1. **How will you use the skills that you've developed during this research project in your future degree programme?**

In this research project I developed a set of skills that I know will be very useful for my future degree. First, I learned how to take a broad and complicated subject and build it into a clear, logical structure. When I was writing my essay, I had to connect evidence from medicine, computer science, and ethics, and that process really trained me to organize information in a way that makes sense.

I also became much more critical about evaluation. Instead of just looking at accuracy, I had to think about whether the results were safe, reliable, and appropriate. That gave me the habit of questioning numbers and always considering the human impact, which is something I want to carry into future projects.

Finally, I practised communicating across disciplines. I had to explain technical issues in a way that could be understood by non-specialists, and that improved my ability to present ideas clearly. Altogether, these skills — structured thinking, critical evaluation, safety awareness, and communication — will directly shape how I approach my graduate studies.

1. **When you were doing your research, did you learn anything that surprised you? Did you learn anything surprising?**

Yes, I did learn something that really surprised me. At the beginning of my research, I assumed that large language models in medicine were mainly limited by accuracy — whether they got the right answer or not. But what really surprised me was that accuracy alone is not enough. In fact, even when a model is correct, doctors may not trust it if it cannot show its reasoning clearly. I found studies where the model performed on par with experienced clinicians, yet its value was questioned because of hallucinations, lack of transparency, or because it sometimes gave overly confident wrong answers.

Another thing that surprised me was how much difference human-AI collaboration can make. Instead of replacing doctors, the best results often came when the model provided suggestions and the clinician made the final judgment. That shifted my perspective — I realized the real opportunity is not just building a “perfect” model, but creating a safe partnership between humans and AI. That insight was unexpected, but it completely changed the way I think about medical AI.

1. **of all the sources that you looked at, which source would you say had the most impact on the direction of your research?**

Of all the sources I read, the one that had the biggest impact on my research was the Nature paper on using large language models for differential diagnosis. What stood out to me was that the authors didn’t just test whether the model could give the right answer; they compared its reasoning and decision-making directly with practising doctors. That made me realize that the real challenge is not only performance on benchmarks but also whether the model can support clinical decision-making in a safe and trustworthy way.

This source shaped the direction of my essay because it pushed me to look beyond simple accuracy. I started to focus more on issues like transparency, hallucinations, and the role of human oversight. It also influenced my overall stance, that AI should be developed to assist doctors rather than to replace them. Without that paper, my essay might have stayed at the level of technical comparisons, but with it, I was able to build a much deeper and more balanced argument.

1. **Are there any limitations to this source?**

That’s a really good question. The Nature paper on MedFound is very recent — it was only published this March — so in a way, it doesn’t yet have the kind of long-term limitations you might see in older studies. It is one of the first large-scale evaluations of an LLM specifically designed for medical diagnosis, so it sets a strong benchmark. Because of that, I wouldn’t say the paper itself has obvious flaws in design; rather, its main limitation is that it is still only an early step.

For example, the study evaluates MedFound in controlled settings, but it doesn’t yet tell us how the model performs in day-to-day clinical environments with all the unpredictability of real patients. It also focuses heavily on diagnostic accuracy, and while that is important, it cannot fully capture issues like patient trust, doctor–AI collaboration, or long-term safety. Another limitation is that because it is so new, it hasn’t yet been replicated or validated by independent groups, so we don’t know how robust the findings are across different hospitals and healthcare systems.

So overall, the limitation is not that the paper is weak, but that it is still very early. It gives us exciting results and a direction forward, but we need more time, more replication, and more practical testing before we can fully judge its clinical value.

1. **Your conclusion states \_\_\_\_\_\_. Can you explain the reasons you reached this conclusion?**

In my conclusion I argued that while large language models like MedFound can reach levels of diagnostic accuracy comparable to doctors, their real value lies in supporting clinicians rather than replacing them. I reached this conclusion for a few reasons. First, throughout my research I saw consistent evidence that accuracy alone is not enough. Even when a model gives the right answer, problems like hallucinations, lack of explainability, or overconfidence make it risky to use without human oversight. Second, studies showed that when AI and doctors work together, the overall quality of reasoning and appropriateness of care actually improves. That convinced me that the best direction is collaboration, not substitution. And finally, I was also influenced by ethical concerns — accountability, patient trust, and the need for transparency. All of these factors pointed me towards the same conclusion: AI should be designed to strengthen clinical decision-making, but the final responsibility and judgment must remain with human doctors.

| **Term** | **Simple Meaning** |
| --- | --- |
| Large Language Model (LLM) | A type of AI trained on huge amounts of text, so it can understand and generate human-like language — like a “talking library.” |
| MedFound | A newly developed medical AI system, designed to help doctors make diagnoses. |
| Diagnostic effectiveness | Not just whether the answer is right, but also whether the reasoning is clear and the patient is kept safe. |
| Top-k / Top-3 accuracy | Whether the correct diagnosis appears in the model’s top guesses. For example, “Top-3” means the right answer is among the three best options. |
| In-distribution vs. Out-of-distribution | Testing on “familiar data” the model has seen before versus testing on “new hospitals or new environments” it has never seen. |
| Hallucination | When the AI makes up medical details that aren’t real but presents them as if they are true. |
| Explainability | How easy it is for people to understand the steps behind the AI’s answer. |
| Accountability mechanisms | Who is responsible if the AI is wrong, and whether there are systems to ensure safe use. |
| Confidence calibration | Whether the AI shows the right level of certainty — not sounding too confident when it’s actually unsure. |