Expert Evaluation of Export Control-Related Question Answering Capabilities of LLMs

1st Rafal Rzepka
Faculty of Information Science and Technology
Hokkaido University
Sapporo, Japan
rzepka@ist.hokudai.ac.jp

2nd Shinji Muraji
Gradutae School of Information Science and Technology
Hokkaido University
Sapporo, Japan
shinjimuraji@ist.hokudai.ac.jp

3rd Akihiko Obayashi
Center for Innovation and Business Promotion
Hokkaido University
Sapporo, Japan
obayashi@mcip.hokudai.ac.jp

2023 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 979-8-3503-4107-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/CSDE59766.2023.10487735

***Abstract*—In this paper we introduce evaluation experiments performed by an expert to assess Large Language Models ability to answer questions related to export control in Japanese language. We compare outputs of two popular models, namely ChatGPT and GPT-4, and measure their accuracy in predicting intention of the question and label of the answer. Prediction results indicate that both models were better in guessing question intention but failed to recognize what type of an answer would be most probable. Furthermore, expert evaluation of outputs generated by both models show that most of the answers contain mistakes which can mislead users asking a dialog system for an expertise. We analyze these outputs and enumerate problems to be tackled in the future.**

***Index Terms*—export control, question answering, large language models**

I. Introduction

Large Language Models (LLMs) has changed many aspects of natural language processing. Machine learning which required vast amounts of samples can be often replaced by LLMs' ability to generate text without the necessity of preparing data, which is costly and time-consuming. Our research goal is to develop an expert dialog system which is capable of answering questions related to export control. Researchers who deal with potentially dangerous materials, viruses or technologies, often require special permits for sharing their work with others. However, the regulations are not always straightforward, and researchers look for guidance from experts in trade security. We aim to address the shortage of experts problems by developing a reliable dialogue system [1] capable of correctly answering questions and engaging in educational conversations on related topics. First trials with implementing older NLP methods were performed in the past [2], [3], but did not yield satisfactory results. Obayashi and Rzepka developed a dataset for question answering related to the issues of trade security [4], and then extended by a part which we decided to utilize for testing LLMs on the topic. We have chosen ChatGPT and GPT-4, two commercial Large Language Models which showed to improve scores on many NLP tasks [5]. Using commercial language models for research is problematic, but we decided to utilize them as the preliminary test bed to acquire insights into LLMs capabilities to be used in the future when sufficiently large open-source models are developed, especially these dedicated to Japanese language. The paper is structured as follows: in the next section (II) we report related research, data we utilized is described in Section III. In Section IV we introduce prompts used in the experiments which are reported in Section V. After presenting the results in Section VI, we discuss our findings (VII). We conclude our paper and describe the next steps in Section VIII.

II. Related Work

A. Expert Systems Era

Answering questions related to specialized fields is a task of a long history. First expert systems for question answering (QA) were developed in 60's, showing that semantic analysis and logical inference have some potential but are questionable when it comes to generality [6]. Two early question-answering systems, BASEBALL [7] and LUNAR [8], were developed to answer questions in specific domains. Former addressed inquiries concerning Major League Baseball, and the latter handled queries regarding the geological examination of rocks brought back from the Apollo Moon missions. Famous expert systems like MYCIN [9], although focusing on knowledge base reasoning capabilities rather tha natural language processing, provided interactive explanation functionalities. These early attempts to solve the QA task pointed out the difficulties coming with subtle aspects of meaning and suggested developing vast dictionaries and correspondingly large grammars. With the advent of machine learning in the 90's, question-answering methods started to include classification task to determine the type of question and the type of answer [10]. Except experimenting with the close-domain QA task, we also investigate automatic classification.

B. Machine Learning Era

The era of machine learning approaches has shifted researchers' attention to open-domain QA task, as they required large amounts of data which was easier to collect from Internet resources like Yahoo! Answers or even automatically generated [11], [12]. Successes of systems like IBM's Watson [13], although with limited form of natural language output, led to development of personal assistants like Siri, Google

979-8-3503-4107-2/23/$31.00 ©2023 IEEE

Assistant, Cortana or Alexa. Such systems are usually limited to answering questions about specific information like weather or the device setup, but can be connected to external sources as WolframAlpha to answer more specialized answers. When it comes to complicated queries, they usually suggest searching the Web, which is cumbersome and time consuming. This problem has been tackled by many after the introduction of transformer architecture [14] and Bidirectional Encoder Representations from Transformers (BERT) [15]. Fine-tuning BERT with specialized data or training from scratch including a domain topic as science [16], biology [17], finance [18] or law [19]. However, it should be noted that these models are trained on vast amounts of documents and dealing with very wide topics not specializing in small subsets of domains like trade security.

C. Large Language Models Era

The advent of large language models has changed this situation, especially the shift from GPT-2 (1.5 billion parameters) [20] to GPT-3 (175 billion parameters) [21] using very large part of the Internet for training. The successor of the third version of Generative Pre-trained Transformer (GPT), namely ChatGPT[1], has been opened for users world-wide, showing its naturalness of text generation, and with the latest version (GPT-4), OpenAI LLMs have shown its usefulness in improving scores in various benchmarks [5]. While the details of GPT-4 models are not revealed to the world, ChatGPT relies heavily on RLHF (Reinforcement Learning with Human Feedback) approach, in which human workers select better answers. This method helps to achieve more naturally-sounding answers but does not guarantee the correctness of outputs. This trend has been confirmed by experiments comparing LLMs' outputs with human experts, for example Karpinska and Iyyer compared translations performed by GPT-3.5 with translations performed by professional translators [22] and showed a variety of errors, especially when English was not involved (e.g. Russian to Japanese in both paragraph and sentence levels). As the use of LLMs is costly, and human evaluation is time consuming, researchers perform experiments on small sets of examples or with existing human-made outputs in a shorter form that can be easily compared with generations of LLMs. For instance, scientific knowledge can be borrowed from test meant for students. Testing LLMs on mathematical reasoning [23], physics [24], medicine [25] or economics [26]. There have been also experiments with the legal domain – Choi et al. [27] have tested ChatGPT using law class exams like Constitutional Law, Employee Benefits or Taxation. The model achieved better results on essay questions than on multiple-choice questions. Although it passed all exams, its scores placed it close to the bottom of the law school students ranking.

D. Tests in Japanese Language

When it comes to Japanese language, testing OpenAI models' expertise in a narrow domains focus on the field of medicine. Kasai et al. introduce IGAKUQA[2] benchmark [28] and test OpenAI models on this set of questions with multiple choices as answers. The results show that GPT-4 otperforms ChatGPT and GPT-3 passing all exams but it also makes crucial mistakes as recommending euthanasia which is prohibited in Japan. Kusonose and colleagues [29] test ChatGPT answering clinical questions on the Japanese Society of Hypertension guidelines – they use Shannon Entropy to measure the degree of uncertainty or randomness in the responses generated by ChatGPT, but human evaluation is not performed. In the legal domain Yu et al. [30] show effectiveness of an entailment task based on the Japanese Bar exam for testing zero-shot/few-shot and fine-tuning approaches, but they work on Japanese texts automatically translated to English. To the authors' best knowledge, no regulation-oriented QA task has ever been investigated and manually evaluated by an expert.

III. QA Dataset

To examine correctness of LLMs in question answering task, we utilize part of existing QA Dataset described in [4]. Because most of the examples coming from original FAQ consist of very long questions, we decided to use only the newly added questions which are shorter and easier to classify when it comes to the intent classification [31]. As the manual check is time consuming, and our expert (third author of this paper) is a sole advisor for the whole university, we limited the test set to 50 randomly selected questions. All queries are related to export control regulations which focus on the export of specific items or technologies, often with dual-use potential (civilian and military). The topics are related to details regarding issues like arms exports, technology transfers, and sensitive information sharing[3].

Except containing questions and answers, the data [31] is annotated with labels indicating what is the intent of a query (three types: original "Yes/No", "Request for Explanation" and "Correctness Confirmation") and what type is the answer (four types: "Regulated", "Not Regulated", "Confirmation Required" and "Not About Regulation"). Examples are presented in Table V-B.

IV. Methodology

The strategy for all our experiments is to create prompts asking two language models, namely ChatGPT[4] and GPT-4[5], to answer or label questions. To acquire as much feedback from the expert as possible, we decided not to use all 50 questions as an input to both models but to divide them into two equal subsets and feed different sets to both models. For this reason it must be noted the comparison between both models is not a strict one – it was performed to assess their relative performance and capture a broader range of possible mistakes.

[1]https://chat.openai.com/chat, GPT-3.5 in the API nomenclature

[2]https://github.com/jungokasai/IgakuQA

[3]The main objectives of export control are to prevent the proliferation of weapons of mass destruction (WMD), safeguard sensitive technologies, protect intellectual property, and maintain strategic trade relationships.

[4]model *gpt3.5-turbo-0613*

[5]model *gpt4-0613*

TABLE I
SAMPLES FROM THE EXPORT CONTROL QA DATASET

Question	Expert Answer	Intent (Q)	Type (A)
We are required to send Ebola hemorrhagic fever virus to the Pasteur Institute, but do we need permission from the Minister of Economy, Trade and Industry?	Please note that all viruses of the genus Ebola virus are regulated.	Yes/No	Regulated
What kind of microscopes are regulated?	Optical microscopes as well as scanning electron microscopes and transmission electron microscopes are not regulated.	Request for Explanation	Not Regulated

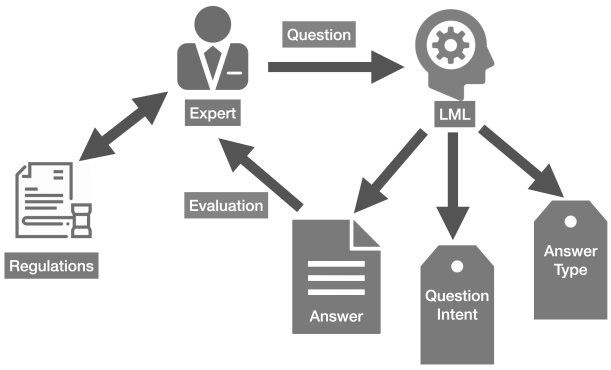

Fig. 1. Automatic labelling and question answering process performed by Large Language Models

A. Question Answering

Questions from dataset introduced in Section III are preceded by the following prompts:
("role": "system", "content": "You are export control expert"), and
("role": "user", "content": "Answer the following question.")
Temperature is set to the default value (1) but due to the costs and time consuming evaluation we do not repeat generations adjusting the randomness of the generated text. As mentioned above, in this research we focus on collecting expert's feedback rather than comparing accuracy of models.

B. Question Intent Labelling

As the dataset contains labels, we also prompt both models to predict them as there are reports about LLMs outperforming humans in annotation tasks [32]. We use the following question intent classification prompt:

> *For each of the following questions, please answer A, B, or C for the intent of the person who asked the question.*
> *A: (someone)*[6] *wants an explanation*
> *B: (someone) wants to confirm that (her/his) understanding is correct*
> *C: (someone) wants you to answer which of two things is correct*
> *Make sure to answer with "A", "B", or "C".*

[6]In Japanese it is usually more natural to omit pronouns.

With this prompt we aim to acquire one of the three labels ("Yes/No", "Request for Explanation" and "Correctness Confirmation").

C. Answer Type Labelling

As explained in the Section III, there are four answer type labels: "Regulated", "Not Regulated", "Confirmation Required" and "Not About Regulation" in the dataset.

Here we decided to divide the prediction task into two sub-tasks. The first is to recognize what type of answer is the expert's answer, and the second one subtask is to predict what the answer type would be from the question only. It is to see how much knowledge a model posses on the topic without being confronted with expert's answer. Prompts for both subtasks are given below:

1) Prediction From Answer:

> *The following is an answer to a question about export control.*
> *Please label the following answer with an A, B, C, or D.*
> *A: It is a subject to a regulation*
> *B: It is not a subject a regulation*
> *C: Necessary to check with company or government*
> *D: Question does not ask for any judgment*
> *Make sure to answer with "A", "B", "C" or "D".*

2) Prediction From Question:

> *The following is a question about export control.*
> *Please label the following questions with an A, B, C, or D.*
> *(the remaining part is identical to 1) above).*

V. EXPERIMENTS

A. Question Answering Capability

We have used the prompts described in Section IV and asked ChatGPT and GPT-4 to answer 50 randomly selected questions related to export control, 25 different questions to every model. Our expert has read all the answers and marked them on Likert scale from 1 to 5, where the lowest value indicates complete incorrectness and the highest one meaning the specialist was satisfied wit the model's answer. The expert was also asked to describe reasons of low scores whenever possible.

B. Label Prediction Capability

Both models were again asked by prompts described in Section IV to predict an intent of a given question and type of the answer. As the labels are given in the dataset we used, we compare the LLM's choices with these labels to calculate prediction correctness.

VI. Experimental Results

Surprisingly, ChatGPT (gpt3.5-turbo) has achieved higher expert scores (3.96 on average) than GPT-4 (3.72). Figures 2 and 3 present detailed scores on two different 25 questions.

When it comes to automatic label prediction tasks, as shown in Table III, this time *gpt-4* outperforms *gpt-3.5-turbo* in all three categories. The notable improvement in gpt-4's ability to deduce answer types from questions alone suggests advancements in contextual understanding, making it a more contextually aware model compared to its predecessor, gpt-3.5-turbo. While gpt-4 demonstrates some knowledge related to export control, the expert evaluator has rejected 1/4 of its answers (1/3 of ChatGPT's answers).

It should be noted that annotation abilities of OpenAI models are far from satisfactory – with ChatGPT not even reaching the correctness of random baseline.

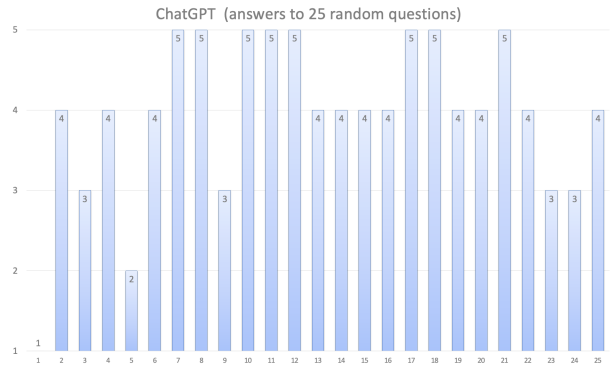

Fig. 2. Expert evaluation of ChatGPT answers (1 being completely incorrect and 5 completely correct)

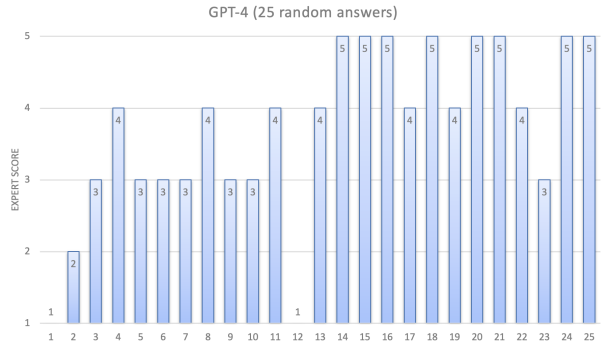

Fig. 3. Expert evaluation of GPT-4 answers (1 being completely incorrect)

VII. Discussion

A. Detected Errors

Here we report some of the expert comments about incorrect or lacking answers from both models.

- *The first sentence is incorrect; section 16 requires a permit only if it pertains to catch-all regulations.*
- *It appears to be a correct response, but does not include a response regarding security export controls.*
- *I have the impression that it is very correct, but the content is total nonsense.*
- *There is no way to refer to the website of the Export Control Service of the Russian Federation or the relevant agency in the first suggestion (confirm export restricted goods). The fourth suggestion ("confirm cultural differences") has nothing to do with export control.*
- *It is not a mistake, but the answer is lacking (ChatGPT answers with "Sorry, I am an AI providing general information and am not familiar with specific laws and regulations. Please check...").*
- *Composition of the answer is hard to understand. (ChatGPT does not seem to understand intent of the query which is not posed as question but statement expressing a will to organize a seminar. Model surprisingly starts with questions, but then continues the answer as a possible points for a talk at the seminar.)*
- *Mechanical cameras are not regulated, but some electronic cameras are.*
- *One needs to check if the liquid is designed for refrigerant in electronic equipment, otherwise it will belong to non-regulated items.*
- *Some parts are correct, but the list regulation and the catch-all regulation are mixed up.*
- *One needs to consider not only import regulations but also export regulations.*
- *It is difficult to say what "inseparable" means, since model gives only interpretation of the notice.*
- *Cargo and technology are listed together and it is not clear. The answer is not related to the topic of structural materials, although it is correct. It says "prohibited", but there is no prohibition on export in this case.*
- *It is unclear where "Appendix 1 of the Revenue Laws and Regulations" came from. The answer is also not clear.*
- *Bearings are regulated as "jikuuke" (synonym of "baearings" written in Chinese ideograms).*
- *Question begins with "in relation to the above", hence is not clear, and makes the answer difficult to understand.*

B. Error Analysis

The expert's comments suggest that there are replies which are very convincing but utterly wrong from the point of export control. Large language models are known to hallucinate, and these hallucinations can even snowball [33]. This phenomenon can be also confirmed as models are able to make up names of not existing regulations. Misunderstanding an intent of a plan to organize a seminar seems interesting – the prompt clearly

TABLE II
EXAMPLES OF MODELS ANSWERS

Question	Model Answer	Model	Expert Score
Is boron regulated?	Boron may have different export restrictions in different countries and regions. Because of the wide range of applications for boron, export restrictions may exist, especially for high-purity boron or certain forms of boron. For specific information on export restrictions, please contact the export control agency or customs and clearance authorities in the country or region where you are located.	ChatGPT	5
We received an inquiry from Company A of Country A, which is located outside of the region specified in Appended Table 3 of the Export Decree, for carbon fiber that falls under only item 16 of Appended Table 1 of the Export Decree. There was no particular wording in the contract or the order form that the product would be used for the development of weapons of mass destruction, etc., and the person in charge informed us that the carbon fiber was to be used as a reinforcing material for construction materials. However, by chance, at a later date, one of our employees had an opportunity to visit the factory of Company A in Country A before exporting the product, and upon visiting, he saw parts that appeared to be missiles on a trailer parked in a corner of the factory. In such a case, do we need to obtain an export license for exporting carbon fiber that falls under the 16th clause of Appended Table 1 of the Export Order?	Carbon fiber that falls under paragraph 16 of Appended Table 1 of the Export Decree normally requires an export license. However, we have received a communication that carbon fiber will be used as a reinforcing material for building materials, without any particular wording in the contract or order form that it will be used for the development of weapons of mass destruction, etc. In this case, based on the legal documents and the content of the communication, it is necessary to confirm that the material does not fall under any specific regulation. However, if there is information at the time of the visit that there are components on the trailer that are believed to be missiles, it is possible that the situation in question is regulated by an export decree. Therefore, additional information should be collected and, if necessary, reported to the relevant authorities for appropriate action. Depending on the circumstances, it may be appropriate to suspend the export of the relevant carbon fiber and either obtain an export license or hold off until the relevant authorities confirm the situation.	ChatGPT	1
We are required to send the Ebola hemorrhagic fever virus to the Pasteur Institute, but do we need permission from the Minister of Economy, Trade and Industry?	There are strict international regulations regarding the sending of viruses with such a high level of danger, as there is concern that they may be used as biological weapons. Generally, permission from the Minister of Economy, Trade and Industry is required based on the Biological Weapons Prevention Law. However, the specific procedures and necessary permission depend on the specifics of the case, so we recommend that you contact the relevant ministry or agency to obtain a definitive answer.	GPT-4	5
Are bearings regulated?	Bearings themselves are not normally subject to export controls. However, certain products in which bearings are incorporated, such as military technology or weapons, may be subject to regulations. Therefore, when exporting, it is important to confirm how the product will be used and what the recipient's purpose is. However, since regulations may change as circumstances, international relations, and policies change, it is necessary to research the latest information on export regulations to determine the specifics of the situation.	GPT-4	1

TABLE III
CORRECT LABEL PREDICTION PERCENTAGE

Model	Question Intention	Answer Type (Answer Given)	Answer Type (Question Only)
random baseline	33.0% (33/100)	25.0% (25/100)	25.0%(25/100)
gpt-3.5-turbo	66.0% (66/100)	23.0% (23/100)	31.0% (31/100)
gpt-4	75.0% (75/100)	27.0% (27/100)	50.0% (50/100)

states to answer a question, but probably most of the RLHF evaluators assessed a will statement as a request for some kind of help in organizing it, therefore the model has overridden the prompt request. Another surprising example is GPT-4 failing to recognize a synonym of a "bearing", a problem that can be solved by looking at Wikipedia links [3]. Although *jikuuke* is almost three times less frequent than *bearingu* in Google search engine, it appears over 6.2 million times in the Internet in very similar contexts.

VIII. CONCLUSION AND FUTURE WORK

In this paper we described our experiments investigating quality of question answering ability of large language models (ChatGPT and GPT-4) in the area of Japanese export control. We performed series of four experiments of two types – question answering and label prediction. The experimental results show that scores of both models are close to 4, which is treated as a satisfactory answer. GPT-4 has made two fatal mistakes that were completely refuted by the expert – e.g. it regarded cameras as not regulated items because it wrongly assumed that the question is about mechanical

cameras, not digital ones. ChatGPT made only one such error.1 In the label prediction tasks, the difference between models' performance was clearer where GPT-4 distinctly outperformed its predecessor. Although both sample sets (50 questions and answers in total) is not a big one, the results showed that both models in both tasks had problems with interpreting some implicit intents of a question. Much worse results of predicting answer from the question suggest that we need to carefully lead our dialog system regarding how the prompt is created. As Dhuliawala et al. [34] showed, chain-of-verification might be a necessary addition to the algorithm in order to limit the possibility of hallucinations which are unforgivable in tasks related to trade security and similar areas. As another future work, we plan to investigate the effectiveness of providing various documents on export control to models which allow larger prompts.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 23K11757.

REFERENCES

[1] A. Obayashi and R. Rzepka, "Towards interactive advisory system for security export control," in *Proceedings of IJCAI Workshop on Language Sense on Computer, Macau*, 2019.

[2] R. Rzepka, D. Shirafuji, and A. Obayashi, "Limits and challenges of embedding-based question answering in export control expert system," *Procedia Computer Science*, vol. 192, pp. 2709–2719, 2021, knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 25th International Conference KES2021.

[3] R. Rzepka, S. Muraji, and A. Obayashi, "Utilizing wikipedia for retrieving synonyms of trade security-related technical terms," in *Proceedings of the 10th Language and Technology Conference*. Adam Mickiewicz University Press, 2023.

[4] A. Obayashi and R. Rzepka, "Annotated question and answer dataset for security export control," in *Proceedings of the 7th Linguistic and Cognitive Approaches to Dialog Agents (LaCATODA 2021) IJCAI 2021 Workshop, CEUR Workshop Proceedings*, vol. 2935. Montreal, Canada, 2021.

[5] R. Mao, G. Chen, X. Zhang, F. Guerin, and E. Cambria, "GPTeval: A survey on assessments of chatgpt and gpt-4," 2023.

[6] R. F. Simmons, "Natural language question-answering systems: 1969," *Commun. ACM*, vol. 13, no. 1, p. 15–30, jan 1970. [Online]. Available: https://doi.org/10.1145/361953.361963

[7] B. F. Green Jr, A. K. Wolf, C. Chomsky, and K. Laughery, "Baseball: an automatic question-answerer," in *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, 1961, pp. 219–224.

[8] W. A. Woods, "Lunar rocks in natural english: Explorations in natural language question answering." vol. 5, p. 521–=569, 1977.

[9] E. Shortliffe, "Mycin: A knowledge-based computer program applied to infectious diseases," in *Proc. 1st SCAMC, IEEE*, 1977, pp. 66–69.

[10] L. Hirschman and R. Gaizauskas, "Natural language question answering: the view from here," *natural language engineering*, vol. 7, no. 4, pp. 275–300, 2001.

[11] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford *et al.*, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[12] E. M. Voorhees and D. M. Tice, "Building a question answering test collection," in *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, 2000, pp. 200–207.

[13] D. Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. A. Kalyanpur, A. Lally, J. W. Murdock, E. Nyberg, J. Prager, N. Schlaefer, and C. Welty, "Building watson: An overview of the deep-qa project," *AI Magazine*, pp. 59–79, 2010.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://www.aclweb.org/anthology/N19-1423

[16] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," 2019.

[17] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, sep 2019. [Online]. Available: https://doi.org/10.1093%2Fbioinformatics%2Fbtz682

[18] D. Araci, "Finbert: Financial sentiment analysis with pre-trained language models," *arXiv preprint arXiv:1908.10063*, 2019.

[19] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "Legal-bert: The muppets straight out of law school," 2020.

[20] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[21] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[22] M. Karpinska and M. Iyyer, "Large language models effectively leverage document-level context for literary translation, but critical errors persist," *arXiv preprint arXiv:2304.03245*, 2023.

[23] S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner, "Mathematical capabilities of chatgpt," *arXiv preprint arXiv:2301.13867*, 2023.

[24] G. Kortemeyer, "Could an artificial-intelligence agent pass an introductory physics course?" *Physical Review Physics Education Research*, vol. 19, no. 1, p. 010132, 2023.

[25] A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash *et al.*, "How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment," *JMIR Medical Education*, vol. 9, no. 1, p. e45312, 2023.

[26] W. Geerling, G. D. Mateer, J. Wooten, and N. Damodaran, "Chatgpt has aced the test of understanding in college economics: Now what?" *The American Economist*, p. 05694345231169654, 2023.

[27] J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz, "Chatgpt goes to law school," *Available at SSRN*, 2023.

[28] J. Kasai, Y. Kasai, K. Sakaguchi, Y. Yamada, and D. Radev, "Evaluating gpt-4 and chatgpt on japanese medical licensing examinations," *arXiv preprint arXiv:2303.18027*, 2023.

[29] K. Kusunose, S. Kashima, and M. Sata, "Evaluation of the accuracy of chatgpt in answering clinical questions on the japanese society of hypertension guidelines," *Circulation Journal*, vol. 87, no. 7, pp. 1030–1033, 2023.

[30] F. Yu, L. Quartey, and F. Schilder, "Legal prompting: Teaching a language model to think like a lawyer," 2022.

[31] A. Obayashi and R. Rzepka, "Expanding export control-related data for expert system," in *Knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022, Procedia Computer Science*, vol. 207, 2022, pp. 3065–3072.

[32] F. Gilardi, M. Alizadeh, and M. Kubli, "Chatgpt outperforms crowd-workers for text-annotation tasks," *arXiv preprint arXiv:2303.15056*, 2023.

[33] M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," *arXiv preprint arXiv:2305.13534*, 2023.

[34] S. Dhuliawala, M. Komeili, J. Xu, R. Raileanu, X. Li, A. Celikyilmaz, and J. Weston, "Chain-of-verification reduces hallucination in large language models," 2023.