No More Pencils No More Books: Capabilities of Generative AI on Irish and UK Computer Science School Leaving Examinations

Joyce Mahon
University College Dublin
Dublin, Ireland
joyce.mahon1@ucdconnect.ie

Brian Mac Namee
University College Dublin
Dublin, Ireland
brian.macnamee@ucd.ie

Brett A. Becker
University College Dublin
Dublin, Ireland
brett.becker@ucd.ie

ABSTRACT

We investigate the capabilities of ChatGPT (GPT-4) on second-level (high-school) computer science examinations: the UK A-Level and Irish Leaving Certificate. Both are national, government-set / approved, and centrally assessed examinations. We also evaluate performance differences in exams made publicly available before and after the ChatGPT knowledge cutoff date, and investigate what types of question ChatGPT struggles with.

We find that ChatGPT is capable of achieving very high marks on both exams and that the performance difference before and after the knowledge cutoff date are minimal. We also observe that ChatGPT struggles with questions involving symbols or images, which can be mitigated when in-text information 'fills in the gaps'. Additionally, GPT-4 performance can be negatively impacted when an initial inaccurate answer leads to further inaccuracies in subsequent parts of the same question. Finally, the element of choice on the Leaving Certificate is a significant advantage in achieving a high grade. Notably, there are minimal occurrences of hallucinations in answers and few errors in solutions not involving images.

These results reveal several strengths and weaknesses of these exams in terms of how generative AI performs on them and have implications for exam design, the construction of marking schemes, and could also shift the focus of what is examined and how.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; • **Social and professional topics → Computer science education**; **Computing education**; **K-12 education**.

KEYWORDS

A-Level; Artificial Intelligence; ChatGPT; examinations; Generative AI; GPT-4; high school; Ireland; K-12; Leaving Certificate; LCCS; school; second-level; UK

ACM Reference Format:
Joyce Mahon, Brian Mac Namee, and Brett A. Becker. 2023. No More Pencils No More Books: Capabilities of Generative AI on Irish and UK Computer Science School Leaving Examinations. In *The United Kingdom and Ireland Computing Education Research (UKICER) conference (UKICER 2023), September 07–08, 2023, Swansea, Wales Uk.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3610969.3610982

UKICER 2023, September 07–08, 2023, Swansea, Wales Uk
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0876-3/23/09.
https://doi.org/10.1145/3610969.3610982

1 INTRODUCTION

In recent years artificial intelligence (AI) and natural language processing (NLP) have seen impressive advances, perhaps only surpassed by societal interest in a product of these advances: large language models (LLMs) and what has become known as Generative AI. Only eight months since its public release, ChatGPT has captured the world's imagination, including predictable speculation about seismic economic changes, the replacement of millions of jobs, and the ubiquitous predictions of 'the robots taking over' and humankind's extinction [42]. More certain are the numerous fields where Generative AI has been used with some impressive results including computer programming where LLMs have delivered on (not always perfect) AI code generation. In this, natural language prompts serve as input to a model, and code is returned. This is potentially revolutionary for computing education, particularly as programming is essential to the study of computing, yet also presents many challenges [3, 29].

GPT-4 is the latest model from OpenAI which like earlier models greatly improves upon its predecessor [9] – although the age of greater performance being achieved by training set scaling has been flagged as ending by AI leaders including the CEO of Open AI [26]. In this study, we assess the capabilities of ChatGPT (GPT-4) on two second-level school leaving exams: The Cambridge Assessment International Education (CIE) A-Level Computer Science (A-Level CS) exam used in most of the UK and several other countries, and the Irish Leaving Certificate Computer Science (LCCS) exam.

In this work our interest goes beyond gauging the performance of ChatGPT particularly as related work indicates that ChatGPT should perform quite well [9] – although how well has not yet been measured. There are differences between these exams that go beyond content. For instance, the LCCS features student choice in the form of optional questions, unlike the A-Level CS. Additionally, 'cascading' questions (where sub-questions build upon each other) feature more heavily in the A-Level CS. We are also interested in how the OpenAI knowledge cutoff date of September 2021 affects the performance of ChatGPT. This may lead to insight on the true 'capability transfer' of ChatGPT, beyond simple 'memorisation'. We aim to answer the following research questions:

RQ1 How does ChatGPT perform on the A-Level Computer Science and the Leaving Certificate Computer Science examinations?
RQ2 What impact do optional and cascading questions have on ChatGPT's performance?

RQ3 Does ChatGPT's performance differ between exams made available before and after the current knowledge cutoff date of September 2021?

2 BACKGROUND

The A-Level is offered in many subjects and is used as a school leaving qualification by educational bodies in England, Wales and Northern Ireland[1], and well over a dozen other countries including India and Pakistan. Performance on the A-Levels is a large factor in determining access to university education for school leavers. The Irish Leaving Certificate is also a subject-based qualification that for most school leavers is the main factor in university entrance.[2]

A-Level and LCCS exams are set and assessed centrally by government approved bodies, and both are typically studied during the final two years of second-level schooling, equivalent to grades 11-12 (the final two years of high school) in the US and many other regions – approximate student ages of 17–18 years [7, 21]. In both systems, the Computer Science subject is optional to students who may choose these from a large selection of other subjects.

There are a number of examining boards that offer different A-Level CS curricula (for example, AQA [1] and OCR [10]). We selected the CIE examining board, as over 10,000 schools in dozens of countries utilise over 55 subjects offered by CIE at AS[3] and A-Level [19]. The CIE A-Level CS is examined through four papers: Paper 1 - Theory Fundamentals; Paper 2 - Fundamental Problem-solving and Programming Skills; Paper 3 - Advanced Theory; and Paper 4 - Practical. Candidates are required to use Java, Visual Basic or Python (all in console mode) [20]. These four papers are completed over 7.5 hours.

The LCCS exam consists of two papers, and we focus on the Higher Level (HL) exam[4] as most students take this (86% in 2022, 91% in 2021) [11] which attracts more points that count towards university entry for a given grade compared to the Ordinary Level exam. The first paper consists of two sections (A and B). Section A features 'short-answer' questions and Section B features 'long answer' questions. The second paper, which includes just one section (C), features programming questions. These two papers count towards 70% of a student's final mark and are allocated 2.5 hours. The remaining 30% is determined by a practical project [4]. This component is completed in advance of the LCCS papers over a period of months under the supervision of the teacher [35].

Comprehensive accounts of the Computer Science subject in the UK are provided by Brown et al. [7, 8] and in Ireland by Faherty et al [21] and Becker et al. [4].

3 RELATED WORK

Despite a spate of recent activity, much related work on LLMs is not yet published in archival form – a lot of the literature is currently on arXiv. In this section we review existing work directly related to our research questions, noting that at a high level it is known that AI presents many opportunities and challenges in computing education [5], with implications for both teachers and students [30, 39].

Finnie-Ansley et al. evaluated OpenAI Codex (a GPT-3-based model tuned to generate computer programs) on university-level introductory programming (often referred to as CS1 [24]) exams and found that it performed amongst the top quartile of human students that sat the same exam [22]. A year later the same authors reported similar performance on Data Structures and Algorithms (also known as CS2 [24]) exams [23]. Surprisingly versatile, LLMs have also shown utility in overcoming traditional barriers to learning programming such as programming error messages [27], and at alternative question styles such as Parsons problems [40].

The GPT-4 technical report by OpenAI states that GPT-4 "exhibits human-level performance on the majority of these professional and academic exams" [37], where 'these' include: several US Advanced Placement (AP) exams, US Standard Aptitude Test (SAT) exams, US Graduate Record Examinations (GRE), the US Law School Admission Test (LSAT), and the Uniform Bar Exam, a US law examination adopted by 39 US states. GPT-4 demonstrated a substantial improvement over GPT-3.5 on most of these, a selection of which are provided in Table 1.

Table 1: GPT-3.5 vs GPT-4 on several exams from [37].

Exam	GPT-3.5 percentile	GPT-4 percentile
Uniform Bar Exam	~10th	~90th
LSAT	~40th	~83rd
GRE Quantitative	~25th	~62nd
AP Macroeconomics	~33rd	~84th

Researchers have explored where GPT models struggle, including on programming questions. Dziri et al. [18] found that transformer models such as GPT-*x* perform better on single step reasoning tasks and face challenges when it comes to effectively combining multiple steps for complex problems where error propagation often occurs. Finnie-Ansley et al. [23] found that Codex struggled with more complex prompts supporting the observation that performance tends to decrease exponentially as the number of basic building blocks in a question increases - first made by Chen et al. [12] in the paper that introduced Codex.

Surprisingly we could not find empirical work on the effect of the GPT knowledge cutoff date although Bubeck et al. note that both GPT-4 and GPT-3.5 can suffer from effects of the cutoff date including using out-of-date information in answers and a refusal to answer other questions [9]. The GPT-4 technical report [37] makes no useful mention of the cutoff date, although user forums report inconsistent results when probing the topic[5]. There are GPT-4 plugins that attempt to overcome this, however we did not utilise these as explained in Section 4.

4 METHOD

For the A-Level CS exams we used 2021 CIE A-Level Computer Science past papers (9618/11, /21, /31 and /41) and their marking schemes [25]. These are the only CIE papers publicly available without restriction. For the LCCS we experimented with the 2021 papers

[1] Scotland uses a different examination, the Scottish Advanced Higher.

[2] Other factors such as university/program choice also affect access.

[3] AS levels are similar to the first year of an A-level course.

[4] The LCCS is offered in two levels - 'Ordinary' and 'Higher'.

[5] community.openai.com/t/knowledge-cutoff-date-of-september-2021/66215

(for a temporal comparison with the available A-Level papers), and the 2022 papers (for pre/post cutoff date comparisons). The first paper contains “Section A&B / HL”, and the second contains “Section C /HL” [34].

All experiments used GPT-4 (May 12 2023 version) which we refer to as ‘ChatGPT’ for the remainder of this paper. ChatGPT does offer browser access (allowing ChatGPT to search the ‘live’ internet) as well as several plugins. However we did not utilise these as they are in beta and could be inconsistent making our results more difficult to replicate.

Each sub-question of each exam paper was provided verbatim as a prompt to ChatGPT without any rephrasing or clarification and responses were saved. The solutions were assessed by one author who is a second-level teacher using the official marking schemes, subsequently checked by two university CS lecturers (the other two authors) acting as verifiers. For programming aspects we utilised Python as it is the only language used for the LCCS in 2021 [16] and 2022 [17]. We therefore chose to focus on Python for the A-Level CS to allow direct comparison. We have made all prompts (questions) and ChatGPT responses (answers) available.[6]

5 RESULTS

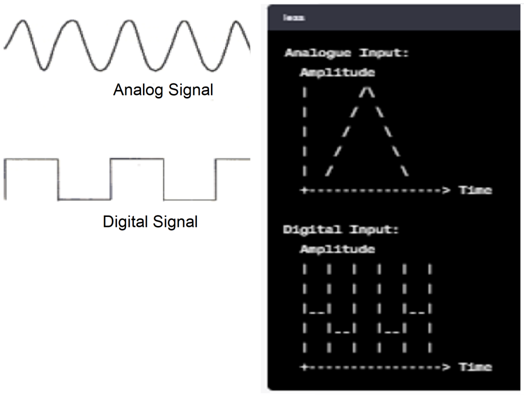
5.1 RQ1: Performance on A-Level CS & LCCS

Our first research question was: *How does ChatGPT perform on the A-Level Computer Science and the Leaving Certificate Computer Science examinations?* To answer this we focused on the 2021 LCCS HL papers and the 2021 CIE A-Level CS papers.

Grades. On the 2021 LCCS papers, ChatGPT achieved 100%, resulting in a ‘H1’ grade. For comparison, only 14% of students achieved a H1 in 2021 [11]. As this exam is 70% of the overall grade for this subject, a grade of H3 would be achieved before even considering the LCCS coursework. In the 2021 CIE A-Level ChatGPT achieved 80.3% resulting in an ‘A’ grade. In 2021 56% of students received an A or A* grade [32] where A* > A. The fact that ChatGPT did much better on the LCCS is largely due to the fact that unlike the A-Level there are many optional questions (see Section 5.2 for details).

Correct Answers. The 2021 LCCS consists of 16 questions in three Sections (A, B and C), broken down into a total of 62 sub-questions. As expected, the majority of these sub-questions (55/62 or 89%) were answered correctly. However, there were some noteworthy unanticipated answers:

- **Symbols in questions.** For Question 14 b) iv) in Section B, the symbol ≠ was not recognised, and it appeared as a placeholder or unknown character. Additionally, round brackets were not recognised in part (v). Despite this, the correct answer was given for both cases.
- **Diagrams in questions.** For Question 8 b) in Section A and Question 13 a) in Section B, ChatGPT was unable to interpret the accompanying diagrams. However, there was enough contextual information provided in the question to return the correct answers.



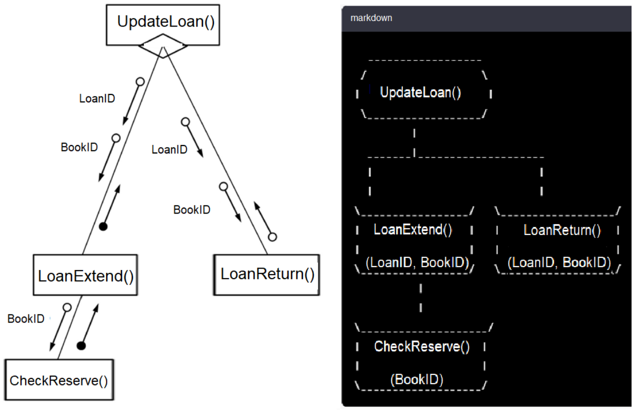
```
Analogue Input:
  Amplitude
  |      /\
  |     /  \
  |    /    \
  |   /      \
  |  /        \
  +-----------------> Time

Digital Input:
  Amplitude
  |  |  |  |  |  |
  |  |  |  |  |  |
  |__|  |  |  |__|
  |  |__|  |__|  |
  |  |  |  |  |  |
  +-----------------> Time
```

Figure 1: Question 15 a) iii), LCCS 2021 HL paper. The ‘diagram showing the analogue and digital output’ from the marking scheme is shown (left – white background), beside the ChatGPT output (right – black background).

- **Diagrams in solutions.** For Question 15 a) iii) in Section B ChatGPT successfully drew a diagram with analogue and digital output - see Figure 1.

The A-Level CS 2021 consists of 27 mandatory questions contained within 4 examination papers, within which are 96 sub-questions. There was a slightly smaller proportion of sub-questions answered correctly by ChatGPT (84/96 or 88%). The following observations were notable:

- **Diagrams in questions.** For Q2a) of the 9618/11 paper, a diagram required matching utility software types to their definitions. Despite ChatGPT not being able to view the information diagrammatically, this was correctly answered as the text contained within could be copied and entered into the prompt.
- **Diagrams in solutions.** For Question 2 a) in the 9618/21 paper, ChatGPT successfully drew a structure chart - see Figure 2.



```
  ------------------
 /                  \
|   UpdateLoan()     |
 \__________________/
          |
   _______|___________________
  |                           |
  |                           |
\________________/   \________________/
|  LoanExtend()  |   |  LoanReturn()  |
|                |   |                |
\(LoanID, BookID)/   \(LoanID, BookID)/
         |
         |
\__________________/
|  CheckReserve()  |
|                  |
\_____(BookID)_____/
```

Figure 2: Question 2 a), A-Level CS 2021 9618/21 paper. The ‘structure chart’ solution from the marking scheme is shown (left – white background), beside the ChatGPT output (right – black background).

[6] doi.org/10.5281/zenodo.8179457

Table 2: Duration of each exam in hours, the marks allocated for each section/paper and the number of questions that must be answered in each section/paper, the number of sub-questions in total in each section/paper, the sub-questions that were answered incorrectly by ChatGPT (GPT-4) in each section/paper, and the marks lost as a result.

Exam/Year/Sections	Exam duration (hours)	Total marks allocated	Number of required questions	Number of correct sub-questions	Incorrect answers	Marks lost
Leaving Certificate Computer Science 2021						
Section A: Short Questions	1.5	30	6/12	19/22	Q5 a); Q5 b); Q9 b)	8
Section B: Long Questions		30	1/3	30/34	Q14 a) iv); Q14 a) v); Q14 b) i); Q14 b) ii)	11
Section C: Programming	1	50	1/1	6/6	–	0
Leaving Certificate Computer Science 2022						
Section A: Short Questions	1.5	30	6/12	15/18	Q5; Q11 a), Q12	13
Section B: Long Questions		30	1/3	26/29	Q13 b) i), Q14 b) i), Q15 c) i)	13
Section C: Programming	1	50	1/1	7/7	–	0
A-Level Computer Science (CIE) 2021						
9618/11: Theory Fundamentals	1.5	75	8/8	27/30	Q3 b); Q4 c) i); Q6 c)	9
9618/21: Fundamental Problem Solving & Prog. Skills	2	75	7/7	19/20	Q1 b); Q2 c); Q3 a) iii); Q4 a); Q4 c) ii)	20
9618/31: Advanced Theory	1.5	75	9/9	20/27	Q1 a); Q1 b); Q1 c); Q5 a); Q7 a); Q7 b); Q7 c)	22
9618/41: Practical	2.5	75	3/3	18/19	Q1 c) ii)	1

Incorrect Answers. Both exams contained a number of components that ChatGPT was unable to adequately address. The incorrect sub-questions are listed in Table 2. Interpreting images and generating diagrams caused the most issues in the LCCS 2021 exam:

- **Diagrams and images in questions.** The flowchart in Question 5 a) in Section A caused difficulties for ChatGPT, as it could not see the image and requested it as text – see Figure 3. A similar issue was seen with the table in Question 9 b) in Section A, and the diagram in Question 14 b) i) in Section B of the exam. ChatGPT did, however, attempt to create a diagram, despite not being able to see the diagram it was asked to complete - see Figure 4.
- **Strategic thinking.** Questions 14 a) iv) and v) in Section B were also answered incorrectly. Question 14 involves problem solving with a two-player game. Interestingly, some of the earlier and later subsections of this question were answered correctly.

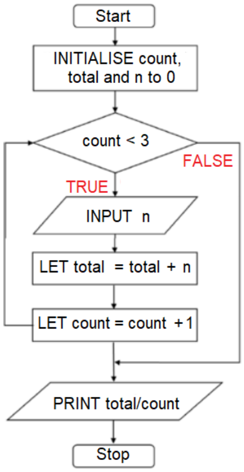
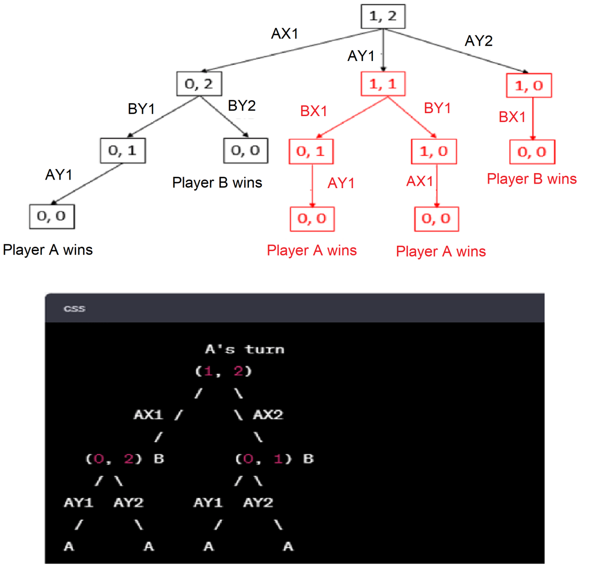


Figure 3: Question 5 a), Section A, LCCS 2021 HL paper. ChatGPT was unable to see the flowchart, and a trace table could not be generated. It requested the algorithm be provided as text or to describe the steps of the algorithm. (ChatGPT did attempt a diagram in Question 14 b) i) - see Figure 4).

Some similar and some different challenges were encountered in the CIE A-Level CS 2021 examinations:

- **Misinterpreted tables.** In Question 3 b) in paper 9618/11 ChatGPT made errors in the trace table it was asked to create for the program currently in memory. Four tables were provided and the table information may have been inadvertently mixed up in the prompt.
- **Incorrect solutions.** In Question 3 a) iii) in paper 9618/21, ChatGPT suggested the use of csv or tsv files when asked to suggest how data may be stored in a text file. ChatGPT gave an incorrect response in Question 6 c) in paper 9618/11. The marking scheme determined that the correct 'validation check' was an 'existence (check)', whereas ChatGPT gave the answer 'lookup check'. It also gave incorrect responses in Question 1 in paper 9618/31 in questions relating to 'normalised floating-point representation'. In Question 4 c) ii) in paper 9618/21, ChatGPT advised moving the wrong line of pseudocode.
- **Symbols in questions and answers.** In Question 4 c) i) in the 9618/11 paper, ChatGPT successfully added some check marks (✓) but not at suitable points in the solution to be correct answers. Also in Question 1 b) from the 9618/21 paper, ChatGPT gave an incorrect solution to a table which contained a number of arrow symbols (⟵). ChatGPT also could not correctly complete the trace table in Question 4 a) in paper 9618/21 that used the symbol ▿ to represent a space character. In addition, it could not write the Boolean expression in Question 7 c) in paper 9618/31.
- **Diagrams in questions.** The diagram in Question 5 a) and tables in Question 8 b) in paper 9618/31 could not be seen by ChatGPT. Similarly, the logic circuit in Question 7 in the same paper was not visible.
- **Diagrams or screenshots in answers.** In Question 2 c) in paper 9618/21, ChatGPT could not create a flowchart. In paper 9618/41 ChatGPT could not provide the screenshots that were requested in the exam paper. It did, however, return the correct code so that

a user could generate the correct output on a Python console, with the exception of Q1 c) ii).



```
              A's turn
             (1, 2)
             /   \
       AX1 /     \ AX2
         /         \
  (0, 2) B       (0, 1) B
   / \           / \
AY1  AY2     AY1  AY2
 /     \       /     \
A       A     A       A
```

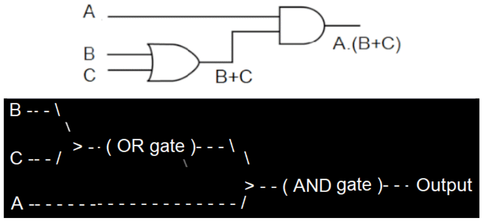
Figure 4: Question 14 b) i), Section B, 2021 LCCS HL paper. The diagram solution from the marking scheme is shown (top – white background), above the ChatGPT output (bottom – black background). The exam paper contained a partially completed diagram. ChatGPT generated a new (incorrect) diagram. In Question 5 a) ChatGPT did not attempt a trace table for an unseen flowchart - see Figure 3.

5.2 RQ2: Optional & Cascading Questions

Our second research question was: *What impact do optional and cascading questions have on ChatGPT's performance?* To answer this we compare the 2021 LCCS HL and 2021 CIE A-Level CS papers.

Optional Questions: On the A-Level, all questions are required unlike the LCCS which contains a total of 16 questions, only 8 of which are required – see Table 2. This element of choice is a large factor in the extremely high performance of ChatGPT on the LCCS discussed in Section 5.1.

Cascading Questions: Some questions involve a series of sub-questions, where the response to one sub-question influences the next sub-question (or set of sub-questions). The 2021 CIE A-Level CS papers contain slightly more cascading questions than the 2021 LCCS papers, and ChatGPT errors in cascading questions occur more frequently in the A-Level papers. Some examples in the LCCS 2021 paper include Question 14 b) i) where a diagram was not visible, and as a result, part ii) could not be answered; and Question 5 where an incorrect trace table in part a) meant that ChatGPT could not answer part b) of the question. In the A-Level CS paper 9618/31 a logic circuit image in Question 7 a) caused a similar issue. Also, ChatGPT could not answer Question 1 a) to c) in paper 9618/31 in a question relating to 'normalised floating-point representation'.



```
B - - - \
          \
C - - - /   > - - ( OR gate ) - - - \
                                      \
                                       > - - ( AND gate ) - - - Output
A - - - - - - - - - - - - - - - - - - /
```

Figure 5: Question 4 b), Section A, 2022 LCCS HL paper. The 'circuit diagram' solution from the marking scheme is shown (top – white background), above the ChatGPT output (bottom – black background).

5.3 RQ3: Effect of Knowledge Cutoff Date

Our third research question was: *Does ChatGPT's performance differ between exams made available before and after the current knowledge cutoff date of September 2021?* To answer this we compare the 2021 LCCS HL papers to the 2022 LCCS HL papers.

The 2022 LCCS exam has the same structure as the 2021 LCCS exam, but it is broken down into a total of 54 sub questions - see Table 2. The majority of these sub questions (48/54 or 89%) were answered correctly by ChatGPT (a slightly higher percentage than the 2021 LCCS of 88% of sub-questions). In this exam ChatGPT performs almost identically to the 2021 exam, receiving 296/300 marks (99%) – a H1. In 2022 only 19% of students received a H1 on this exam [11]. Further observations regarding these questions:

- **Identifying patterns.** In Question 11 b) ChatGPT, was able to describe the pattern emerging in the diagram (ancestry tree) in part a), despite being unable to view or extend it. Contextual information in the question allowed it to return the correct answer.
- **Diagrams in answers.** ChatGPT was unable to create a wireframe in Q13 b) i), Section B. However, in Q4 b) in Section A, ChatGPT could draw a circuit diagram – see Figure 5.

6 DISCUSSION

In both exams ChatGPT answered the majority of questions correctly (obtaining correspondingly high marks). It also often attempts to propose solutions in the absence of complete information. In some instances, ChatGPT is capable of generating charts or diagrams - see Figures 1, 2, 4 and 5. We also observed negligible differences between marks and solutions on papers made publicly available before and after the knowledge cutoff date.

One lesson learned was that when entering questions as prompts, they should be numbered/ labelled as ChatGPT may refer to them in subsequent sub-questions. Additionally, the process of evaluating the correctness of code generated by ChatGPT presents a substantial challenge due to the inherent flexibility in coding. There exist multiple viable approaches to answer most questions, and the preferred approach(es) may have evolved over time due to developments in the programming language or practices.

In our study, questions that were left unanswered or partially answered by ChatGPT typically involved images or symbols in the question that the model was unable to interpret (note we chose against using plugins for the reasons discussed in Section 4). We experimented with describing an image in the prompt but this was usually ineffective. We also experimented with copying text from images where possible, and including it to the prompt. This was often successful - copying the text in the flowchart in Figure 3 into the prompt allowed ChatGPT to generate the correct solution.

In AI, the term "hallucination" describes the generation of output that appears convincing but is factually untrue or unrelated to the current context [6]. For instance, ChatGPT has been known to cite sources that do not exist [15]. Notably, there are minimal occurrences of hallucinations in the model's output from the LCCS and CIE A-Level CS exams.

Compared to the A-Level CS exam, the LCCS exam offers a more flexible examination format in the form of optional questions, allowing students to select those questions they feel most confident in answering. This resulted in ChatGPT doing exceptionally well in this exam. Cascading questions were used slightly more frequently in the 2021 A-Level exams, compared to the 2021 LCCS exam (although there was increased use of them in the 2022 LCCS exam). This approach can simulate real-world problem-solving situations, assessing a student's capacity for logical thought, adaptation, and building upon prior solutions. However, it also poses a problem in that a wrong answer to one sub-question could affect the answers to subsequent sub-questions. Cascading questions may potentially add a layer of complexity to the exams, demanding a deeper comprehension of the subject area.

There is no coursework component in the CIE A-Level CS, which may appeal to students who prefer a more traditional assessment, but it means they miss out on the hands-on experience that the coursework component of the LCCS (or other A-Level CS exams offered by other boards) offer. It is also likely that coursework mitigates academic misconduct concerns, a topic that has been brought into the spotlight by ChatGPT [14].

ChatGPT often provided much more detailed responses than those in official marking schemes and alternative solutions in different programming languages. An implication of this is that tools like ChatGPT could contribute to expanding the scope of marking schemes by offering alternative solutions that may not have been considered initially – a likely welcome aid to marking programming exams with inherent flexibility in correct (and incorrect) answers.

As of August 2022, there were just 34 accredited CS teachers in Ireland [33] despite the LCCS being offered in 114 schools [13]. The UK has also been experiencing a CS teacher shortage for years [38]. Employers have concerns regarding the shortage of skilled individuals in computing roles, and there are inquiries about whether parents and educators unintentionally discourage students from pursuing a career in computing [36]. ChatGPT's versatility as an educational tool is evident across diverse educational settings. In the future tools such as ChatGPT may provide personalised learning experiences, helping students actively engage with material. With the use of plugins, these tools are becoming adaptable and more accessible, and may soon provide tailored feedback.

7 LIMITATIONS

This study is not without limitations. Government approved assessments are notoriously opaque and it is possible that we marked ChatGPT's answers differently than they would be by approved assessors. However we did use official marking schemes. The primary marker was a second-level teacher whose grading was verified by two university lecturers.

Although the LCCS and A-Level exams are publicly accessible, they are in PDF format. It is not known if the 2021 LCCS exam which was available prior to the knowledge cutoff date was in fact integrated into the GPT-4 model. Additionally, both the 2021 and 2022 LCCS exams are only available for download once several drop-down menus have been navigated, forming another potential barrier to these exams being included in the GPT-4 model. This could explain the similar performance observed between the (pre-cutoff) 2021 exam and the (post-cutoff) 2022 exam. We believe this limitation is not crucial as future AI models are likely to eliminate knowledge cutoff dates and integrate more data into their models. Auto-GPT and other plugins can already search the internet in real-time, bypassing the knowledge cutoff of ChatGPT. However, concerns of potential contamination remain due to limited information about ChatGPT's internal workings. [41].

Another limitation is that 30% of the LCCS is based on a group software development project such as a webpage with a database backend. We did not examine this. This requires a video, documentation, and other artefacts. Assessing ChatGPT on this would require a different approach. For instance, tools such as AutoGPT [31] are likely more performant on more complicated assessment such as this and would likely do better than ChatGPT. Other models such as Deepmind Alphacode [28] are trained on programming competition data and would also likely perform better on larger software development projects, providing an avenue for future work.

8 CONCLUSION & FUTURE WORK

ChatGPT has potential to serve as a valuable tool in education, providing support to teachers and learners that has long been sought by the AI in education community [2]. Despite the opportunities, challenges including the ethical, bias, and logistical issues such as student over-reliance abound [5]. The fact that ChatGPT is capable of answering many exam questions should not undermine the integrity of the LCCS and A-Level CS exams, as they are still conducted under invigilation. However, in terms of exam design implications, tools such as ChatGPT may encourage a shift towards assessing processes rather than exam-style answers.

If using ChatGPT as a learning aid, it may be useful to offer multiple choice solutions to ChatGPT within the prompt, so that the model defaults to the best answer, thus limiting the potential for responses with hallucinations. Perhaps in the longer-term future, we might re-imagine aspects of CS exams to include interactive components where students communicate with ChatGPT during the exam to gather information, discuss tactics, or seek clarifications. As a log of activity can be recorded, this could steer assessment more towards process and away from rote learning.

Going forward we anticipate utilising Generative AI to evaluate coursework and future exams to see if more observations that may be helpful in the future may result.

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183 and Huawei Ireland.

REFERENCES

[1] AQA:. 2019. AQA AS and A-Level Computer Science Specifications. https://www.aqa.org.uk/subjects/computer-science-and-it/as-and-a-level/computer-science-7516-7517/specification-at-a-glance

[2] Brett Becker. 2017. Artificial Intelligence in Education: What Is It, Where Is It Now, Where Is It Going. *Ireland's Yearbook of Education* 2018 (2017), 42–46.

[3] Brett A. Becker. 2021. What Does Saying That 'Programming is Hard' Really Say, and About Whom? *Commun. ACM* 64, 8 (jul 2021), 27–29. https://doi.org/10.1145/3469115

[4] Brett A. Becker, Steven Bradley, Joseph Maguire, Michaela Black, Tom Crick, Mohammed Saqr, Sue Sentance, and Keith Quille. 2023. *Computing Education Research in the UK & Ireland.* Springer International Publishing, Cham, 421–479, Ireland, along with a scientometric study of research outputs. https://doi.org/10.1007/978-3-031-25336-2_19

[5] Brett A. Becker, Paul Denny, James Finnie-Ansley, Andrew Luxton-Reilly, James Prather, and Eddie Antonio Santos. 2023. Programming Is Hard - Or at Least It Used to Be: Educational Opportunities and Challenges of AI Code Generation. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. ACM, NY, NY, USA, 500–506. https://doi.org/10.1145/3545945.3569759

[6] Gernot Beutel, Eline Geerits, and Jan T Kielstein. 2023. Artificial Hallucination: GPT on LSD? *Critical Care* 27, 1 (2023), 148.

[7] Neil Christopher Charles Brown, Michael Kölling, Tom Crick, Simon Peyton Jones, Simon Humphreys, and Sue Sentance. 2013. Bringing Computer Science Back into Schools: Lessons from the UK *(SIGCSE '13)*. ACM, NY, NY, USA, 269–274. https://doi.org/10.1145/2445196.2445277

[8] Neil C. C. Brown, Sue Sentance, Tom Crick, and Simon Humphreys. 2014. Restart: The Resurgence of Computer Science in UK Schools. *ACM Trans. Comput. Educ.* 14, 2, Article 9 (jun 2014), 22 pages. https://doi.org/10.1145/2602484

[9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. (March 2023). https://www.microsoft.com/en-us/research/publication/sparks-of-artificial-general-intelligence-early-experiments-with-gpt-4/

[10] OCR: Oxford Cambridge and RSA Examinations. 2021. A-Level Specification. Computer Science. https://www.ocr.org.uk/qualifications/as-and-a-level/computer-science-h046-h446-from-2015/

[11] careersportal.ie. 2022. https://careersportal.ie/school/lc_marks_distribution.php

[12] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, et al. 2021. Evaluating Large Language Models Trained on Code. arXiv:2107.03374 [cs.LG]

[13] Cornelia Connolly and Colette Kirwan. 2023. Capacity for, Access to, and Participation in Computer Science Education in Ireland, University of Galway (Ed.). . https://doi.org/10.13025/bccm-2c38

[14] Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2023. Chatting and cheating: Ensuring Academic Integrity in the Era of ChatGPT. *Innovations in Education and Teaching International* (2023), 1–12.

[15] Paul Denny, Brett A. Becker, Juho Leinonen, and James Prather. 2023. Chat Overflow: Artificially Intelligent Models for Computing Education - RenAIssance or ApocAIypse?. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) *(ITiCSE 2023)*. ACM, NY, NY, USA, 3–4. https://doi.org/10.1145/3587102.3588773 Video: www.youtube.com/watch?v=KwVcRXQc3IU, Slides: brettbecker.com/publications/#iticse23keynote.

[16] DES. 2020. *Assessment Arrangements For Junior Cycle and Leaving Certificate Examinations 2021.* https://www.tui.ie/_fileupload/assessment-arrangements-junior-cycle-and-leaving-certificate-examinations-2021%20FINAL.pdf

[17] DES. 2021. *Assessment Arrangements For Junior Cycle and Leaving Certificate Examinations 2022.* https://assets.gov.ie/85054/e0afa83e-e701-40d2-8dfe-8baba98a01ca.pdf

[18] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. 2023. Faith and Fate: Limits of Transformers on Compositionality. arXiv:2305.18654 [cs.CL]

[19] Cambridge Assessment International Education. 2021. Implementing the Curriculum with Cambridge A Guide for School Leaders. https://www.cambridgeinternational.org/support-and-training-for-schools/teaching-cambridge-at-your-school/implementing-the-curriculum-with-cambridge/

[20] Cambridge Assessment International Education. 2021. Syllabus. Cambridge International AS & A Level Computer Science 9618. Version 2. https://www.cambridgeinternational.org/programmes-and-qualifications/cambridge-international-as-and-a-level-computer-science-9618/

[21] Roisin Faherty, Karen Nolan, Keith Quille, Brett Becker, and Elizabeth Oldham. 2023. A Brief History of K-12 Computer Science Education in Ireland. *International Journal of Computer Science Education in Schools* 6, 1 (Mar. 2023), 3–34. https://doi.org/10.21585/ijcses.v6i1.148

[22] James Finnie-Ansley, Paul Denny, Brett A. Becker, Andrew Luxton-Reilly, and James Prather. 2022. The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. In *Proceedings of the 24th Australasian Computing Education Conference* (Virtual Event, Australia) *(ACE '22)*. ACM, NY, NY, USA, 10–19. https://doi.org/10.1145/3511861.3511863

[23] James Finnie-Ansley, Paul Denny, Andrew Luxton-Reilly, Eddie Antonio Santos, James Prather, and Brett A. Becker. 2023. My AI Wants to Know If This Will Be on the Exam: Testing OpenAI's Codex on CS2 Programming Exercises. In *Proceedings of the 25th Australasian Computing Education Conference* (Melbourne, VIC, Australia) *(ACE '23)*. ACM, NY, NY, USA, 97–104. https://doi.org/10.1145/3576123.3576134

[24] Matthew Hertz. 2010. What Do "CS1" and "CS2" Mean? Investigating Differences in the Early Courses. In *Proceedings of the 41st ACM Technical Symposium on Computer Science Education* (Milwaukee, Wisconsin, USA) *(SIGCSE '10)*. ACM, NY, NY, USA, 199–203. https://doi.org/10.1145/1734263.1734335

[25] Cambridge International. 2021. Cambridge International AS & A Level Computer Science (9618). https://www.cambridgeinternational.org/programmes-and-qualifications/cambridge-international-as-and-a-level-computer-science-9618/past-papers/

[26] Will Knight. 2023. OpenAI's CEO Says the Age of Giant AI Models is Already Over. https://www.wired.com/story/openai-ceo-sam-altman-the-age-of-giant-ai-models-is-already-over/

[27] Juho Leinonen, Arto Hellas, Sami Sarsa, Brent Reeves, Paul Denny, James Prather, and Brett A. Becker. 2023. Using Large Language Models to Enhance Programming Error Messages. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1* (Toronto ON, Canada) *(SIGCSE 2023)*. ACM, NY, NY, USA, 563–569. https://doi.org/10.1145/3545945.3569770

[28] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-Level Code Generation with Alphacode. *Science* 378, 6624 (2022), 1092–1097.

[29] Andrew Luxton-Reilly, Simon, Ibrahim Albluwi, Brett A. Becker, Michail Giannakos, Amruth N. Kumar, Linda Ott, James Paterson, Michael James Scott, Judy Sheard, and Claudia Szabo. 2018. Introductory Programming: A Systematic Literature Review. In *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education* (Larnaca, Cyprus) *(ITiCSE 2018 Companion)*. ACM, NY, NY, USA, 55–106. https://doi.org/10.1145/3293881.3295779

[30] Stephen MacNeil, Joanne Kim, Juho Leinonen, Paul Denny, Seth Bernstein, Brett A. Becker, Michel Wermelinger, Arto Hellas, Andrew Tran, Sami Sarsa, James Prather, and Viraj Kumar. 2023. The Implications of Large Language Models for CS Teachers and Students. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 2* (Toronto ON, Canada) *(SIGCSE 2023)*. ACM, NY, NY, USA, 1255. https://doi.org/10.1145/3545947.3573358

[31] Marek Mardosewicz. 2023. What is AutoGPT and How Can I Benefit From It? https://lablab.ai/blog/what-is-autogpt-and-how-can-i-benefit-from-it

[32] Clare McDonald. 2021. Number of Students Taking A-level Computing Rose in 2021: Computer Weekly. https://www.computerweekly.com/news/252505123/Number-of-students-taking-A-level-computing-rose-in-2021

[33] Sean Murray. 2023. Most Computer Science Teachers Not Accredited, Report Finds. https://www.irishexaminer.com/news/arid-41101681.html

[34] State Examinations Commission/Coimisiún na Scrúduithe Stáit. 2022. Leaving Certificate Computer Science, Higher Level Section A&B and Section C. English version. https://www.examinations.ie/exammaterialarchive/

[35] State Examinations Commission/Coimisiún na Scrúduithe Stáit. 2023. LC Computer Science Coursework Projct Brief 2023 Higher and Ordinary. https://www.examinations.ie/?l=en&mc=ex&sc=he#ComputerScience58

[36] Karen Nolan and Keith Quille. 2023. Why Kids Find Computing Jobs to be a Turnoff. https://www.rte.ie/brainstorm/2023/0529/1386218-children-students-computing-careers-jobs-ireland/

[37] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[38] Person. 2023. Bridging the Gap in Computer Science Teachers. https://www.bcs.org/articles-opinion-and-research/bridging-the-gap-in-computer-science-teachers/

[39] James Prather, Brent N. Reeves, Paul Denny, Brett A. Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2023. "It's Weird That it Knows What I Want": Usability and Interactions with Copilot for Novice Programmers. arXiv:2304.02491 [cs.HC]

[40] Brent Reeves, Sami Sarsa, James Prather, Paul Denny, Brett A. Becker, Arto Hellas, Bailey Kimmel, Garrett Powell, and Juho Leinonen. 2023. Evaluating the Performance of Code Generation Models for Solving Parsons Problems With Small Prompt Variations. In *Proceedings of the 2023 Conference on Innovation and Technology in Computer Science Education V. 1* (Turku, Finland) *(ITiCSE 2023)*. ACM, NY, NY, USA, 299–305. https://doi.org/10.1145/3587102.3588805

[41] Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, and Eneko Agirre. 2023. Did ChatGPT Cheat on Your Test? https://hitz-zentroa.github.io/lm-contamination/blog/

[42] Michael Wooldridge. 2020. *The Road to Conscious Machines: The Story of AI.* Penguin UK.