indicators



Fig. 14. Second low pressure manometer photo.



Fig. 15. First oil pressure gauge image.



Fig. 16. Second oil pressure gauge.

Knowledge: Prompts designed to draw logical conclusions. Use of encyclopedic and current knowledge (news, news). Purpose is to figure out if LLM possesses specific knowledge or and can perform complex tasks. Example:

- “In what space have we taken the photograph?”
- “What happens if the temperature reading exceeds 120 degrees Celsius?”
- “Convert the pressure that you identify and given in psi to units of measurement of the bar.”

So, we gathered a number of photos which we divided into the following categories:

- Displays showing text and/or graphics
- Gauges
- A combination of objects to be identified including many different objects (panels).

We then posed very specific prompts to the two language models and recorded their responses. This made it possible to statistically analyze them, compare the responses and draw useful conclusions about their reliability.

7. RESULTS

The prompt that was given to the LLM to check Screens was: “Identify the text you see in the photo” as shown in [Table I](#).

The prompt that was given to the two LLMs to check Gauges was: “Where does the needle point to?” as shown in [Table II](#).

The prompt that was given to the LLM to check Panels was: “Write me in detail what do you see and recognize in this photo” as shown in [Table III](#).

8. ANALYSIS

Concerning the first group of images which depict various textual clues on screens, it can be said that the answers to the question “Identify the text you see in the photo” were answered correctly in most cases.

As shown in [Table V](#), ChatGPT 4o showed an overall text and number recognition success rate of 92%, while the

The prompts we could set in LLM can ask questions of different types such as the following. Prompts for simple text recognition (OCR) example:

- “Identify the text you see in the photo”
- “Which numbers appear in the picture?”

Recognition: Prompts to identify the objects depicted. We mean the ability to recognize the objects but their properties (color, shape, material, etc.), the number of objects, etc. Example:

- “What color is the circle in the center of the image?”
- “Which of the two circles is darker?”
- “How many buttons are there in the picture?”

Spatial awareness: Prompts for recognizing the spatial position of objects and the correlation between the position of objects, background and text. Example:

- “What is the temperature T2?”
- “What value does the clock hand point to?”
- “What is to the right of the switch?”
- “What is the highest temperature you can see?”

TABLE III: CATEGORY PANELS

Fig. 17

Component	Real indications	ChatGPT 4o	Claude 3.5
<i>Indicators</i>			
Zundung	Off	Off	Off
Betrieb	On	On	On
Kesselstoerung	Off	On	On
<i>Switches</i>			
Brennstoffwahl	Gas	Gas	–
Betriebswahl	Druckabsenkung	0 (Off)	Center
Brennersteuerung	Betrieb	Aus (Off)	Right
Verbundregler	Auto	Auto	AUT
ECO-Steuerung	Auto	Auto	AUT
Abgasklappe	Auto	Auto	AUT
Betriebsstunden	97066h	97698.8 h	–
Betriebsstunden	14779.07 h	14721.1 h	–
Starts Brenner	2346	66322	–

Fig. 18

Component	Real indications	ChatGPT 4o	Claude 3.5
<i>Indicators</i>			
Ventil Auf	On	Off	On
Ventil Zu	Off	Off	Off
Betrieb Pumpe 1	On	Off	On
Betrieb Pumpe 2	Off	Off	Off
Betrieb	On	On	On
<i>Switches</i>			
Motorventil	Auto	Auto	AUT
Vorwahl	“1”	0 (Off)	“0”
Betriebswahl	Auto	Hand	AUT
Absalzung	Ein	Ein	AUS

Claude 3.5 model showed a slightly lower success rate of 91%.

In the Gauges category the success rates were significantly lower. Table VI summarizes the results on the identification of the needle position in relation to the indications of the circular scale. This is a difficult task and methodologies for increasing the success rates should be further investigated.

We can see the Right and the Wrong Answers in Table IV.

ChatGPT 4o managed to correctly read 40% of the clues while Claude 3.5 managed to correctly read only 20% of the clues.

Finally in the general category of recognizing a panel of instruments and indications we had an average success rate. The ChatGPT 4o had a 70% correct recognition rate while the Claude 3.5 had a 60% rate.

8.1. Comments

The results of this study demonstrate the high accuracy of LLMs in tasks such as Optical Character Recognition (OCR), where ChatGPT 4o achieved a 92% accuracy rate, and Claude 3.5 Sonnet achieved 91%. While this is promising from a technical perspective, these results also hold significant potential for educational applications. For example, in fields like mechanical engineering, where students must read and interpret data from digital displays, these AI models could be integrated into virtual labs or educational tools to assist students in understanding and interpreting real-world data.

Hence the next step is to be able to see if spatial recognition and correlation of the texts we read is also feasible. That is, on a screen where we have the number “23” under “T1” will the LLM be able to understand that $T1 = 23$ points?

Regarding Gauges we notice a particular difficulty in identifying the position of the needle in relation to the scale peripheral to an arc. Success rates were very low regardless of the language model. Here we should look for different ways of approaching this. Perhaps a combination of LLMs and simple programming in Python code using appropriate libraries should be used.

In the control panels category we had a moderate success rate. Here we could segment the whole process into steps. First step could be the identification and numbering of the different elements identified in the panel. Then depending on the element a different path of detailed identification can be followed.

These limitations suggest that while LLMs are highly effective for text-based tasks, further refinements are needed in order to improve their performance in scenarios requiring spatial reasoning or object identification, such as reading complex instrumentation or control panels. Future developments in Vision AI, combined with improved prompt engineering, could help overcome these challenges and create more sophisticated educational tools.



Fig. 17. First part of panel indicators.



Fig. 18. Second part of panel indicators.

9. REAL WORLD APPLICATIONS

Successful image recognition from experimental device displays and components may have positive implications for other real-world applications. As already mentioned the use of this capability in the recognition and reading of laboratory experimental device readings may be utilized in the process of distant learning.

Other application areas are the manufacturing industry for monitoring the production process and prevent downtime, the energy sector for managing power networks and plants with control of systems at remote locations, the monitoring of medical equipment and readings remotely, the engine monitoring in the shipping and aviation industry

In general, remote reading of displays and controls offers the possibility of optimizing the operation of machinery, reducing maintenance costs, preventing breakdowns and increasing efficiency in many critical sectors of the economy and industry [10].

10. FUTURE PROSPECTS

Possible future prospects include a more detailed categorization-grouping of objects from a variety of control panels and display screens that we want to be able to successfully recognize and read as well as the provision of additional data to evolve the model. We may also explore the utilization of custom made LLMs and/or using a combination of results from queries in LLM and custom algorithms. Finally, future work could focus on the utilization of sequencing prompts for successful image reading using appropriate prompt engineering [11].

TABLE IV: RIGHT/WRONG ANSWERS FOR SCREENS CHATGPT

No	Right	Wrong	Total	% (Right)
Fig. 1	39	1	40	97.5%
Fig. 2	46	3	49	93.9%
Fig. 3	46	4	50	92.0%
Fig. 4	21	6	27	77.8%
Fig. 5	6	0	6	100.0%
Fig. 6	13	0	13	100.0%
Fig. 7	10	2	12	83.3%
	181	16	197	92%

TABLE V: RIGHT/WRONG ANSWERS FOR SCREENS CLAUDE 3.5

No	Right	Wrong	Total	% (Right)
Fig. 1	32	8	40	80.0%
Fig. 2	47	2	49	95.9%
Fig. 3	50	0	50	100.0%
Fig. 4	27	0	27	100.0%
Fig. 5	6	0	6	100.0%
Fig. 6	6	7	13	46.2%
Fig. 7	11	1	12	91.7%
	179	18	197	91%

TABLE VI: RIGHT/WRONG ANSWERS FOR GAUGES

No	ChatGPT 4o	Claude 3.5
	Right/Wrong	Right/Wrong
Fig. 1	R	R
Fig. 2	W	W
Fig. 3	W	W
Fig. 4	R	R
Fig. 5	W – W	W – W
Fig. 6	W	W
Fig. 7	R	W
Fig. 8	W	W
Fig. 9	R	W
	4/10 (40 %)	2/10 (20 %)

11. CONCLUSIONS

The future of Vision AI in LLMs like ChatGPT and Claude is promising, with ongoing advancements enhancing their capabilities significantly. The integration of visual understanding into these models allows for more complex and contextually rich interactions, making them more versatile and effective across various applications.

One notable development is the combination of vision and language models, exemplified by OpenAI’s GPT-4, which incorporates visual inputs to improve understanding and reasoning abilities. This multimodal approach enables the model to perform tasks that require both text and image comprehension, such as identifying objects within images, interpreting visual scenes, and providing detailed descriptions.

Additionally, advancements in pre-training techniques, such as those used in MiniGPT-4, show how aligning visual information with language models can enhance performance on vision-language tasks. This involves using large datasets of image-text pairs to train models, allowing them to generate coherent and contextually accurate responses based on visual inputs [12].

The research community is also exploring new methods to improve the alignment and instruction-following capabilities of these models in the multimodal space, which will further enhance their utility and applicability in real-world scenarios [13].

Hence, Vision AI, holds great promise for both technical applications and education. This study demonstrates the potential of these models to accurately recognize and interpret diverse display and gauge images in a variety of settings, with potential implications for industries ranging from manufacturing to healthcare. However, the educational applications are equally significant. Vision AI could have a transformative effect in STEM education by providing interactive learning environments, improving accessibility for students with disabilities, and automating the evaluation of technical skills.

The continuous development of the above technologies allows for more effective and inclusive educational tools, to engage students with complex visual materials in ways that were previously impossible. The integration of Vision AI into educational settings has the potential to transform the learning process and improve interaction with visual data, making education more dynamic, personalized, and accessible for all.

FUNDING

The research was funded by the Special Account for Research Funds of the University of West Attica (E.L.K.E. Pa.D.A.).

CONFLICT OF INTEREST

The authors declare that they do not have any conflict of interest.

REFERENCES

- [1] Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, *et al.* ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Differ.* 2023 Apr 1;103(1):102274. doi: 10.1016/j.lindif.2023.102274.
- [2] Amazon Web Services, Inc What are large language models?—LLM AI explained—AWS. [cited 2024 Jul 19]. Available from: <https://aws.amazon.com/what-is/large-language-model/>.
- [3] Jabeen S, Li X, Amin MS, Bourahla O, Li S, Jabbar A. A review on methods and applications in multimodal deep learning. *ACM Trans Multimedia Comput Commun Appl.* 2023 Oct 17;19(2s):76:1–41. doi: 10.1145/3545572.
- [4] Wu J, Gan W, Chen Z, Shicheng W, Yu P. Multimodal large language models: a survey. 2023. Available from: <https://arxiv.org/abs/2311.13165>.
- [5] Canales Luna J. 8 Top open-source LLMs for 2024 and their uses [Internet]. 2024 Aug 8 [cited 2024 Jul 25]. Available from: <https://www.datacamp.com/blog/top-open-source-llms>.
- [6] Zakari RY, Owusu JW, Wang H, Qin K, Lawal ZK, Dong Y. VQA and visual reasoning: an overview of recent datasets, methods and challenges [Internet]. arXiv. 2022 [cited 2024 Jul 25]. Available from: <http://arxiv.org/abs/2212.13296>.
- [7] Wu C, Yin S, Qi W, Wang X, Tang Z, Duan N. Visual ChatGPT: talking, drawing and editing with visual foundation models [Internet]. arXiv. 2023 [cited 2024 Jul 25]. Available from: <http://arxiv.org/abs/2303.04671>.
- [8] Khoshafah F. ChatGPT for arabic-english translation: evaluating the accuracy. 2023. Available from: <https://www.researchsquare.com/article/rs-2814154/v1>.

- [9] Huang J, Chang KCC. Towards reasoning in large language models: a survey [Internet]. arXiv. 2022 [cited 2024 Jul 29]. Available from: <http://arxiv.org/abs/2212.10403>.
- [10] Madani K. Artificial neural networks based image processing & pattern recognition: from concepts to real-world applications. *2008 First Workshops on Image Processing Theory, Tools and Applications [Internet]*. pp. 1–9. 2008 [cited 2024 Oct 9]. Available from: <https://ieeexplore.ieee.org/abstract/document/4743797>.
- [11] Karwa S. Exploring multimodal large language models: a step forward in AI. Medium. 2023 [cited 2024 Jul 24]. Available from: <https://medium.com/@cout.shubham/exploring-multimodal-large-language-models-a-step-forward-in-ai-626918c6a3ec>.
- [12] Zhu D, Chen J, Shen X, Li X, Elhoseiny M. MiniGPT-4: enhancing vision-language understanding with advanced large language models [Internet]. arXiv; 2023 [cited 2024 Jul 24]. Available from: <https://arxiv.org/abs/2304.10592>.
- [13] Wang X, Chen G, Qian G, Gao P, Wei XY, Wang Y, *et al.* Large-scale multi-modal pre-trained models: a comprehensive survey. *Mach Intell Res.* 2024 Aug;20(4):447–82. doi: 10.1007/s11633-022-1410-8.