

Article

Not peer-reviewed version

Can ChatGPT Solve Undergraduate Exams from Logistics Studies? An Investigation

[Sven Franke](#)*, [Christoph Pott](#), Jérôme Rutinowski, [Markus Pauly](#), [Christopher Reining](#), Alice Kirchheim

Posted Date: 11 January 2025

doi: 10.20944/preprints202501.0843.v1

Keywords: GPT-4o; GPT-4o mini; o1-preview; Educational assessment; Warehousing; LLM



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Can ChatGPT Solve Undergraduate Exams from Logistics Studies? An Investigation

Sven Franke ^{1,*} , Christoph Pott ¹ , Jérôme Rutinowski ¹ , Markus Pauly ^{2,3} ,
Christopher Reining ^{1,4}  and Alice Kirchheim ^{1,4} 

¹ TU Dortmund University, Department of Mechanical Engineering, Germany; christoph.pott@tu-dortmund.de (C.P.); jerome.rutinowski@tu-dortmund.de (J.R.); christopher.reining@tu-dortmund.de (C.R.); alice.kirchheim@tu-dortmund.de (A.K.)

² TU Dortmund University, Department of Statistics, Germany; markus.pauly@tu-dortmund.de

³ Research Center Trustworthy Data Science and Security of the University Alliance Ruhr, Germany

⁴ Fraunhofer Institute for Material Flow and Logistics IML, Germany

* Correspondence: sven.franke@tu-dortmund.de; Tel.: +49-231-755-4863

Abstract: The performance of Large Language Models, such as ChatGPT, generally increases with each further model release. In this paper, we investigate whether and how well different ChatGPT models solve the exams of three different logistics undergraduate courses. We contribute to the discussion of ChatGPT's existing logistics knowledge, particularly in the field of warehousing. Both, the free version (GPT-4o mini) and the chargeable version (GPT-4o) completed three different logistics exams using three different prompting techniques (with and without role assignment as logistics expert or student). The o1-preview model was also used (without role assignment) for six attempts. The tests were repeated three times. A total of 60 tests were completed and compared to in-class results of logistics students. The results show that a total of 46 tests were passed. The best attempt solved 93% of an exam correctly. Compared to the students from the respective semester, ChatGPT outperforms students in one exam. On the other two exams the students perform better on average.

Keywords: GPT-4o; GPT-4o mini; o1-preview; Educational assessment; Warehousing; LLM

1. Introduction

In 2023 OpenAI's ChatGPT broke the turing test [1]. The development of Large Language Models (LLMs) is rapid [2] and an LLMs model performance generally improves with each new version. The community also tests and evaluates new models using different methods. In this regard, OpenAI's models are in top positions [3,4]. They use, for example, a graduate-level Google Proof Q&A Benchmark with over 400 science questions, reading comprehension questions and professional translation tests. Even though many consider the LLMs performance tests to be highly informative, there is criticism that the tests can be cheated with the right training data [5]. Moreover, it is well known that LLMs tend to hallucinate, have a bias or even produce false results c.f., [6–8]. Nevertheless, LLMs like ChatGPT are used for a wide range of applications [9–11].

However, despite their widespread use in various studies, there is little research in the field of logistics, particularly in warehousing. In logistics, there are many use cases for decision support or even automation using AI [12–14]. Yet, little is known about how much ChatGPT knows or can answer about warehousing. This contribution tests ChatGPT's knowledge of warehousing using three different exams from a german University's undergraduate logistics degree. The answers of the LLM are evaluated as if it was a student and they are compared with the results of the actual written exam of the winter semester 2022/2023 at TU Dortmund University. We use the exams Warehouse Management Systems (WMS), Material Flow Systems I (MFS I) and Material Flow Systems II (MFS II), which cover the basics of warehousing and other logistics-related topics. In September 2024, OpenAI published the o1-preview and o1-mini models. At the launch of the new models, only 30 and 50 prompts per

week were possible, respectively [15]. Due to the limitations of the latest models, the GPT-4o mini and GPT-4o models are used for the study in addition to the o1-preview. At the time of the investigation, the GPT-4o mini model is free of charge and the GPT-4o model can also be used free of charge for a dynamic number of prompts. After that, it can only be used with a paid subscription. This work contributes to understanding ChatGPT's capability to replicate logistics and warehousing knowledge. In addition, it becomes known how well different types of exam questions are handled and where limitations are, either in the way of answering qualitative questions or in solving math problems.

2. Related Work

Over four generations, language models developed from specific task helpers in the 1990s to general-purpose task solvers in the 2020s [16]. During this evolution process, they continuously improved their task solving capacity. LLMs represent the latest generation of language models. As a category of foundation models, LLMs are designed to understand and generate natural (human) language [17]. To achieve this, they assign a probability to a sequence of words [16,18]. This prediction process is based on LLM training with ever larger data sets and was accelerated by increased computational capabilities and technological breakthroughs, such as transformers [19–21]. It enables LLMs such as Open AI's ChatGPT to answer questions with human-like capabilities [22,23]. More recent LLMs such as GPT-4, however, have undergone a paradigm shift and focus not only on generating text data, but on solving complex, diverse tasks which is considered a step towards artificial general intelligence [16,24,25]. Hence, they have aroused interest in both industry and academia [24,26].

Latest with the launch of ChatGPT in 2022, a discourse on the impact of LLMs on education arose [27–30]. Significant branches of the discussion are how LLMs can support instructors in their teaching, how they can support student learning, and issues related to LLMs in education [27–29,31–34]. Along with this discussion, many scholars since have risen the obvious question if LLMs are capable of passing exams at university level.

As for ChatGPT-3, results were very different depending on the domain. ChatGPT-3.5 performed outstandingly in the domains of economics [35] and critical and higher-order thinking [36]; outstandingly to satisfactorily in programming [37,38]; satisfactorily in English language comprehension [39]; barely satisfactorily to unsatisfactorily in law [40,41] and medical education [42–46]; unsatisfactorily in mathematics [47], software testing [48], and multiple choice based exams across subjects [49].

In comparison, testing has shown that GPT-4 can solve novel and difficult tasks from diverse fields of expertise, covering natural sciences, life sciences, engineering or social sciences [25]. In the field of medical education GPT-4 achieved clearly better results than GPT-3 [50–54]. The same is true for higher education programming and university-level coding courses [55,56], mechanical engineering education [57] or law education [58].

3. Materials and Methods

The ChatGPT web interface is used for all tests. The models used are GPT-4o mini, GPT-4o and o1-preview. All tests are performed in November 2024. At that time, no API (Application Programming Interface) could be used for the o1-preview model, so all questions are prompted to the web interface manually. For a better comparability and exam performance evaluation, all tests are repeated a total of three times. The WMS, MFS I and MFS II exams from the 2022/2023 winter semester represent the use case. The first step is to translate the respective exams into English. After the translation, all exam questions are asked to the respective model. Three different roles or prompting techniques are used for GPT-4o mini and GPT-4o. Role assignment has an influence on how ChatGPT behaves when answering questions [11,59]. Therefore, in the first test, we are investigating by asking the questions without an initial prompt. In the second test, the model is asked to put itself in the role of a logistics expert (Prompt LE). Finally, the model is asked to take on the role of a logistics student (Prompt LS). These prompts are used at the beginning of the session:

LE Please put yourself in the role of a logistics expert and answer all the questions below.

LS Please put yourself in the role of a logistics student and answer all the questions below.

This procedure is repeated two more times. A total of 60 attempts are carried out using this procedure. 21 times WMS and MFS II each and 18 times MFS I. Since the MFS I exam includes image interpretation in the exam, the o1-preview model cannot be used because it does not allow uploading images at the time of testing. The o1-preview model is therefore only used for WMS and MFS II without prompting due to its limited usability. The investigations using the o1-preview model as a logistics student and logistics expert are not conducted. Finally, all exams are evaluated using the sample solution of the exams. The grading system for the exams can be found in the Appendix in Table A1. In Germany a grading system from 1.0 (best) to 5.0 (worst) is used, which has been adapted to the international form from A (best) to F (worst) for clarity. Both systems use equivalent grade increments. The overview of the procedure for the tests and findings is summarized in Figure 1. All answers of the respective runs are publicly accessible in the protocol under [60].

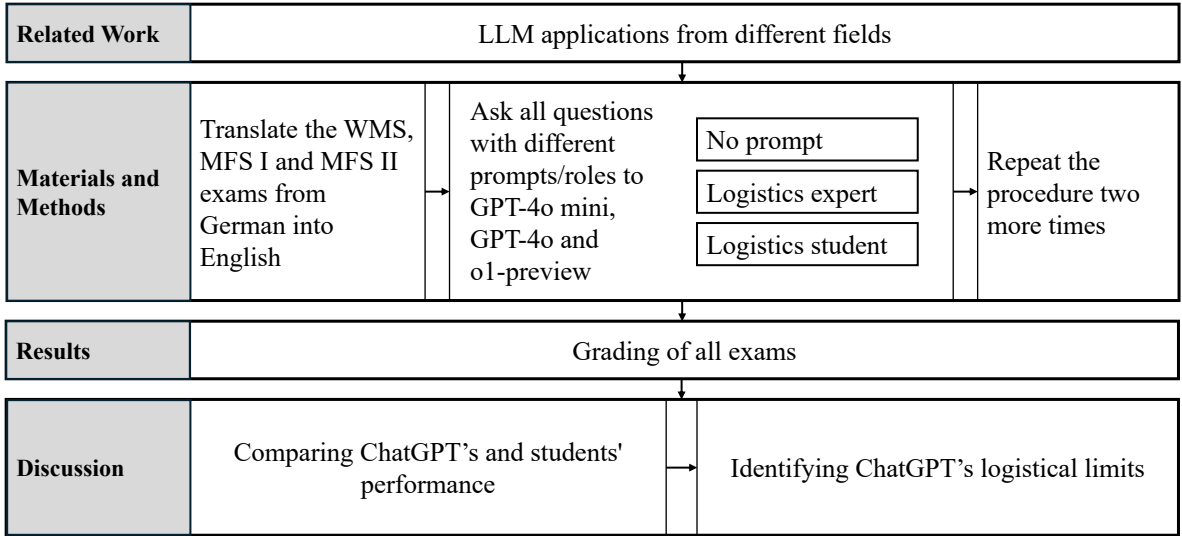


Figure 1. Overview of the process and paper structure.

According to the module description, the lecture WMS provides basic knowledge on the practical use of information technology (IT) in logistics [61]. The content deals with the computer-aided management and monitoring of logistics processes in a warehouse. Accompanying this, logistical data processing is covered with a focus on methods for evaluating, preparing and presenting company data using standard programs. The WMS exam consists of 32 questions, which are divided into eight different topic blocks. The questions include 24 single and eight multiple choice questions. In terms of content, knowledge questions, logic questions, programming questions and simple math questions are asked. The points awarded per question range from one to four points. When translating them, two questions from block seven of the WMS exam had to be adapted. Question 7.1 uses specific abbreviations of German words, which were adapted to the English abbreviations. Question 7.3 asks for an interpretation of English terms which have a different connotation in German (so called False Friends). In the adapted question, the context for the different interpretations in German has been added.

The MFS I course covers the equipment and systems required for internal logistics in materials handling technology. Students learn about the systematic classification of devices, their structure and their essential characteristics as well as their application criteria. In addition to the interaction of the material flow, they learn which standards, guidelines and laws are important for the operation of these devices and systems. The MFS I exam consists of a total of 17 questions, divided into five content blocks. Block one contains two single choice and 11 multiple choice questions. Content from all lectures is tested here. Block two contains a free text question on the topic of order picker guidance systems. The last three blocks are math problems with several sub tasks [61].

In the MFS II lecture, students learn the methods, procedures and instruments required for planning and operating warehousing systems. The aim of the course is to plan and optimize material flow systems, to design and use the necessary IT and to create the organizational processes and structures. Attending the MFS I lecture in advance is recommended but not mandatory. Students can choose their subjects independently and therefore attend MFS II before MFS I or, depending on their studies, choose only one of the two. The MFS II exam also consists of 17 questions, divided into four content blocks. Block one contains questions on storage systems, block two on the lecture fundamentals, block three on order picking and block four contains mixed questions. The question types consist of single choice, multiple choice, free text questions and two math problems [61]. For all three exams a total of 60 points are awarded. For student participants, the duration of each exam is 60 minutes.

4. Results

From a total of 60 attempts, ChatGPT passed 46. All 21 attempts of the WMS exams were passed. The grades range from A- to C. In the MFS I exam, 17 out of 18 attempts were passed. Here, only D+ and D grades were assigned to the attempts that were passed. Of the 21 attempts in MFS II, eight were also passed. One attempt was passed with a C-. The remaining grades are D+ and D. An overview of the distribution of grades across the exams is shown in Figure 2.

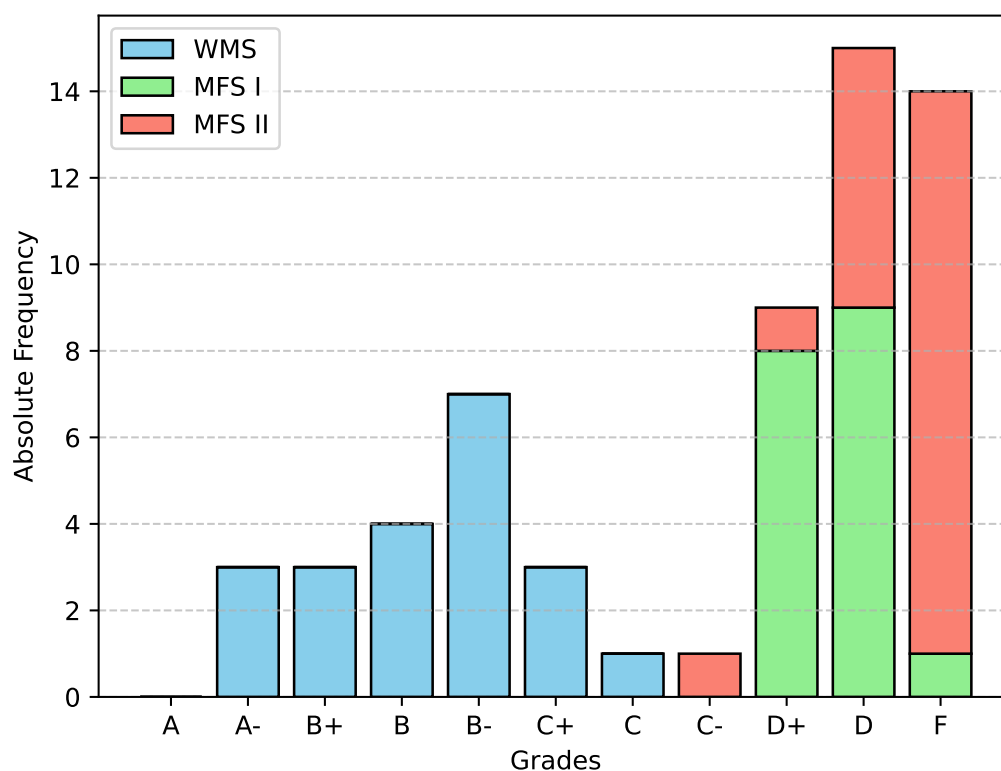


Figure 2. Stacked bar chart showing the distribution of grades across all runs. The different exams are highlighted in different colours: WMS in blue, MFS I in green and MFS II in red.

Looking at the differences between the models, GPT-4o mini passes 66.7% of the 27 attempts made to the model. GPT-4o passes 81.5% (22) of the 27 attempts. All six attempts with the o1-preview model are successful. Within the different prompts, the approach without prompt has the highest success rate with 83.3% of attempts passing. The passed attempts as a logistics expert are 77.8% and as a logistics student 66.7% (14 and 12 passed attempts out of 18 attempts respectively). It should be noted that the

attempts of the o1-preview model are included in the approach without prompting. Without these attempts, the success rate is also 77.8% (14 out of 18 attempts). Table 1 shows the summary of all exam attempts with the median and range of the points. A detailed summary of the results of the three individual runs can be found in the Appendix under Tables A2–A4.

Table 1. Summary measures of the exam results of the three different exams (WMS, MFS I, MFS II), GPT models (4o mini, 4o and o1-preview) and prompts (NP, LE, LS). The percentage and grade columns display the results corresponding to the median exam score shown in the third column.

Version	Prompt ¹	Median	Passed	Percentage	Grade	Points Range
WMS						
GPT-4o mini	NP	42	Yes	70%	B-	4
GPT-4o mini	LE	41	Yes	68%	C+	3
GPT-4o mini	LS	42	Yes	70%	B-	1
GPT-4o	NP	49	Yes	82%	B+	6
GPT-4o	LE	44	Yes	73%	B-	5
GPT-4o	LS	48	Yes	80%	B	2
o1-preview	NP	56	Yes	93%	A-	5
MFS I						
GPT-4o mini	NP	25.5	Yes	43%	D	2
GPT-4o mini	LE	24.5	Yes	41%	D	1.5
GPT-4o mini	LS	26.5	Yes	44%	D	6
GPT-4o	NP	29	Yes	48%	D+	1
GPT-4o	LE	29	Yes	48%	D+	3
GPT-4o	LS	28.5	Yes	48%	D+	4.5
o1-preview	NP	Model can't analyze pictures or answer all questions.				
MFS II						
GPT-4o mini	NP	18.5	No	31%	F	3.5
GPT-4o mini	LE	19	No	32%	F	10.5
GPT-4o mini	LS	19.5	No	33%	F	1
GPT-4o	NP	24.5	Yes	41%	D	2
GPT-4o	LE	21	No	36%	F	7.5
GPT-4o	LS	22	No	37%	F	3
o1-preview	NP	25.5	Yes	43%	D	9

with NP for No Prompt, LE for Logistics Expert and LS for Logistics Student.

5. Discussion

Summary: Comparing the results of ChatGPT to the results of the students from the winter semester 2022/2023, ChatGPT is superior in the WMS exam. For the other two exams, the students performed better. The concrete results are discussed below.

WMS: A total of six students participate in the WMS exam in the winter semester 2022/2023, all of whom passed the exam. The low number of participants is due to the fact that the associated lecture is regularly offered in the summer semester. Students still have the opportunity to take each exam once per semester. Participation in the off-semester exam varies greatly. In comparison, 40 students write the exam in the summer semester 2022. 38 pass the exam. The grades A and A- are each awarded once. The remaining 36 grades range from B to D (one B, six B-, eight C+, four C, six C-, seven D+ and four D). The students' average grade is C. The distribution of the six winter semester grades is as follows: one B+, and the remaining five grades are evenly distributed between C+ to D. Thus, the students achieve an average grade of C. In comparison, ChatGPT performs better with median grades between A- and C+ across all models and prompts. On average, the median grade is B. The particularly good performance in the exam can be attributed to both, the question types and the content. Single choice and multiple choice questions leave little room for interpretation and can be answered clearly. As

the content of the exam not only asks for specific logistics knowledge but also logic and IT, ChatGPT probably can benefit from its logical reasoning abilities [62].

MFS I: The MFS I exam is taken by 74 students in the winter semester. Of these, 69 pass the exam. The grades are distributed as follows: seven A-, four B+, eight B, ten B-, nine C+, eight C, ten C-, eight D+, five D and five F. In comparison, ChatGPT performs below the students average in the MFS I exam with median grades between D+ and D. In MFS I, most single choice and multiple choice questions are also answered correctly. The free text task is correctly solved in every attempt. The three math problems, worth a total of 31 points, have a high influence on the overall score. ChatGPT provides some plausible solutions here, often earning partial points for correctly handling subsequent errors or the right reasoning when interpreting results. However, as tasks become more complicated, the number of correct answers decreases. In calculation task 5.1, all ChatGPT models scored zero points across all runs. This task relies on logistical principles that require accurate knowledge from literature to correctly apply the calculation method.

MFS II: A total of 37 students participate in the MFS II exam in the winter semester 2022/2023, 29 of whom pass the exam. This means that 78.38% pass and 21.62% (eight participants) fail. The grade distribution is as follows: one B-, six D+, six D, seven C+, seven C, two C- and eight F. Students also perform better than ChatGPT on the MFS II exam. Here, its median grade is between D (two times) and F (five times). Across the three different runs, large point variations are observed for the two LE prompts and the o1-preview model. The difficulty of this exam lies not only in the math problems, worth 27 points, but also the single choice and multiple choice questions. These include an additional rule: a correct combination earns two points, while a wrong combination gives zero points. The lack of partial points and the inclusion of more specific logistics questions compared to WMS and MFS I exams results in a higher difficulty level. In comparison, the ChatGPT models often complete the free-text tasks with a high to full score. As with MFS I, partial points can be awarded for two calculation tasks. These are often incorrect as the ChatGPT models did not apply the correct procedures or formulas.

5.1. Chatgpt's Logistical Limits

General: ChatGPT's performance on the exams highlights its limitations in a logistical context. The calculation tasks show that even simple dimensioning of storage techniques cannot be solved. Many of the calculation principles are based on basic logistics literature and/or standards c.f., [63,64]. ChatGPT does not apply these calculation rules. In many cases, partial points are awarded in the exam, as ChatGPT provides a calculation path allowing partial points because of subsequent errors. However, there are also calculations where only the final result is given without explanations. As these are almost always incorrect, zero points are awarded here. There are also cases where only the final result is answered correctly but without explanation (see MFS II run 2, Question 1.3, o1-preview). ChatGPT's answer in part b) is 16 compartments, without any further explanation. In the example, the answer is correct, but it is not possible to determine whether a correct calculation has taken place or the guess is correct. It can also be observed that during prompting the math problems are not always answered to the end. In a few cases, the answer has to be regenerated until all partial answers are given.

Images: The results also show that the GPT-4o mini and GPT-4o models have problems with the interpretation of images. In the MFS I exam, two tasks use a picture to ask which warehouse technology is involved. Task 1.3 is answered incorrectly in all runs. Task 1.9, which asks for the exact name of the technology and the corresponding categorization, is always answered incorrectly by the GPT-4o model. It is noteworthy that the weaker model GPT-4o mini recognizes the categorization correctly in all cases and therefore receives partial points.

Contradictions: In the WMS exam, ChatGPT sometimes contradicts itself within its own responses. An example is given in Figure 3 which displays GPT-4o's answer to Question 6.2 of the WMS exam under the LS prompt. The same contradictory answer is given by GPT-4o mini.

A fair die, i.e. all numbers are equally likely, is thrown with the numbers 1, 2, 3, 4, 5 and 6.
Which event has the highest probability? [Single Choice] (2/60)

- The number rolled is 1.
- The number of dice rolled is odd.
- The number rolled is greater than 3.
- The number of dice rolled is at most equal to 4.



The correct answer is:

c. The number rolled is greater than 3.

Explanation:

- The probability of rolling a number greater than 3 (i.e., 4, 5, or 6) is $\frac{3}{6} = 0.5$.
- The probability of rolling a specific number (e.g., 1) is $\frac{1}{6}$.
- The probability of rolling an odd number (1, 3, or 5) is also $\frac{3}{6} = 0.5$.
- The probability of rolling a number at most equal to 4 (i.e., 1, 2, 3, or 4) is $\frac{4}{6} = 0.67$.

Therefore, option **d** ("The number of dice rolled is at most equal to 4") actually has the highest probability, not option **c**



Figure 3. Question 6.2 of the WMS exam and GPT-4o's answer as a logistics student (prompt LS).

The model answers option c) first. After explaining the individual possibilities, it comes to the conclusion that answer d) is correct. As answer c) was the first to be given in bold, it was also scored and ChatGPT therefore receives zero points. The correct answer is d).

Model differences and similarities: As can already be seen from the examples (no explanation on answers, missing image interpretation and Figure 3), we observe discrepancies in all three models (for example, Question 2.2 in the MFS II exam is only answered correctly by the o1-preview model). From the results in Table 1, it can be seen that the scores achieved increase from GPT-4o mini to GPT-4o and o1-preview. The GPT-4o mini model performed worst in all three exams and the o1-preview achieved the best grades. However, there are also some questions that cannot be answered by any of the models. An example is Question 1.1 in the WMS exam:

Which cross-docking process has picking as a key feature? [Single Choice] (1/60)

- Cross-docking with collection in the storage system
- Cross-docking as a throughput system
- Cross-docking with breaking up the load units
- Cross-docking with pick families (clustering)

The correct answer is c). Relatively simple questions, which are based on logic-based or generally valid assumptions, are always answered correctly by all models. An example is Question 4.5 in the MFS II exam:

Give two examples of the risk of industrial trucks tipping over. [2 pt]

Correct answers are among others driving too fast on bends, driving with a raised load, driving against obstacles. As long as the answer shows a plausible reason, it will be considered correct. There are few differences within the prompts (NP, LE, LS). From the median values in Table 1, it can be seen that the LE prompt achieves the worst results, except for MFS I model GPT-4o. From the given answers, it is

often only minor details that differ. Only once does ChatGPT refer literally to the initial prompt. In the MFS II exam, run 3, GPT-4o starts the answer with the sentence: As a logistics student, I interpret the designation 07-12-2 for a storage location as a structured label that provides essential information about the pallet's position in the warehouse.

Clear differences in response length are difficult to distinguish. In most cases, the responses within the prompts are approximately the same length. In a few cases, the lengths of the responses differ significantly. However, here the prompts alternate and there is not one prompt that always results in the longest responses.

5.2. Limitations and Future Work

The work has limitations in that only three different exams from one semester have been answered from ChatGPT. In addition, these were evaluated using a sample solution, but by one reviewer (the first author) only. The tested logistics knowledge cannot provide a conclusive assessment of the training data from ChatGPT, but it does show the limitations of the models for simple logistical questions. It also makes sense to test the latest models, the o1-preview and o1-mini, in more detail, which was only possible to a limited extent at the time of testing. In the future, it would be important to test for further logistics knowledge related to warehousing beyond exams. Tests should also not be limited to ChatGPT models.

We believe that LLMs can become valuable tools for supporting planning, data analysis or even to completely automate specific process steps. However, this requires precise and weighted data with which the models are trained. In logistics and warehouse planning, a great amount of experience-based knowledge is used. LLMs must recognize which data is given and specifically ask for parameters that have the greatest influence on the problem. Only when these relationships have been learned can an LLM replace planning tasks. To test this, it is useful to ask both consultants and various LLMs specific questions from logistics planning in the future and compare the results. The need for better models is also recognized by OpenAI itself and a Reinforcement Fine-Tuning Research Program was recently launched [65].

6. Conclusions

In summary, the question of whether ChatGPT can pass undergraduate exams from logistics is answered positively. The WMS exam is solved above average compared to students, while the MFS I-II exams are completed with noticeable weaknesses. To obtain correct answers, the question type is crucial. Single choice, multiple choice and free text questions can often be solved reliably. Calculations that use formulas from the relevant literature, however, cannot be solved well. Moreover, when the questions become highly domain-specific, ChatGPT quickly reaches its limits. The role assigned to the model influences its performance: overall, the answers as a logistics student are the ones with the worst results. Prompting as a logistics expert or without role assignment lead to better outcomes. To sum up, a basic knowledge of warehousing is present in the ChatGPT models, but there is still a lot of potential for improvement. A consequence for future exams is that questions should be designed to test knowledge that can only be solved by humans. This requires follow-up studies to determine the areas in which humans will remain indispensable in logistics. In addition, more LLM applications should be developed for logistics planning. Furthermore, it is necessary that images with logistical content can be better interpreted.

Author Contributions: Conceptualization, S.F. and M.P.; methodology, S.F.; investigation, S.F.; data curation, S.F. and M.P.; writing—original draft preparation, S.F. and C.P.; writing—review and editing, C.P. and C.R. and M.P. and A.K.; visualization, J.R.; supervision, C.R.; project administration, S.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All answers of the respective runs from ChatGPT are publicly accessible in the protocol under [60].

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

API	Application Programming Interface
IT	Information Technology
LE	Logistics Expert
LLM	Large Language Model
LS	Logistics Student
MFS I	Material Flow Systems I
MFS II	Material Flow Systems II
NP	No Prompt
WMS	Warehouse Management Systems

Appendix A

Table A1. Evaluation scheme for all exams.

Points in percent	Grade	German grading system
94% - 100%	A	1.0
88% - 93.99%	A-	1.3
82% - 87.99%	B+	1.7
76% - 81.99%	B	2.0
70% - 75.99%	B-	2.3
64% - 69.99%	C+	2.7
58% - 63.99%	C	3.0
52% - 57.99%	C-	3.3
46% - 51.99%	D+	3.7
40% - 45.99%	D	4.0
0% - 39.99%	F	5.0

Table A2. Run 1 of the exam attempts.

Version	Prompt ¹	Total Points	Passed	Percentage	Grade
WMS					
GPT-4o mini	NP	42	Yes	70%	B-
GPT-4o mini	LE	41	Yes	68%	C+
GPT-4o mini	LS	42	Yes	70%	B-
GPT-4o	NP	47	Yes	78%	B
GPT-4o	LE	44	Yes	73%	B-
GPT-4o	LS	48	Yes	80%	B
o1-preview	NP	56	Yes	93%	A-
MFS I					
GPT-4o mini	NP	25.5	Yes	43%	D
GPT-4o mini	LE	25.5	Yes	43%	D
GPT-4o mini	LS	22.5	No	38%	F
GPT-4o	NP	30	Yes	50%	D+
GPT-4o	LE	30	Yes	50%	D+
GPT-4o	LS	28.5	Yes	48%	D+
o1-preview	NP	Model can't analyze pictures or answer all questions.			
MFS II					
GPT-4o mini	NP	22	No	37%	F
GPT-4o mini	LE	26.5	Yes	44%	D
GPT-4o mini	LS	18.5	No	31%	F
GPT-4o	NP	25	Yes	42%	D
GPT-4o	LE	28	Yes	47%	D+
GPT-4o	LS	21	No	35%	F
o1-preview	NP	25	Yes	42%	D

¹ with NP for No Prompt, LE for Logistics Expert and LS for Logistics Student.

Table A3. Run 2 of the exam attempts.

Version	Prompt ¹	Total Points	Passed	Percentage	Grade
WMS					
GPT-4o mini	NP	42	Yes	70%	B-
GPT-4o mini	LE	43	Yes	72%	B-
GPT-4o mini	LS	42	Yes	70%	B-
GPT-4o	NP	49	Yes	82%	B+
GPT-4o	LE	48	Yes	80%	B
GPT-4o	LS	49	Yes	82%	B+
o1-preview	NP	56	Yes	93%	A-
MFS I					
GPT-4o mini	NP	26.5	Yes	44%	D
GPT-4o mini	LE	24.5	Yes	41%	D
GPT-4o mini	LS	28.5	Yes	48%	D+
GPT-4o	NP	29	Yes	48%	D+
GPT-4o	LE	27	Yes	45%	D
GPT-4o	LS	24.5	Yes	41%	D
o1-preview	NP	Model can't analyze pictures or answer all questions.			
MFS II					
GPT-4o mini	NP	18.5	No	31%	F
GPT-4o mini	LE	16	No	27%	F
GPT-4o mini	LS	19.5	No	33%	F
GPT-4o	NP	23	No	38%	F
GPT-4o	LE	21.5	No	36%	F
GPT-4o	LS	24	Yes	40%	D
o1-preview	NP	34	Yes	57%	C-

¹ with NP for No Prompt, LE for Logistics Expert and LS for Logistics Student.

Table A4. Run 3 of the exam attempts.

Version	Prompt ¹	Total Points	Passed	Percentage	Grade
WMS					
GPT-4o mini	NP	38	Yes	63%	C
GPT-4o mini	LE	40	Yes	67%	C+
GPT-4o mini	LS	41	Yes	68%	C+
GPT-4o	NP	53	Yes	88%	A-
GPT-4o	LE	43	Yes	72%	B-
GPT-4o	LS	47	Yes	78%	B
o1-preview	NP	51	Yes	85%	B+
MFS I					
GPT-4o mini	NP	24.5	Yes	41%	D
GPT-4o mini	LE	24	Yes	40%	D
GPT-4o mini	LS	26.5	Yes	44%	D
GPT-4o	NP	29	Yes	48%	D+
GPT-4o	LE	29	Yes	48%	D+
GPT-4o	LS	29	Yes	48%	D+
o1-preview	NP	Model can't analyze pictures or answer all questions.			
MFS II					
GPT-4o mini	NP	18.5	No	31%	F
GPT-4o mini	LE	19	No	32%	F
GPT-4o mini	LS	19.5	No	33%	F
GPT-4o	NP	24.5	Yes	41%	D
GPT-4o	LE	20.5	No	34%	F
GPT-4o	LS	22	No	37%	F
o1-preview	NP	25.5	Yes	43%	D

¹ with NP for No Prompt, LE for Logistics Expert and LS for Logistics Student.

References

1. Bieber, C. ChatGPT broke the Turing test-the race is on for new ways to assess AI. *Nature* **2023**, 619, 686–689.
2. Wu, C.; Tang, R. Performance Law of Large Language Models. *arXiv preprint arXiv:2408.09895* **2024**.
3. Chatbot Arena LLM Leaderboard: Community-driven Evaluation for Best LLM and AI chatbots. <https://lmarena.ai/?leaderboard>. Visited on 2024-12-05.
4. GitHub OpenAI Repository Simple-evals. <https://github.com/openai/simple-evals?tab=readme-ov-file#benchmark-results>. Visited on 2024-12-05.
5. Zhou, K.; Zhu, Y.; Chen, Z.; Chen, W.; Zhao, W.X.; Chen, X.; Lin, Y.; Wen, J.R.; Han, J. Don't make your llm an evaluation benchmark cheater. *arXiv preprint arXiv:2311.01964* **2023**.
6. Rutinowski, J.; Franke, S.; Endendyk, J.; Dormuth, I.; Roidl, M.; Pauly, M. The Self-Perception and Political Biases of ChatGPT. *Human Behavior and Emerging Technologies* **2024**, 2024, 7115633.
7. Snyder, B.; Moisescu, M.; Zafar, M.B. On early detection of hallucinations in factual question answering. In Proceedings of the Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2024, pp. 2721–2732.
8. Ji, Z.; Yu, T.; Xu, Y.; Lee, N.; Ishii, E.; Fung, P. Towards mitigating LLM hallucination via self reflection. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, pp. 1827–1843.
9. Palen-Michel, C.; Wang, R.; Zhang, Y.; Yu, D.; Xu, C.; Wu, Z. Investigating LLM Applications in E-Commerce. *arXiv preprint arXiv:2408.12779* **2024**.
10. Huang, Y.; Gomaa, A.; Semrau, S.; Haderlein, M.; Lettmaier, S.; Weissmann, T.; Grigo, J.; Tkhayat, H.B.; Frey, B.; Gaipf, U.; et al. Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: Potentials and challenges for ai-assisted medical education and decision making in radiation oncology. *Frontiers in Oncology* **2023**, 13, 1265024.
11. Weber, E.; Rutinowski, J.; Pauly, M. Behind the Screen: Investigating ChatGPT's Dark Personality Traits and Conspiracy Beliefs. *arXiv preprint arXiv:2402.04110* **2024**.

12. Chen, J.; Zhao, W. Logistics automation management based on the Internet of things. *Cluster Computing* **2019**, *22*, 13627–13634.
13. Gouda, A.; Ghanem, A.; Reining, C. DoPose-6d dataset for object segmentation and 6d pose estimation. In Proceedings of the 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2022, pp. 477–483.
14. Franke, S.; Bommert, A.; Brandt, M.J.; Kuhlmann, J.L.; Olivier, M.C.; Schorning, K.; Reining, C.; Kirchheim, A. Smart pallets: Towards event detection using imus. In Proceedings of the 2024 IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA). IEEE, 2024, pp. 1–4.
15. OpenAI. Model Release Notes. <https://help.openai.com/en/articles/9624314-model-release-notes>. Visited on 2024-12-05.
16. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv preprint arXiv:2303.18223* **2023**.
17. IBM. What are large language models (LLMs)? <https://www.ibm.com/think/topics/large-language-models>. Visited on 2024-12-15.
18. Carlini, N.; Tramer, F.; Wallace, E.; Jagielski, M.; Herbert-Voss, A.; Lee, K.; Roberts, A.; Brown, T.; Song, D.; Erlingsson, U.; et al. Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (USENIX Security 21), 2021, pp. 2633–2650.
19. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **2020**, *21*, 1–67.
20. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435* **2023**.
21. Vaswani, A. Attention is all you need. *Advances in Neural Information Processing Systems* **2017**.
22. Kirchenbauer, J.; Geiping, J.; Wen, Y.; Katz, J.; Miers, I.; Goldstein, T. A watermark for large language models. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 17061–17084.
23. OpenAI. Introducing ChatGPT. <https://openai.com/index/chatgpt/>. Visited on 2024-12-12.
24. Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* **2024**, *15*, 1–45.
25. Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y.T.; Li, Y.; Lundberg, S.; et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* **2023**.
26. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* **2021**.
27. Adeshola, I.; Adepoju, A.P. The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments* **2023**, pp. 1–14.
28. Sok, S.; Heng, K. ChatGPT for education and research: A review of benefits and risks. *Cambodian Journal of Educational Research* **2023**, *3*, 110–121.
29. Lo, C.K. What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences* **2023**, *13*, 410.
30. Montenegro-Rueda, M.; Fernández-Cerero, J.; Fernández-Batanero, J.M.; López-Meneses, E. Impact of the implementation of ChatGPT in education: A systematic review. *Computers* **2023**, *12*, 153.
31. Grassini, S. Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences* **2023**, *13*, 692.
32. Rahman, M.M.; Watanobe, Y. ChatGPT for education and research: Opportunities, threats, and strategies. *Applied Sciences* **2023**, *13*, 5783.
33. Halaweh, M. ChatGPT in education: Strategies for responsible implementation. *Contemporary educational technology* **2023**, *15*.
34. Kasneci, E.; Seßler, K.; Küchemann, S.; Bannert, M.; Dementieva, D.; Fischer, F.; Gasser, U.; Groh, G.; Günnemann, S.; Hüllermeier, E.; et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and individual differences* **2023**, *103*, 102274.
35. Geerling, W.; Mateer, G.D.; Wooten, J.; Damodaran, N. Is ChatGPT smarter than a student in principles of economics. *Available at SSRN* **2023**, 4356034.
36. Susnjak, T.; McIntosh, T.R. ChatGPT: The end of online exam integrity? *Education Sciences* **2024**, *14*, 656.

37. Stutz, P.; Elixhauser, M.; Grubinger-Preiner, J.; Linner, V.; Reibersdorfer-Adelsberger, E.; Traun, C.; Wallentin, G.; Wöhs, K.; Zuberbühler, T. Ch (e) atGPT? an anecdotal approach addressing the impact of ChatGPT on teaching and learning Giscience **2023**.
38. Buchberger, B. Is ChatGPT smarter than master's applicants. *Research Institute for Symbolic Computation: Linz, Austria* **2023**, pp. 23–04.
39. de Winter, J.C. Can ChatGPT pass high school exams on English language comprehension? *International Journal of Artificial Intelligence in Education* **2023**, pp. 1–16.
40. Choi, J.H.; Hickman, K.E.; Monahan, A.B.; Schwarcz, D. ChatGPT goes to law school. *J. Legal Educ.* **2021**, *71*, 387.
41. Hargreaves, S. Words Are Flowing Out Like Endless Rain Into a Paper Cup': ChatGPT &. *Law School Assessments, SSRN Electronic Journal* **2023**.
42. Kung, T.H.; Cheatham, M.; Medenilla, A.; Sillos, C.; De Leon, L.; Elepaño, C.; Madriaga, M.; Aggabao, R.; Diaz-Candido, G.; Maningo, J.; et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS digital health* **2023**, *2*, e0000198.
43. Gilson, A.; Safranek, C.W.; Huang, T.; Socrates, V.; Chi, L.; Taylor, R.A.; Chartash, D.; et al. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR medical education* **2023**, *9*, e45312.
44. Fijačko, N.; Gosak, L.; Štiglic, G.; Picard, C.T.; Douma, M.J. Can ChatGPT pass the life support exams without entering the American heart association course? *Resuscitation* **2023**, *185*.
45. Wang, X.; Gong, Z.; Wang, G.; Jia, J.; Xu, Y.; Zhao, J.; Fan, Q.; Wu, S.; Hu, W.; Li, X. ChatGPT performs on the Chinese national medical licensing examination. *Journal of medical systems* **2023**, *47*, 86.
46. Huh, S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: A descriptive study. *J Educ Eval Health Prof* **2023**, *20*, 1.
47. Frieder, S.; Pinchetti, L.; Griffiths, R.R.; Salvatori, T.; Lukasiewicz, T.; Petersen, P.; Berner, J. Mathematical capabilities of chatgpt. *Advances in neural information processing systems* **2024**, *36*.
48. Jalil, S.; Rafi, S.; LaToza, T.D.; Moran, K.; Lam, W. Chatgpt and software testing education: Promises & perils. In Proceedings of the 2023 IEEE international conference on software testing, verification and validation workshops (ICSTW). IEEE, 2023, pp. 4130–4137.
49. Newton, P.; Xiromeriti, M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. *Assessment & Evaluation in Higher Education* **2024**, *49*, 781–798.
50. Oh, N.; Choi, G.S.; Lee, W.Y. ChatGPT goes to the operating room: Evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research* **2023**, *104*, 269–273.
51. Rizzo, M.G.; Cai, N.; Constantinescu, D. The performance of ChatGPT on orthopaedic in-service training exams: A comparative study of the GPT-3.5 turbo and GPT-4 models in orthopaedic education. *Journal of Orthopaedics* **2024**, *50*, 70–75.
52. Currie, G.M. GPT-4 in nuclear medicine education: Does it outperform GPT-3.5? *Journal of Nuclear Medicine Technology* **2023**, *51*, 314–317.
53. Takagi, S.; Watari, T.; Erabi, A.; Sakaguchi, K.; et al. Performance of GPT-3.5 and GPT-4 on the Japanese medical licensing examination: Comparison study. *JMIR Medical Education* **2023**, *9*, e48002.
54. Jin, H.K.; Lee, H.E.; Kim, E. Performance of ChatGPT-3.5 and GPT-4 in national licensing examinations for medicine, pharmacy, dentistry, and nursing: A systematic review and meta-analysis. *BMC Medical Education* **2024**, *24*, 1013.
55. Savelka, J.; Agarwal, A.; An, M.; Bogart, C.; Sakr, M. Thrilled by your progress! large language models (gpt-4) no longer struggle to pass assessments in higher education programming courses. In Proceedings of the Proceedings of the 2023 ACM Conference on International Computing Education Research-Volume 1, 2023, pp. 78–92.
56. Yeadon, W.; Peach, A.; Testrow, C. A comparison of human, GPT-3.5, and GPT-4 performance in a university-level coding course. *Scientific Reports* **2024**, *14*, 23285.
57. Tian, J.; Hou, J.; Wu, Z.; Shu, P.; Liu, Z.; Xiang, Y.; Gu, B.; Filla, N.; Li, Y.; Liu, N.; et al. Assessing Large Language Models in Mechanical Engineering Education: A Study on Mechanics-Focused Conceptual Understanding. *arXiv preprint arXiv:2401.12983* **2024**.
58. Katz, D.M.; Bommarito, M.J.; Gao, S.; Arredondo, P. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A* **2024**, *382*, 20230254.

59. White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D.C. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* **2023**.
60. Franke, S. Test Results: Can ChatGPT Solve Undergraduate Exams from Logistics Studies? An Investigation. Zenodo, Dec. 12, 2024, <https://doi.org/doi:10.5281/zenodo.14412298>.
61. TU Dortmund University. Module Description Bachelor's degree in Logistics. https://mb.tu-dortmund.de/storages/mb/r/Formulare/Studiengaenge/B.Sc._Logistik.pdf. Visited on 2024-12-05.
62. Liu, H.; Ning, R.; Teng, Z.; Liu, J.; Zhou, Q.; Zhang, Y. Evaluating the logical reasoning ability of chatgpt and gpt-4. *arXiv preprint arXiv:2304.03439* **2023**.
63. Großeschallau, W. *Materialflussrechnung: Modelle und Verfahren zur Analyse und Berechnung von Materialflusssystemen*; Springer-Verlag, 2013.
64. Ten Hompel, M.; Schmidt, T.; Dregger, J. *Materialflusssysteme: Förder-und Lagertechnik*; Springer-Verlag, 2018.
65. OpenAI. OpenAI's Reinforcement Fine-Tuning Research Program. <https://openai.com/form/rft-research-program/>, 2024. Visited on 2024-12-07.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.