

Review

Applications of Large Language Models in Pathology

Jerome Cheng 

Department of Pathology, University of Michigan, Ann Arbor, MI 48105, USA; jeromech@med.umich.edu

Abstract: Large language models (LLMs) are transformer-based neural networks that can provide human-like responses to questions and instructions. LLMs can generate educational material, summarize text, extract structured data from free text, create reports, write programs, and potentially assist in case sign-out. LLMs combined with vision models can assist in interpreting histopathology images. LLMs have immense potential in transforming pathology practice and education, but these models are not infallible, so any artificial intelligence generated content must be verified with reputable sources. Caution must be exercised on how these models are integrated into clinical practice, as these models can produce hallucinations and incorrect results, and an over-reliance on artificial intelligence may lead to de-skilling and automation bias. This review paper provides a brief history of LLMs and highlights several use cases for LLMs in the field of pathology.

Keywords: large language model; generative pretrained transformer; bidirectional encoder representations from transformers; artificial intelligence; natural language processing; GPT; BERT; Mistral; Llama; Gemma; surgical pathology

1. Introduction

LLMs are based on the transformer neural network architecture introduced in the 2017 paper “Attention Is All You Need” written by Vaswani et al. [1] This paper significantly changed the landscape of natural language processing by providing the fundamental architecture used in modern large language models. Although it was not evident at the time of publication, it was soon discovered that transformers were very powerful in multiple natural language tasks that previously required different algorithms to solve. The first version of generative pretrained transformers (GPT) was released by OpenAI in June 2018 [2]. Later in the same year, Bidirectional Encoder Representations from Transformers (BERT) was released by Google researchers [3]. Since then, the progress of transformers has gone a long way, and every few months, models are being released that outperform earlier models. Initially, transformers were intended to assist only in natural language tasks, such as language translation (using sequence to sequence transformers), named entity recognition [4], and text summarization, until Dosovitskiy et al. discovered that transformers may be used for image recognition by using image embeddings with position encodings as an input for standard transformers [5]. Another breakthrough was achieved in 2020, when OpenAI released the 175 billion parameter GPT3 that performed very well in multiple metrics, showing that a very high parameter count may be key to the development of general language models [6]. Later in 2022, when ChatGPT3.5 was introduced to the public, it brought about a quick surge in popularity of LLMs due to its simple and user-friendly interface, while demonstrating impressive performance in natural language understanding, and providing seemingly intelligent answers to prompts/questions. Even experts in the field were completely astounded when ChatGPT was able to simulate human understanding and generate human-like responses at a very high level [7]. Similar to how we treat other deep neural networks as a black box, we still do not fully comprehend how a neural network trained only to predict the next word (based on probabilities) can be so powerful. It can follow instructions, explain text, summarize text, carry out meaningful conversations, write computer programs, generate reports, perform text manipulations, and



Citation: Cheng, J. Applications of Large Language Models in Pathology. *Bioengineering* **2024**, *11*, 342. <https://doi.org/10.3390/bioengineering11040342>

Academic Editor: Zeynettin Akkus

Received: 14 March 2024

Revised: 27 March 2024

Accepted: 29 March 2024

Published: 31 March 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

is proficient at generating well-written prose in different styles. In 2023, GPT4 was released, containing over 1 trillion parameters. It was significantly better than its predecessor in multilingual capabilities, contextual understanding, and reasoning skills [8].

The earlier generation of LLMs (e.g., BERT) had fewer parameters, and therefore were trainable on relatively smaller resources (e.g., 16 TPU chips) [3]. BERT (Bert large) has 345 million parameters, which is small compared to GPT4 and many recently released LLMs. Notwithstanding, BERT is still a very useful model, and its relatively small size makes it accessible to individuals and groups with fewer resources. Several groups have been able to train LLMs from scratch. Geiping and Goldstein showed that it was possible to train a language model with similar performance to BERT by using only a single consumer GPU in a single day [9]. Mitchell et al. trained a BERT transformer model with 275,605 pathology reports, which included 121 million words, to extract data from pathology reports using eight Nvidia Tesla V100 GPUs [10]. Modern LLMs have larger training set and compute requirements. To put things into perspective, Llama 65b took Meta approximately 21 Days to train using 2048 A100 GPUs with 80 GB of RAM [11], while Llama 2 70b, which is open sourced, took Meta a total of 1,720,320 GPU hours to train using the same type of GPUs [12]. These pre-trained models are referred to as foundation models. Foundation models may be fine-tuned for specific tasks through supervised learning (e.g., using a question-and-answer dataset) with much smaller GPU resources. A human feedback loop where a person decides whether an LLM-generated response is desirable or not can further refine model performance [13].

Various groups have published studies in almost every medical and non-medical field on the use of ChatGPT and other LLMs in their respective specialties, with mixed success. This review paper will focus on the applications of large language models in the field of pathology.

2. Education

ChatGPT shows considerable potential in education. LLMs can assist medical educators with curriculum development, writing presentations [14], formulation of course syllabuses, developing case scenarios, crafting learning plans, and developing proposals [15]. A curriculum may be created by prompting ChatGPT to create a specific curriculum with its associated requirements. In one study, ChatGPT was asked to provide models for medical education DEI (diversity, equity, and inclusion) curriculum development in a radiology residency program. ChatGPT was subsequently asked to create a curriculum based on one of the recommended models, with follow-up prompts asking ChatGPT to create surveys, goals, topics, and implementation steps [15].

ChatGPT can provide an interactive educational experience, summarize important concepts [16], and give instant access to information [17]. It can tailor the educational experience according to an individual's needs by providing personalized responses to questions with immediate feedback [18]. ChatGPT will provide a detailed explanation if asked to explain the management of hyperlipidemia. ChatGPT may be asked to summarize topics by prompting it to "provide a summary" of the desired topic. Alternatively, an entire article may be copied onto the chat box as part of a prompt that states "summarize the following:". Follow-up prompts can clarify specific points, errors, and inconsistencies. Subsequent prompts can be made to create multiple choice questions based on the article summary. Given a list of signs and symptoms, it can recommend a set of differential diagnoses and engage in a brainstorming session where students can ask follow-up questions [19]. LLMs can generate and organize material in different formats to facilitate learning. By specifying "in table format" in the prompt, information about certain topics can be summarized in a table (Figure 1). It can generate outlines, multiple-choice questions with answers and explanations [20], and even simulate conversations between people about a certain topic. There is a possibility that some of the information is incorrect due to the tendency of LLMs to hallucinate [19], so LLM-generated information must be verified with other sources.

As an AI, I can model the table as if I were an experienced pathologist presenting common carcinoma types with their immunostains and molecular tests. Here's a concise representation for 12 carcinoma types:

Carcinoma Type	Immunostains	Molecular Tests
Breast Carcinoma	ER, PR, HER2, GATA3, CK7/CK8	HER2 amplification, BRCA1/2 mutations
Lung Adenocarcinoma	TTF-1, Napsin A, CK7	EGFR mutations, ALK rearrangements
Lung Squamous Cell Carcinoma	p63, p40, CK5/6	None standardly used
Colorectal Carcinoma	CDX2, CK20, MLH1, MSH2, MSH6, PMS2	KRAS/NRAS/BRAF mutations, MSI status
Prostate Carcinoma	PSA, PSAP, NKX3.1, AMACR	TMPRSS2-ERG fusions, PTEN loss
Hepatocellular Carcinoma	HepPar1, Arginase-1, AFP	TERT promoter mutations
Renal Cell Carcinoma (Clear Cell)	PAX8, RCC, CD10, CAIX	VHL gene mutations
Ovarian Serous Carcinoma	WT1, PAX8, CA125 (serum)	BRCA1/2 mutations, TP53 mutations
Pancreatic Ductal Adenocarcinoma	CA19-9 (serum), CK19, CEA, SMAD4/DPC4	KRAS mutations
Bladder Urothelial Carcinoma	Uroplakin III, CK20, GATA3, p63	FGFR3 mutations, TERT promoter mutations
Squamous Cell Carcinoma of the Head & Neck	p63, p16, CK5/6	HPV status, TP53 mutations
Gastric Adenocarcinoma	CDH1, CK7, MUC5AC, HER2	HER2 amplification, PD-L1 expression

The immunostains listed here are often used to aid in the classification and sometimes the site of origin of various carcinomas. For molecular tests, many are used to guide treatment decisions as well as for prognostic purposes. Carcinomas, being diversified by their origin and behavior, may require different sets of molecular tests for a full analysis. The tests applied often depend on clinical guidelines and the specifics of a given case, including potential responsiveness to targeted therapies or other treatment implications.

Figure 1. ChatGPT-4 Turbo followed instructions appropriately after being given the following prompt: “You are an experienced pathologist. Give me a list of 12 carcinomas with associated immunostains and molecular tests in table format”.

In a paper using ChatGPT to answer pathology-related questions, such as “explain why transfusion-related diseases are avoidable”, it was able to provide credible responses in most cases, achieving an overall median rating of 4.08 out of 5 [21]. In a recently published study, the domain-specific knowledge of ChatGPT 3.5 in pathology was assessed to be the same level as a staff pathologist, while ChatGPT 4 exceeded that of a trained pathologist [17]. There may be instances where ChatGPT may not have information about certain topics in its knowledge base. In these cases, an article may be copied as part of the input prompt, and ChatGPT may be asked to summarize the article, or provide the important points within the text. The desired length of the summary may be specified, but there is no guarantee ChatGPT will follow it.

Not all studies pertaining to ChatGPT and education were positive. Ngo et al. [20] used ChatGPT 3.5 to generate questions for an immunology course, but it was able to generate correct questions with answers and explanations in only 32% (19 of 60) of cases. In a separate study, ChatGPT4 was tested with the 2022 American Society for Clinical Pathology resident question bank and it did not fare very well, scoring 60.42% in clinical pathology, 54.94% in anatomic pathology, and garnering an overall score of 56.98% [22]. Munoz-Zuluaga et al. tested GPT4 on 65 questions related to clinical laboratory medicine, and only 50.7% of answers were evaluated to be correct [23]. In another study, clinical chemistry faculty and trainees outperformed ChatGPT (versions 3.5 and 4) in answering 35 clinical chemistry questions [24]. Considering the poor performance of ChatGPT in answering medical questions in several studies, one should not trust everything generated by ChatGPT, and information should be cross-referenced with other reputable sources, such as textbooks.

3. Information Extraction

Cancer registries and research studies will benefit tremendously from automated information extraction, where manual review and encoding of pathology reports are often necessary when these reports are in free text format [25]. Choi et al. showed that LLM-assisted extraction of structured information (e.g., tumor location, surgery type, histologic grade) from pathology and ultrasound reports can lead to significant time and cost savings, compared to manual methods [26]. The accuracy of GPT 3.5 in this study was mostly in the 80s and 90s (overall accuracy of 87.7%), therefore suggesting that manual supervision is still needed. Considering the study used GPT 3.5, the numbers would probably be better if newer LLMs, such as GPT4, were used.

Information extraction from clinical notes and free-text reports is challenging using traditional methods; regular expressions [27] or multiple text matching rules have to be written, and these rules can be error-prone, especially when unstructured free text is involved. Determining the presence or absence of cancer within a report is prone to error, as looking for cases with “carcinoma” could retrieve cases of “negative for carcinoma” being phrased in different ways [28]. This is one task LLMs have proven to be good at, since they are trained to learn the meaning of words, in addition to looking for the presence or absence of a particular word. Transformers have been used for predicting CPT codes [29], naming entities in breast pathology reports [30], extracting clinical information [26], and extracting structured information from free-text pathology reports [31]. One study used retrieval-augmented generation, which enabled GPT4 to extract inclusion/exclusion criteria from free-text clinical reports, leading to increased efficiency and reduced costs [32]. Zhang et al. fine-tuned a BERT model to extract various concepts (e.g., site, degree of differentiation) from breast cancer reports, achieving an overall precision of 0.927 and recall of 0.939 [30]. Yang et al. trained a BERT model that can accurately predict the type of rejection and IFTA (interstitial fibrosis and tubular atrophy) in renal pathology reports [33]. Liu et al. developed a BERT deidentification pipeline using 2100 pathology reports, achieving a best F1-score of 0.9659 in identifying sensitive health information. It was deployed in a teaching hospital and managed to process over 8000 unstructured notes in real time [34]. Santos et al. released an open source pre-trained transformer model (PathologyBERT), which was trained on 347,173 pathology reports. It was found to be accurate in breast cancer severity classification, and may be utilized for other information extraction and classification tasks involving pathology reports [35].

Many earlier studies using transformers involve pre-training and fine-tuning of the BERT transformer architecture [35]. With modern chat-based LLMs such as ChatGPT, it is possible to extract information without model finetuning. LLMs can be prompted to identify the location of the primary cancer within a report, or determine the presence of lymphovascular invasion [36]. In using LLMs to extract text, providing a correctly phrased user prompt can have a significant impact on the quality of the output. For instance, it may

be necessary to instruct the LLM “You are a pathologist. Provide a concise answer.” if a shorter output is desired [37].

4. Text Classification

Scientific document classification is a labor-intensive task with diverse applications [38]. Since transformers are trained to learn contextual information between words, they can be used to classify documents using learned meanings and sentence embeddings [39]. A modified version of BERT (Sentence-BERT) used siamese and triplet network structures to derive meaningful sentence embeddings that may be compared using vector similarity metrics [40]. Embeddings are useful for text similarity search, text clustering, and text classification [41]. Transformer models are not necessarily superior to CNN and hierarchical self-attention networks in text classification, where the identification of keywords and phrases can be more significant, rather than the contextual meaning of words and sentences [42].

Several studies used LLMs for the classification of pathology reports and scientific literature. One study used BERT-derived embeddings with active learning approaches to classify and cluster pathology reports according to specific diagnosis and diagnostic category [43]. Fijacko et al. performed multinomial classification of abstract titles using the ChatGPT-4 application programming interface (API), through a python function call with predefined prompts, demonstrating the effectiveness of LLM-based approaches in bibliometric analysis [44]. Using optical character recognition to convert pathology reports into a textual format, Kefeli and Tatonetti trained several BERT-based models for TNM stage and cancer type classification [45,46]. Fang and Wang used several BERT models pre-trained on scientific literature for multi-label topic classification, achieving F1-scores over 90% [47].

5. Report and Content Generation

Many physicians, including pathologists, spend a significant amount of time writing notes and reports. LLMs can assist in medical report writing and produce presentations, potentially leading to an increase in efficiency [48]. LLMs can help automate the process of generating pathology reports [49] and summarize case visits [50]. It can populate a template using unstructured text, extract data from different sources, and combine data into a single report [51]. Reports may be reformatted by specifying the new output format, and the language can be simplified by asking it to avoid medical terms [52].

There is ongoing debate about the appropriateness of using ChatGPT in writing research manuscripts, and many scientists disapprove of this practice [53]. ChatGPT is trained on diverse and vast amounts of internet text, but it is not aware of the specific source of data it is trained on [54]. When it is prompted to provide a reference, it will respond that it cannot cite specific studies directly. Prompted to write a paper with citations, it will write a paper where some of the references may be fictitious. One study assessed the authenticity and accuracy of references ChatGPT (version 3.5) cited in 30 medical papers it generated, and it was revealed 47% were fabricated, 46% were authentic but inaccurate, and only 7% were without errors [55]. In another study, Naik et al. used ChatGPT to create a case report on synchronous bilateral breast cancer; where the generated explanations were sensible, but several errors were present in the citations [56].

Caution should be exercised when submitting papers with LLM-generated content to publishers, as journal policies vary regarding the use of LLMs. The ethical boundaries and acceptability of using AI in writing is still being discussed [57]. Some journals allow ChatGPT to be listed as a contributor in the acknowledgements section, while others explicitly prohibit listing ChatGPT as an author [58]. However, it is unclear how journals will be able to identify submissions that are AI-generated, as tools developed for this purpose can misidentify real abstracts as AI-generated [59], and researchers themselves have trouble differentiating between AI-generated and original content [60].

6. Prompt Engineering

Prompt engineering is the process of designing questions and instructions in order to obtain the best response from an LLM [61]. The appropriate combination of words is crucial to eliciting the most accurate and relevant response [62], although different LLMs can have a variable response to the same prompt [63]. Some trial and error is often necessary to determine the most suitable prompt for a specific task [64]. It is recommended to be specific, provide context/examples, phrase questions differently, and assign a role to the LLM [65] (e.g., you are a pathologist). In a study that extracted symptoms from medical narratives, few-shot prompting (with examples of desired input and output) had higher sensitivity and specificity than zero-shot prompting (without examples of input/output) [66]. A zero-shot prompt is one that does not contain any training data [67]. Kojima et al. found out that adding “Let’s think step by step” to a question significantly improved the performance of LLMs in zero-shot reasoning [68]. In one study identifying the presence of metastatic cancer in discharge summaries, it was revealed that clear, concise prompts with reasoning steps significantly improved performance [69]. A study by Abdullahi et al. showed that LLMs performed better with multiple choice prompts that narrowed the search space, as opposed to open-ended prompts [70].

7. Programming

With the aid of LLMs, developing software for pathology-related projects can become easier. LLMs will enable pathologists with little or no programming experience to design and create programs [71]. AI has numerous potential impacts in computer programming: it can increase human productivity, automate tasks, reduce errors, document processes, and assist in bug detection [72,73]. It can also aid in data preparation and the development of pathology data visualization tools, websites, and artificial intelligence software using different languages, including Python and Matlab [74]. However, human validation is required to ensure proper performance of LLM-generated code [75]. ChatGPT can help you write a program to split a whole slide image into smaller images, simply by prompting it with “write a program for tiling a whole slide image into 224 by 224 pixel tiles”. It can also translate code from one programming language to another, which is highly valuable for someone translating legacy code to a modern language, or switching deep learning code to a different framework [76].

With a few lines of code, it is possible to generate API calls to GPT4 and other LLMs for inference. Access to the ChatGPT API is not free, so if cost or data privacy is a concern [77], many open source models are available, which are already very capable, but not as performant as GPT4, in part due to the smaller parameter size of these open models. In spite of this, many relatively simple tasks (e.g., text summarization, information extraction from templates) can be performed on open source models such as Mistral [78], Llama2, and Gemma.

8. Clinical Pathology

ChatGPT and other LLMs can assist with the interpretation of laboratory tests [23], but they do not have the level of knowledge necessary to replace the judgment of medical personnel [79], so caution must be exercised when using LLMs to interpret laboratory tests [24]. In an assessment involving 10 clinical microbiology case scenarios, ChatGPT’s performance was above average at best, and unsatisfactory in a few cases [80]. In another study, healthcare personnel were found to be better than ChatGPT at identifying fluid contamination of basic metabolic panel results [81]. Yang et al. evaluated ChatGPT-4 on blood smear images, and found it was able to identify 88% of normal blood cells and 54% of abnormal blood cells [79]. Kumari et al. [82] conducted a comparison of LLM performance in answering 50 hematology-related questions, where ChatGPT 3.5 achieved the highest score of 3.15 out of 4. The results were promising but indicated a need for LLM validation due to potential for inaccuracies [82].

LLMs can generate incorrect transfusion medicine recommendations. ChatGPT erroneously recommended Rho(D) immune globulin to prevent the development of anti-K antibodies during pregnancy [83]. In another study, several LLMs were given blood transfusion-related questions; Google Bard, GPT 3.5, and GPT4 achieved Receiver Operating Characteristic Curve (ROC AUC) scores of 0.65, 0.90, and 0.92, respectively [84].

9. Multi-Modal Large Language Models

LLMs that are trained with text and other types of data are referred to as multi-modal large language models (MLLMs) [85]. Of particular interest are recently developed MLLMs that can provide textual descriptions of microscopical images. MLLMs can also be tailored for object detection [86], and therefore may be used for counting cells and mitosis counting. Off-the-shelf MLLMs, such as GPT4v, are not well-trained on pathology images, so are not very suitable for analyzing histopathological images. In one study involving 100 colorectal polyp photomicrographs, ChatGPT achieved a sensitivity of 74% and specificity of 36% in adenoma detection [87]. Sievert et al. trained ChatGPT with 16 oropharyngeal confocal laser endomicroscopy images (8 with squamous cell carcinoma, 8 with normal mucosa) and it was tested with 139 images (83 with squamous cell carcinoma and 56 with healthy normal mucosa), achieving an overall accuracy of 71%, demonstrating an ability for few-shot learning. It was interesting that ChatGPT aborted the experiment when the terms “healthy” and “malignant” were used and recommended consulting a medical professional, so alternative terms had to be used to pursue the experiment [88]. However, it is highly likely that future versions of ChatGPT will fare better at interpreting histopathology images as more publicly available image/text pathology datasets become part of its training set.

Some groups fine-tuned MLLMs with surgical pathology images, with promising results. Tsuneki et al. combined convolutional neural network derived features with a recurrent neural network to generate histopathological descriptions of adenocarcinoma cases [89]. Their best model achieved a BLEU-4 score of 0.324. Sengupta and Brown demonstrated that whole slide image descriptions can be generated by combining encodings from a pre-trained vision transformer with a pre-trained BERT model, achieving a BLEU-4 score of 0.5818 [90]. Additionally, their best performing model predicted the tissue type with 79.98% accuracy, and patient gender with 66.36% accuracy. Sun et al. utilized 207,000 pathology image–text pairs to fine-tune a pre-trained OpenAI CLIP base model [91], and combined it with a 13B parameter LLM to develop PathAsst [92], which exhibited an impressive performance in interpreting pathology images. In another study, Yu et al. developed a vision-language pathology artificial intelligence (AI) assistant, named PathChat, by using 100 million histology images, 1.18 million pathology image–caption pairs, and 250,000 visual language instructions. Despite only using a 13B model combined with a vision transformer, PathChat [93] outperformed ChatGPT4v in interpreting pathology microscopic images, which was not totally unexpected, but this reinforces the notion that model size is not the only criteria for predicting performance. For a list of other MLLMs, Zhang et al. compiled a list of foundation models in healthcare, which includes several other MLLMs trained with pathology material [94].

Recently, there have been concerns that AI tools such as the aforementioned will replace pathologists, but the consensus is that these AI-based tools will supplement and augment the performance of pathologists and laboratory personnel, which would lead to increased efficiency and cost savings [95]. AI tools can be assigned to routine and boring tasks, like mitosis counting, cell counting, looking for metastasis in lymph nodes, or tasks that computers are inherently good at, such as calculating the tumor percentage on a digital slide. Additionally, adopting AI technologies is expected to improve the detection of rare events, diagnostic accuracy, and report quality [96].

10. Challenges and Limitations

Although LLMs show tremendous potential in improving the practice of pathology, LLMs can be prone to bias and knowledge plagiarism [97], and can also make mistakes.

The core neural network of LLMs has no information beyond the time of its training, and most cannot access internet data, except for a few products that allow the LLM to search the internet for additional information [98]. Therefore, most LLMs will not be knowledgeable of recently updated clinical guidelines. LLMs can produce hallucinations [99] and recommend products/tools that may not exist. Retrieval augmented generation (RAG) methods offer promise in mitigating some of these deficiencies [100]. RAG comprises a set of techniques and methodologies that enable a large language model to query an external data source, potentially improving the accuracy and relevance of responses while reducing the incidence of hallucinations and inappropriate responses [101]. Building guardrails into LLMs is another way of reducing the occurrence of improper responses [102].

LLMs can automatically generate reports based on provided data but, due to the unpredictable nature of LLMs, care must be undertaken in integrating these technologies into pathology practice, to mitigate the risk of introducing errors into patient reports. LLM-generated reports may contain inaccuracies, biases, and fictitious content. One study noted that ChatGPT 3.5 produced a high rate of fabricated references, so particular attention should be given to LLM-generated references [55]. In its current state, there is a need for manual supervision to ensure no improper content is introduced into LLM-created content, and it is not advisable to incorporate LLMs into a fully automated workflow.

Some groups have voiced concerns that over-reliance on AI tools for pathology case sign-out may lead to deskilling, burnout, and diminished knowledge of the histology and mechanisms of disease [103,104]. Pathologists could end up using an AI tool to sign out a case without critically analyzing the case [105]. Automation bias is a closely related problem, which is the propensity of humans to believe that AI is right, in spite of information to the contrary [106]. Therefore, it is important for medical practitioners to be cognizant that their own judgments are valuable and be aware of the limitations of AI tools they use in practice. Inadequate understanding of how artificial intelligence models work is a key contributing factor to automation bias [106]. An AI-recommended diagnosis may convince a pathologist to alter a correct diagnosis into an incorrect diagnosis suggested by an AI-based tool [107]. Pathologists need to be informed that AI tools make predictions based on large amounts of mathematical computations, and that these predictions are prone to error. Adding a confidence score to predictions may help a pathologist focus on cases where an AI tool is “unsure” of its recommendations [107]. On the other hand, an AI tool may also associate a high confidence score with a wrong diagnosis, so it would be advisable for a pathologist to consult other experts when faced with difficult cases. Knowing how these models are trained can help clarify misconceptions about AI [108].

Implementation of LLM tools in healthcare settings are plagued by concerns over cost, computing power, and data privacy. Although GPT4 currently performs better than open source LLMs, there are circumstances where open LLMs are preferable, due to lack of adherence to HIPAA (Health Insurance Portability and Accountability Act) regulations, and other data security concerns [109] involving commercial LLMs. Smaller LLMs, such as the open source FastChat-T5 3B-parameter model can be run locally on a personal computer, while preserving patient privacy [36]. Other small LLM alternatives include TinyLlama [110], Microsoft’s Phi-2, and Google’s Gemma. Small LLMs come with their own set of challenges. These LLMs give less accurate responses, and some can, rarely, produce nonsensical output when given a certain prompt, repeating the same character over and over again.

11. Conclusions

The capabilities of LLMs are growing at a very fast rate, and LLMs will continue to get better. Smaller and open source LLMs are not as powerful, but these are still adequate for many natural language tasks and may be run locally within an institutional server. Eliciting the correct response from LLMs can benefit from carefully crafted prompts, which may involve some trial and error. LLMs, together with MLLMs, can transform many aspects of pathology practice, freeing up time from repetitive, routine, and boring tasks, such as

cell counting, mitosis counting, report generation, information extraction, and medical coding. On a cautionary note, it is important that pathologists understand AI-based tools can make errors while sounding confident and be aware of the phenomenon of automation bias. LLMs can hallucinate, generating fictitious content and references. When in doubt, output from LLMs should be doublechecked with information from reliable sources.

12. Future Directions

It is perceivable that both open and proprietary LLMs/MLLMs will continue to improve. Additional guardrails and training incorporating human feedback is envisioned to improve LLM and MLLM performance, while reducing the incidence of erroneous responses [93]. Some errors produced by LLMs may be attributable to biases and errors present in its training set, so efforts will be undertaken to improve training data quality [111]. More publicly available pathology text and text/image datasets will become available, which will open up opportunities for training more accurate and powerful models. More open source models of varying parameter sizes will be released, including smaller LLMs that can function without a GPU, such as TinyLlama, and the 2B parameter Gemma. Guidelines will be developed to address the ethical aspects of LLM use in scientific writing and other aspects of daily practice [98]. Additional safeguards will be built into LLMs to curtail improper usage.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
2. Yenduri, G.; Srivastava, G.; Maddikunta, P.K.R.; Jhaveri, R.H.; Wang, W.; Vasilakos, A.V.; Gadekallu, T.R. Generative Pre-Trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *arXiv* **2023**, arXiv:2305.10435.
3. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
4. Zeng, K.G.; Dutt, T.; Witowski, J.; Kranthi Kiran, G.V.; Yeung, F.; Kim, M.; Kim, J.; Pleasure, M.; Moczulski, C.; Lopez, L.J.L.; et al. Improving Information Extraction from Pathology Reports Using Named Entity Recognition. *Res. Sq.* **2023**, rs.3.rs-3035772. [[CrossRef](#)]
5. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2021**, arXiv:2010.11929.
6. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
7. Castelvetti, D. Are ChatGPT and AlphaCode Going to Replace Programmers? *Nature* **2022**. [[CrossRef](#)] [[PubMed](#)]
8. Baktash, J.A.; Dawodi, M. Gpt-4: A Review on Advancements and Opportunities in Natural Language Processing. *arXiv* **2023**, arXiv:2305.03195.
9. Geiping, J.; Goldstein, T. Cramming: Training a Language Model on a Single GPU in One Day. In Proceedings of the International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
10. Mitchell, J.R.; Szepietowski, P.; Howard, R.; Reisman, P.; Jones, J.D.; Lewis, P.; Fridley, B.L.; Rollison, D.E. A Question-and-Answer System to Extract Data From Free-Text Oncological Pathology Reports (CancerBERT Network): Development Study. *J. Med. Internet Res.* **2022**, *24*, e27210. [[CrossRef](#)] [[PubMed](#)]
11. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv* **2023**, arXiv:2302.13971.
12. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv* **2023**, arXiv:2307.09288.
13. Yang, H.S.; Wang, F.; Greenblatt, M.B.; Huang, S.X.; Zhang, Y. AI Chatbots in Clinical Laboratory Medicine: Foundations and Trends. *Clin. Chem.* **2023**, *69*, 1238–1246. [[CrossRef](#)]
14. Owens, B. How Nature Readers Are Using ChatGPT. *Nature* **2023**, *615*, 20. [[CrossRef](#)] [[PubMed](#)]
15. Peacock, J.; Austin, A.; Shapiro, M.; Battista, A.; Samuel, A. Accelerating Medical Education with ChatGPT: An Implementation Guide. *MedEdPublish* **2023**, *13*, 64. [[CrossRef](#)] [[PubMed](#)]

16. Cheng, S.-L.; Tsai, S.-J.; Bai, Y.-M.; Ko, C.-H.; Hsu, C.-W.; Yang, F.-C.; Tsai, C.-K.; Tu, Y.-K.; Yang, S.-N.; Tseng, P.-T.; et al. Comparisons of Quality, Correctness, and Similarity Between ChatGPT-Generated and Human-Written Abstracts for Basic Research: Cross-Sectional Study. *J. Med. Internet Res.* **2023**, *25*, e51229. [\[CrossRef\]](#) [\[PubMed\]](#)
17. Wang, A.Y.; Lin, S.; Tran, C.; Homer, R.J.; Wilsdon, D.; Walsh, J.C.; Goebel, E.A.; Sansano, I.; Sonawane, S.; Cockenpot, V.; et al. Assessment of Pathology Domain-Specific Knowledge of ChatGPT and Comparison to Human Performance. *Arch. Pathol. Lab. Med.* **2024**. [\[CrossRef\]](#) [\[PubMed\]](#)
18. 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. [\[CrossRef\]](#)
19. Safranek, C.W.; Sidamon-Eristoff, A.E.; Gilson, A.; Chartash, D. The Role of Large Language Models in Medical Education: Applications and Implications. *JMIR Med. Educ.* **2023**, *9*, e50945. [\[CrossRef\]](#) [\[PubMed\]](#)
20. Ngo, A.; Gupta, S.; Perrine, O.; Reddy, R.; Ershadi, S.; Remick, D. ChatGPT 3.5 Fails to Write Appropriate Multiple Choice Practice Exam Questions. *Acad. Pathol.* **2024**, *11*, 100099. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Sinha, R.K.; Deb Roy, A.; Kumar, N.; Mondal, H. Applicability of ChatGPT in Assisting to Solve Higher Order Problems in Pathology. *Cureus* **2023**, *15*, e35237. [\[CrossRef\]](#)
22. Geetha, S.D.; Khan, A.; Khan, A.; Kannadath, B.S.; Vitkovski, T. Evaluation of ChatGPT Pathology Knowledge Using Board-Style Questions. *Am. J. Clin. Pathol.* **2023**, aqad158. [\[CrossRef\]](#)
23. Munoz-Zuluaga, C.; Zhao, Z.; Wang, F.; Greenblatt, M.B.; Yang, H.S. Assessing the Accuracy and Clinical Utility of ChatGPT in Laboratory Medicine. *Clin. Chem.* **2023**, *69*, 939–940. [\[CrossRef\]](#)
24. Ibrahim, R.B.; Chokkalla, A.K.; Levett, K.; Gustafson, D.; Olayinka, L.; Kumar, S.; Devaraj, S. ChatGPT-Exploring Its Role in Clinical Chemistry. *Ann. Clin. Lab. Sci.* **2023**, *53*, 835–839. [\[PubMed\]](#)
25. Blumenthal, W.; Alimi, T.O.; Jones, S.F.; Jones, D.E.; Rogers, J.D.; Benard, V.B.; Richardson, L.C. Using Informatics to Improve Cancer Surveillance. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 1488–1495. [\[CrossRef\]](#)
26. Choi, H.S.; Song, J.Y.; Shin, K.H.; Chang, J.H.; Jang, B.-S. Developing Prompts from Large Language Model for Extracting Clinical Information from Pathology and Ultrasound Reports in Breast Cancer. *Radiat. Oncol. J.* **2023**, *41*, 209–216. [\[CrossRef\]](#)
27. Schadow, G.; McDonald, C.J. Extracting Structured Information from Free Text Pathology Reports. *AMIA Annu. Symp. Proc.* **2003**, *2003*, 584–588. [\[PubMed\]](#)
28. Cheng, J. Neural Network Assisted Pathology Case Identification. *J. Pathol. Inform.* **2022**, *13*, 100008. [\[CrossRef\]](#)
29. Levy, J.; Vattikonda, N.; Haudenschild, C.; Christensen, B.; Vaickus, L. Comparison of Machine-Learning Algorithms for the Prediction of Current Procedural Terminology (CPT) Codes from Pathology Reports. *J. Pathol. Inform.* **2022**, *13*, 3. [\[CrossRef\]](#)
30. Zhang, X.; Zhang, Y.; Zhang, Q.; Ren, Y.; Qiu, T.; Ma, J.; Sun, Q. Extracting Comprehensive Clinical Information for Breast Cancer Using Deep Learning Methods. *Int. J. Med. Inform.* **2019**, *132*, 103985. [\[CrossRef\]](#)
31. Truhn, D.; Loeffler, C.M.; Müller-Franzes, G.; Nebelung, S.; Hewitt, K.J.; Brandner, S.; Bressem, K.K.; Foersch, S.; Kather, J.N. Extracting Structured Information from Unstructured Histopathology Reports Using Generative Pre-trained Transformer 4 (GPT-4). *J. Pathol.* **2024**, *262*, 310–319. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Unlu, O.; Shin, J.; Mailly, C.J.; Oates, M.F.; Tucci, M.R.; Varugheese, M.; Waghlikar, K.; Wang, F.; Scirica, B.M.; Blood, A.J.; et al. Retrieval Augmented Generation Enabled Generative Pre-Trained Transformer 4 (GPT-4) Performance for Clinical Trial Screening. *medRxiv* **2024**, 2024.02.08.24302376. [\[CrossRef\]](#)
33. Yang, T.; Sucholutsky, I.; Jen, K.-Y.; Schonlau, M. exKidneyBERT: A Language Model for Kidney Transplant Pathology Reports and the Crucial Role of Extended Vocabularies. *PeerJ Comput. Sci.* **2024**, *10*, e1888. [\[CrossRef\]](#)
34. Liu, J.; Gupta, S.; Chen, A.; Wang, C.-K.; Mishra, P.; Dai, H.-J.; Wong, Z.S.-Y.; Jonnagaddala, J. OpenDeID Pipeline for Unstructured Electronic Health Record Text Notes Based on Rules and Transformers: Deidentification Algorithm Development and Validation Study. *J. Med. Internet Res.* **2023**, *25*, e48145. [\[CrossRef\]](#)
35. Santos, T.; Tariq, A.; Das, S.; Vayalapati, K.; Smith, G.H.; Trivedi, H.; Banerjee, I. PathologyBERT—Pre-Trained Vs. A New Transformer Language Model for Pathology Domain. In Proceedings of the AMIA Annual Symposium Proceedings, Washington, DC, USA, 5–9 November 2022.
36. Lee, D.T.; Vaid, A.; Menon, K.M.; Freeman, R.; Matteson, D.S.; Marin, M.P.; Nadkarni, G.N. Development of a Privacy Preserving Large Language Model for Automated Data Extraction from Thyroid Cancer Pathology Reports. *medRxiv* **2023**. [\[CrossRef\]](#)
37. Sushil, M.; Zack, T.; Mandair, D.; Zheng, Z.; Wali, A.; Yu, Y.-N.; Quan, Y.; Butte, A.J. A Comparative Study of Zero-Shot Inference with Large Language Models and Supervised Modeling in Breast Cancer Pathology Classification. *Res. Sq.* **2024**, rs.3.rs-3914899. [\[CrossRef\]](#)
38. Xu, Y.; Zhu, J.-Y.; Chang, E.I.-C.; Lai, M.; Tu, Z. Weakly Supervised Histopathology Cancer Image Segmentation and Classification. *Med. Image Anal.* **2014**, *18*, 591–604. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Vithanage, D.; Yu, P.; Wang, L.; Deng, C. Contextual Word Embedding for Biomedical Knowledge Extraction: A Rapid Review and Case Study. *J. Healthc. Inform. Res.* **2024**, *8*, 158–179. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.
41. Ghinassi, I.; Wang, L.; Newell, C.; Purver, M. Comparing Neural Sentence Encoders for Topic Segmentation across Domains: Not Your Typical Text Similarity Task. *PeerJ Comput. Sci.* **2023**, *9*, e1593. [\[CrossRef\]](#) [\[PubMed\]](#)

42. Gao, S.; Alawad, M.; Young, M.T.; Gounley, J.; Schaefferkoetter, N.; Yoon, H.J.; Wu, X.-C.; Durbin, E.B.; Doherty, J.; Stroup, A.; et al. Limitations of Transformers on Clinical Text Classification. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3596–3607. [\[CrossRef\]](#) [\[PubMed\]](#)
43. Mu, Y.; Tizhoosh, H.R.; Tayebi, R.M.; Ross, C.; Sur, M.; Leber, B.; Campbell, C.J.V. A BERT Model Generates Diagnostically Relevant Semantic Embeddings from Pathology Synopses with Active Learning. *Commun. Med.* **2021**, *1*, 11. [\[CrossRef\]](#)
44. Fijačko, N.; Creber, R.M.; Abella, B.S.; Kocbek, P.; Metličar, Š.; Greif, R.; Štiglic, G. Using Generative Artificial Intelligence in Bibliometric Analysis: 10 Years of Research Trends from the European Resuscitation Congresses. *Resusc. Plus* **2024**, *18*, 100584. [\[CrossRef\]](#)
45. Kefeli, J.; Tatonetti, N. Benchmark Pathology Report Text Corpus with Cancer Type Classification. *medRxiv* **2023**, 2023.08.03.23293618. [\[CrossRef\]](#)
46. Kefeli, J.; Tatonetti, N. Generalizable and Automated Classification of TNM Stage from Pathology Reports with External Validation. *medRxiv* **2023**, 2023.06.26.23291912. [\[CrossRef\]](#)
47. Fang, L.; Wang, K. Multi-Label Topic Classification for COVID-19 Literature with Bioformer. *arXiv* **2022**, arXiv:2204.06758v1.
48. Zhou, Z. Evaluation of ChatGPT's Capabilities in Medical Report Generation. *Cureus* **2023**, *15*, e37589. [\[CrossRef\]](#)
49. Shah, A.; Wahood, S.; Guermazi, D.; Brem, C.E.; Saliba, E. Skin and Syntax: Large Language Models in Dermatopathology. *Dermatopathology* **2024**, *11*, 101–111. [\[CrossRef\]](#) [\[PubMed\]](#)
50. Hart, S.N.; Hoffman, N.G.; Gershkovich, P.; Christenson, C.; McClintock, D.S.; Miller, L.J.; Jackups, R.; Azimi, V.; Spies, N.; Brodsky, V. Organizational Preparedness for the Use of Large Language Models in Pathology Informatics. *J. Pathol. Inform.* **2023**, *14*, 100338. [\[CrossRef\]](#)
51. Grewal, H.; Dhillon, G.; Monga, V.; Sharma, P.; Buddhavarapu, V.S.; Sidhu, G.; Kashyap, R. Radiology Gets Chatty: The ChatGPT Saga Unfolds. *Cureus* **2023**, *15*, e40135. [\[CrossRef\]](#) [\[PubMed\]](#)
52. Russe, M.F.; Reiser, M.; Bamberg, F.; Rau, A. Improving the Use of LLMs in Radiology through Prompt Engineering: From Precision Prompts to Zero-Shot Learning. *Rofo* **2024**. [\[CrossRef\]](#) [\[PubMed\]](#)
53. Stokel-Walker, C. ChatGPT Listed as Author on Research Papers: Many Scientists Disapprove. *Nature* **2023**, *613*, 620–621. [\[CrossRef\]](#) [\[PubMed\]](#)
54. Briganti, G. How ChatGPT Works: A Mini Review. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 1565–1569. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Bhattacharyya, M.; Miller, V.M.; Bhattacharyya, D.; Miller, L.E. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus* **2023**, *15*, e39238. [\[CrossRef\]](#) [\[PubMed\]](#)
56. Naik, H.R.; Prather, A.D.; Gurda, G.T. Synchronous Bilateral Breast Cancer: A Case Report Piloting and Evaluating the Implementation of the AI-Powered Large Language Model (LLM) ChatGPT. *Cureus* **2023**, *15*, e37587. [\[CrossRef\]](#)
57. Gao, C.A.; Howard, F.M.; Markov, N.S.; Dyer, E.C.; Ramesh, S.; Luo, Y.; Pearson, A.T. Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers. *NPJ Digit. Med.* **2023**, *6*, 75. [\[CrossRef\]](#)
58. Mojadeddi, Z.M.; Rosenberg, J. The Impact of AI and ChatGPT on Research Reporting. *N. Z. Med. J.* **2023**, *136*, 60–64.
59. Rashidi, H.H.; Fennell, B.D.; Albahra, S.; Hu, B.; Gorbett, T. The ChatGPT Conundrum: Human-Generated Scientific Manuscripts Misidentified as AI Creations by AI Text Detection Tool. *J. Pathol. Inform.* **2023**, *14*, 100342. [\[CrossRef\]](#) [\[PubMed\]](#)
60. Else, H. Abstracts Written by ChatGPT Fool Scientists. *Nature* **2023**, *613*, 423. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Polak, M.P.; Morgan, D. Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering. *Nat. Commun.* **2024**, *15*, 1569. [\[CrossRef\]](#) [\[PubMed\]](#)
62. Leypold, T.; Schäfer, B.; Boos, A.; Beier, J.P. Can AI Think Like a Plastic Surgeon? Evaluating GPT-4's Clinical Judgment in Reconstructive Procedures of the Upper Extremity. *Plast. Reconstr. Surg. Glob. Open* **2023**, *11*, e5471. [\[CrossRef\]](#) [\[PubMed\]](#)
63. Wang, L.; Chen, X.; Deng, X.; Wen, H.; You, M.; Liu, W.; Li, Q.; Li, J. Prompt Engineering in Consistency and Reliability with the Evidence-Based Guideline for LLMs. *NPJ Digit. Med.* **2024**, *7*, 41. [\[CrossRef\]](#) [\[PubMed\]](#)
64. Cheng, S.-W.; Chang, C.-W.; Chang, W.-J.; Wang, H.-W.; Liang, C.-S.; Kishimoto, T.; Chang, J.P.-C.; Kuo, J.S.; Su, K.-P. The Now and Future of ChatGPT and GPT in Psychiatry. *Psychiatry Clin. Neurosci.* **2023**, *77*, 592–596. [\[CrossRef\]](#) [\[PubMed\]](#)
65. Meskó, B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. *J. Med. Internet Res.* **2023**, *25*, e50638. [\[CrossRef\]](#)
66. Wei, W.I.; Leung, C.L.K.; Tang, A.; McNeil, E.B.; Wong, S.Y.S.; Kwok, K.O. Extracting Symptoms from Free-Text Responses Using ChatGPT among COVID-19 Cases in Hong Kong. *Clin. Microbiol. Infect.* **2024**, *30*, 142.e1–142.e3. [\[CrossRef\]](#) [\[PubMed\]](#)
67. Ge, J.; Li, M.; Delk, M.B.; Lai, J.C. A Comparison of Large Language Model versus Manual Chart Review for Extraction of Data Elements from the Electronic Health Record. *medRxiv* **2023**, 2023.08.31.23294924. [\[CrossRef\]](#)
68. Kojima, T.; Gu, S.S.; Reid, M.; Matsuo, Y.; Iwasawa, Y. Large Language Models Are Zero-Shot Reasoners. *Adv. Neural Inf. Process. Syst.* **2023**, *35*, 22199–22213.
69. Zhang, X.; Talukdar, N.; Vemulapalli, S.; Ahn, S.; Wang, J.; Meng, H.; Murtaza, S.M.B.; Leshchiner, D.; Dave, A.A.; Joseph, D.F.; et al. Comparison of Prompt Engineering and Fine-Tuning Strategies in Large Language Models in the Classification of Clinical Notes. *medRxiv* **2024**, 2024.02.07.24302444. [\[CrossRef\]](#)
70. Abdullahi, T.; Singh, R.; Eickhoff, C. Learning to Make Rare and Complex Diagnoses With Generative AI Assistance: Qualitative Study of Popular Large Language Models. *JMIR Med. Educ.* **2024**, *10*, e51391. [\[CrossRef\]](#)
71. Agarwal, A.; Chan, A.; Chandel, S.; Jang, J.; Miller, S.; Moghaddam, R.Z.; Mohylevskyy, Y.; Sundaresan, N.; Tufano, M. Copilot Evaluation Harness: Evaluating LLM-Guided Software Programming. *arXiv* **2024**, arXiv:2402.14261.

72. Coello, C.E.A.; Alimam, M.N.; Kouatly, R. Effectiveness of ChatGPT in Coding: A Comparative Analysis of Popular Large Language Models. *Digital* **2024**, *4*, 114–125. [[CrossRef](#)]
73. Hellas, A.; Leinonen, J.; Sarsa, S.; Koutchme, C.; Kujanpää, L.; Sorva, J. Exploring the Responses of Large Language Models to Beginner Programmers' Help Requests. In Proceedings of the 2023 ACM Conference on International Computing Education Research V.1, Chicago, IL, USA, 7–11 August 2023; pp. 93–105.
74. King, M.R.; Abdulrahman, A.M.; Petrovic, M.I.; Poley, P.L.; Hall, S.P.; Kulapatana, S.; Lamantia, Z.E. Incorporation of ChatGPT and Other Large Language Models into a Graduate Level Computational Bioengineering Course. *Cell. Mol. Bioeng.* **2024**, *17*, 1–6. [[CrossRef](#)] [[PubMed](#)]
75. Poldrack, R.A.; Lu, T.; Beguš, G. AI-Assisted Coding: Experiments with GPT-4. *arXiv* **2023**, arXiv:2304.13187.
76. Yan, W.; Tian, Y.; Li, Y.; Chen, Q.; Wang, W. CodeTransOcean: A Comprehensive Multilingual Benchmark for Code Translation. *arXiv* **2023**, arXiv:2310.04951.
77. Rao, D. The Urgent Need for Healthcare Workforce Upskilling and Ethical Considerations in the Era of AI-Assisted Medicine. *Indian J. Otolaryngol. Head Neck Surg.* **2023**, *75*, 2638–2639. [[CrossRef](#)] [[PubMed](#)]
78. Jiang, A.Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D.S.; Casas, D.d.l.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; et al. Mistral 7B. *arXiv* **2023**, arXiv:2310.06825.
79. Yang, W.-H.; Yang, Y.-J.; Chen, T.-J. ChatGPT's Innovative Application in Blood Morphology Recognition. *J. Chin. Med. Assoc.* **2024**. [[CrossRef](#)] [[PubMed](#)]
80. Sallam, M.; Al-Salahat, K.; Al-Ajlouni, E. ChatGPT Performance in Diagnostic Clinical Microbiology Laboratory-Oriented Case Scenarios. *Cureus* **2023**, *15*, e50629. [[CrossRef](#)]
81. Spies, N.C.; Hubler, Z.; Roper, S.M.; Omosule, C.L.; Senter-Zapata, M.; Roemmich, B.L.; Brown, H.M.; Gimple, R.; Farnsworth, C.W. GPT-4 Underperforms Experts in Detecting IV Fluid Contamination. *J. Appl. Lab. Med.* **2023**, *8*, 1092–1100. [[CrossRef](#)]
82. Kumari, A.; Kumari, A.; Singh, A.; Singh, S.K.; Juhi, A.; Dhanvijay, A.K.D.; Pinjar, M.J.; Mondal, H. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* **2023**, *15*, e43861. [[CrossRef](#)] [[PubMed](#)]
83. Stephens, L.D. ChatGPT in Transfusion Medicine: A New Frontier for Patients? *Transfusion* **2023**, *63*, 1110–1112. [[CrossRef](#)] [[PubMed](#)]
84. Hurley, N.C.; Schroeder, K.M.; Hess, A.S. Would Doctors Dream of Electric Blood Bankers? Large Language Model-Based Artificial Intelligence Performs Well in Many Aspects of Transfusion Medicine. *Transfusion* **2023**, *63*, 1833–1840. [[CrossRef](#)] [[PubMed](#)]
85. Wu, J.; Gan, W.; Chen, Z.; Wan, S.; Yu, P.S. Multimodal Large Language Models: A Survey. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023.
86. Zang, Y.; Li, W.; Han, J.; Zhou, K.; Loy, C.C. Contextual Object Detection with Multimodal Large Language Models. *arXiv* **2023**, arXiv:2305.18279.
87. Laohawetwanit, T.; Namboonlue, C.; Apornvirat, S. Accuracy of GPT-4 in Histopathological Image Detection and Classification of Colorectal Adenomas. *J. Clin. Pathol.* **2024**, jcp-2023-209304. [[CrossRef](#)]
88. Sievert, M.; Aubreville, M.; Mueller, S.K.; Eckstein, M.; Breininger, K.; Iro, H.; Goncalves, M. Diagnosis of Malignancy in Oropharyngeal Confocal Laser Endomicroscopy Using GPT 4.0 with Vision. *Eur. Arch. Otorhinolaryngol.* **2024**, *281*, 2115–2122. [[CrossRef](#)] [[PubMed](#)]
89. Tsuneki, M.; Kanavati, F. Inference of Captions from Histopathological Patches. In Proceedings of the International Conference on Medical Imaging with Deep Learning, Zurich, Switzerland, 6–8 July 2022.
90. Sengupta, S.; Brown, D.E. Automatic Report Generation for Histopathology Images Using Pre-Trained Vision Transformers and BERT. *arXiv* **2023**, arXiv:2312.01435.
91. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021.
92. Sun, Y.; Zhu, C.; Zheng, S.; Zhang, K.; Sun, L.; Shui, Z.; Zhang, Y.; Li, H.; Yang, L. PathAsst: A Generative Foundation AI Assistant Towards Artificial Intelligence of Pathology. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 20–27 February 2024.
93. Lu, M.Y.; Chen, B.; Williamson, D.F.K.; Chen, R.J.; Ikamura, K.; Gerber, G.; Liang, I.; Le, L.P.; Ding, T.; Parwani, A.V.; et al. A Foundational Multimodal Vision Language AI Assistant for Human Pathology. *arXiv* **2023**, arXiv:2312.07814.
94. Zhang, Y.; Gao, J.; Tan, Z.; Zhou, L.; Ding, K.; Zhou, M.; Zhang, S.; Wang, D. Data-Centric Foundation Models in Computational Healthcare: A Survey. *arXiv* **2024**, arXiv:2401.02458.
95. Shafi, S.; Parwani, A.V. Artificial Intelligence in Diagnostic Pathology. *Diagn. Pathol.* **2023**, *18*, 109. [[CrossRef](#)]
96. Berbis, M.A.; McClintock, D.S.; Bychkov, A.; Van Der Laak, J.; Pantanowitz, L.; Lennerz, J.K.; Cheng, J.Y.; Delahunt, B.; Egevad, L.; Eloy, C.; et al. Computational Pathology in 2030: A Delphi Study Forecasting the Role of AI in Pathology within the next Decade. *eBioMedicine* **2023**, *88*, 104427. [[CrossRef](#)]
97. Yu, H. The Application and Challenges of ChatGPT in Educational Transformation: New Demands for Teachers' Roles. *Heliyon* **2024**, *10*, e24289. [[CrossRef](#)] [[PubMed](#)]

98. Schukow, C.; Smith, S.C.; Landgrebe, E.; Parasuraman, S.; Folaranmi, O.O.; Paner, G.P.; Amin, M.B. Application of ChatGPT in Routine Diagnostic Pathology: Promises, Pitfalls, and Potential Future Directions. *Adv. Anat. Pathol.* **2024**, *31*, 15–21. [[CrossRef](#)] [[PubMed](#)]
99. Gutiérrez-Cirlos, C.; Carrillo-Pérez, D.L.; Bermúdez-González, J.L.; Hidrogo-Montemayor, I.; Carrillo-Esper, R.; Sánchez-Mendiola, M. ChatGPT: Opportunities and Risks in the Fields of Medical Care, Teaching, and Research. *Gac. Med. Mex.* **2023**, *159*, 372–379. [[CrossRef](#)] [[PubMed](#)]
100. Ge, J.; Sun, S.; Owens, J.; Galvez, V.; Gologorskaya, O.; Lai, J.C.; Pletcher, M.J.; Lai, K. Development of a Liver Disease-Specific Large Language Model Chat Interface Using Retrieval Augmented Generation. *medRxiv* **2023**, 2023.11.10.23298364. [[CrossRef](#)]
101. Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Guo, Q.; Wang, M.; et al. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv* **2024**, arXiv:2312.10997.
102. Wang, Y.; Singh, L. Adding Guardrails to Advanced Chatbots. *arXiv* **2023**, arXiv:2306.07500.
103. Fogo, A.B.; Kronbichler, A.; Bajema, I.M. AI's Threat to the Medical Profession. *JAMA* **2024**, *331*, 471. [[CrossRef](#)] [[PubMed](#)]
104. Cheng, J.Y.; Abel, J.T.; Balis, U.G.J.; McClintock, D.S.; Pantanowitz, L. Challenges in the Development, Deployment, and Regulation of Artificial Intelligence in Anatomic Pathology. *Am. J. Pathol.* **2021**, *191*, 1684–1692. [[CrossRef](#)] [[PubMed](#)]
105. Nakagawa, K.; Moukheiber, L.; Celi, L.A.; Patel, M.; Mahmood, F.; Gondim, D.; Hogarth, M.; Levenson, R. AI in Pathology: What Could Possibly Go Wrong? *Semin. Diagn. Pathol.* **2023**, *40*, 100–108. [[CrossRef](#)] [[PubMed](#)]
106. Nguyen, T. ChatGPT in Medical Education: A Precursor for Automation Bias? *JMIR Med. Educ.* **2024**, *10*, e50174. [[CrossRef](#)]
107. Evans, H.; Snead, D. Why Do Errors Arise in Artificial Intelligence Diagnostic Tools in Histopathology and How Can We Minimize Them? *Histopathology* **2024**, *84*, 279–287. [[CrossRef](#)]
108. Emmert-Streib, F.; Yli-Harja, O.; Dehmer, M. Artificial Intelligence: A Clarification of Misconceptions, Myths and Desired Status. *Front. Artif. Intell.* **2020**, *3*, 524339. [[CrossRef](#)] [[PubMed](#)]
109. Gordon, E.R.; Trager, M.H.; Kontos, D.; Weng, C.; Geskin, L.J.; Dugdale, L.S.; Samie, F.H. Ethical Considerations for Artificial Intelligence in Dermatology: A Scoping Review. *Br. J. Dermatol.* **2024**, ljae040. [[CrossRef](#)] [[PubMed](#)]
110. Zhang, P.; Zeng, G.; Wang, T.; Lu, W. TinyLlama: An Open-Source Small Language Model. *arXiv* **2024**, arXiv:2401.02385.
111. Ullah, E.; Parwani, A.; Baig, M.M.; Singh, R. Challenges and Barriers of Using Large Language Models (LLM) Such as ChatGPT for Diagnostic Medicine with a Focus on Digital Pathology—A Recent Scoping Review. *Diagn. Pathol.* **2024**, *19*, 43. [[CrossRef](#)] [[PubMed](#)]

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.