2225 – 6416M

Markscheme

May 2025

Computer science

Higher level

Paper 3

9 pages

Subject details: **Computer science HL paper 3 markscheme**

Mark allocation

Candidates are required to answer **all** questions. Total 30 marks.

General

A markscheme often has more specific points worthy of a mark than the total allows. This is intentional. Do not award more than the maximum marks allowed for that part of a question.

When deciding upon alternative answers by candidates to those given in the markscheme, consider the following points:

- Each statement worth one point has a separate line and the end is signified by means of a semi-colon (;).
- An alternative answer or wording is indicated in the markscheme by a "/"; either wording can be accepted.
- Words in (…) in the markscheme are not necessary to gain the mark.
- If the candidate's answer has the same meaning or can be clearly interpreted as being the same as that in the markscheme then award the mark.
- Mark positively. Give candidates credit for what they have achieved and for what they have got correct, rather than penalizing them for what they have not achieved or what they have got wrong.
- Remember that many candidates are writing in a second language; be forgiving of minor linguistic slips. In this subject effective communication is more important than grammatical accuracy.
- Occasionally, a part of a question may require a calculation whose answer is required for subsequent parts. If an error is made in the first part then it should be penalized. However, if the incorrect answer is used correctly in subsequent parts then **follow through** marks should be awarded. Indicate this with "**FT**".
- Question 4 is marked against markbands. The markbands represent a single holistic criterion applied to the piece of work. Each markband level descriptor corresponds to a number of marks. When assessing with markbands, a "best fit" approach is used, with markers making a judgment about which particular mark to award from the possible range for each level descriptor, according to how well the candidate's work fits that descriptor.

General guidance

Issue	Guidance
Answering more than the quantity of responses prescribed in the questions	• In the case of an "identify" question read all answers and mark positively up to the maximum marks. Disregard incorrect answers. • In the case of a "describe" question, which asks for a certain number of facts *eg* "describe two kinds", mark the first two correct answers. This could include two descriptions, one description and one identification, or two identifications. • In the case of an "explain" question, which asks for a specified number of explanations *eg* "explain two reasons …", mark the first two correct answers. This could include two full explanations, one explanation, one partial explanation *etc*.

1. (a) *Award **[2] max***
*Award **[1 max]** identifying what an LLM is.*
*Award **[1 max]** identifying what an LLM does.*

LLM is an AI algorithm that uses deep learning/typically uses a Transformer NN;
And a large data set to predict new content;

LLMs are a subset of deep learning models (trained on a large volume of data);
To generate human-like language/deploys natural language processing.

Note: Neural Network is not specific enough – deep learning or uses a Transformer NN

(b) *Award **[2] max***
*Award **[1]** for identifying what pre-process does*
*Award **[1] for** identifying the outcomes of pre-processing*
Pre-processing cleans and transforms data/removes noise and inconsistencies of data;
To improve the quality of data and/or accuracy of the model;

Pre-processing standardises data;
To reduce errors/improves models' ability to learn patterns;

Pre-processing can be used to remove irrelevant features/dimensionality reduction;
To reduce computational cost/training time.

2. (a) *Award **[4] max***
*Award **[1]** identifying an advantage of using TPU, and **[1]** for an expansion*

TPUs are application specific integrated circuits (ASIC)/built for ML workloads;
Accelerate machine learning workloads resulting in faster training/processes are optimised;

TPUs are optimised for tensor/matrix operations/perform multiple operations simultaneously;
Common in ML/enabling faster computations;

TPUs use less power for the same workload (compared to general-purpose hardware);
More cost-effective/due to parallelsation;

TPUs are easily scalable;
Ideal for LLM deployment/handle increased workloads;

TPUs accelerate real-time inference;
Enabling faster chatbot responses;

TPUs provide high memory bandwidth;
Improves model training efficiency, supporting large-scale tensor operations.

Note: Don't award a mark for faster computations/training without a reason.

*Mark as **[2]** + **[2]***

(b) *Award **[4] max***
For two processes
*Award **[1]** for naming/identifying the process.*
*Award **[1]** for describing relevance to the chatbot response.*

Lexical analysis:
A process that breaks down text into words/tokens;
To recognise key terms in the user's message;

Syntactic analysis/parsing:
A process that interprets the structure of the message;
Improves intent detection/Helps the chatbot understand how parts of the message relate;

Semantic analysis:
A process that finds meaning of the sentence;
Can be used to infer knowledge/respond with content that matches what the user is asking;

Discourse integration:
Helps understand the meaning in a larger context of the conversation;
Develops connections between utterances in a discourse (context and coherence)/Allows chatbot to respond in a way that builds on previous exchanges;

Pragmatic analysis:
Helps understand the social, legal, situational, and/or cultural context;
Understand how interpretation is different in different situations/adjusts the tone or content of the response based on the user;

Natural Language Understanding (NLU)/Nature Language Pipeline;
Uses intent recognition, entity extraction, and coreference resolution to understand the user's goal (e.g. tokenisation → POS tagging → intent recognition → entity extraction).

*Mark as **[2]** and **[2]***

3. *Award **[6] max***
*Award **[2]** for advantages, **[2]** for disadvantages, **[2]** for an opinion/conclusion*

*For Advantages Award **[1]** for each point, up to **[2]** max*
Privacy protection: Synthetic data contains no personal information, reducing legal and ethical risks;
Cost savings: Reduces or eliminates the need to purchase or license real datasets;
On-demand scalability: Synthetic data can be generated in large volumes, quickly and affordably;
Customisation: It can be engineered to match specific training needs (e.g. varied edge cases);
Bias control: Enables removal of known biases from the training process;
No storage burden: Regenerable on demand, so doesn't require long-term storage of large datasets;
Automatic labelling: Labels can be generated with the data, saving pre-processing time;

*For Disadvantages Award **[1]** for each point, up to **[2]** max*
May lack realism: Synthetic data may not accurately reflect real-world customer behaviour;
Risk of hidden biases: If based on flawed source data or poor algorithms, it can perpetuate;
Labeling bias: Errors in automatically generated labels can reduce model accuracy;
Domain limitations: May not include nuanced or domain-specific terminology (i.e. insurance-specific);
Validation required: Synthetic data must still be tested to ensure it effectively trains the model;

*Award **[2 max]** for an opinion/conclusion*
Must include a reasoned judgement that weighs advantages against disadvantages and reflects the context of RAKT's chatbot. A high-quality conclusion should suggest a logical outcome or strategy.

Example opinion/conclusions:
Although synthetic data can be engineered to closely reflect real-world datasets, each instance remains a fabrication, lacking the nuance and unpredictability of genuine interactions. This raises concerns about whether the chatbot will perform reliably when faced with actual customer queries.

Synthetic data may fail to capture the full complexity and variability of real-world communication. A model that performs well in a synthetic environment may not generalise effectively once deployed, especially in a specialised domain like insurance.

Synthetic data is ideal for training general-purpose chatbots, but RAKT operates in a specialised domain with nuanced insurance language, tone, and legal sensitivity. Without real data, the chatbot risks misunderstanding policy-related queries. In this case, synthetic data alone may not be sufficient.

While synthetic data is a safe and cost-effective option for initial chatbot training, relying on it alone may limit the model's ability to handle complex real customer queries. For RAKT, combining synthetic data with anonymised real interactions could offer both ethical protection and practical accuracy.

4. *Award **[12] max***

RNNs
Designed to process sequential data.
Can be used for speech recognition, generation of music, automatic translation etc.
For predicting the next word or phrase, previous words must be remembered. RNN uses a hidden layer to remember specific information about a sequence.
It allows past outputs to be used as input.
Chatbots may use many-to-many RNN architecture where both input and output are sequences.
The input layer of an RNN receives and processes the input before passing it to the hidden layer.
A recurrent hidden layer processes the input across time steps using shared weights.
RNN evaluates current input as well as what it has learned from past inputs.
Hence RNN is more suitable than a feedforward neural network for sequential tasks, as it can retain information from previous steps.
Backpropagation through time algorithm is applied to RNN with time series data as its input.
The output of the neural network is used to calculate and collect the errors once it has trained on a time set. The network is then rolled back, weights are recalculated and adjusted to reduce the errors.

Disadvantages of RNNs
RNN is slow to train.
Suffers from the vanishing gradient problem, where gradients become too small for effective learning during backpropagation.
RNNs struggle to retain long-term dependencies in sequences, and while LSTM networks help mitigate this issue, they do not fully overcome it.
Training is hard to parallelise because RNNs process inputs sequentially.
Exploding gradients can occur, leading to unstable training if not managed through techniques like gradient clipping.
RNNs are less effective than attention-based models at capturing complex or distant relationships within text, which limits their performance in advanced NLP tasks.

Transformer NNs
Transformer models use parallel processing — they process all inputs simultaneously, making them significantly faster than RNNs (e.g., BERT, GPT-3).
They use a self-attention mechanism to determine the relevance of each part of the input sequence to every other part.
The attention mechanism uses all input positions (not just the last state) to generate output predictions.
They dynamically weigh the importance of input elements, enabling context-aware understanding.

Transformers eliminate the need for recurrence by relying solely on attention, which reduces computational complexity and improves efficiency.
The encoder segment includes an embedding layer that converts input tokens into numerical vectors for processing.
Positional encoding is added to the embeddings to retain information about the order of words in the input sequence.
Multi-head attention splits inputs into queries, keys, and values, allowing the model to attend to information from multiple representation subspaces.
The outputs of each attention head are concatenated and passed through a normalisation layer to stabilise training and improve performance.
Transformer models can infer meaning even when some contextual information is missing, due to their global attention.
Transformers use contextualised word embeddings (e.g., from BERT) to understand not just individual words but their meaning in the surrounding context.

Disadvantages of Transformer NNs
Transformer models are very large and require significant computational resources, making them complex to deploy in real-world applications.
They can inherit biases present in their training data, which may lead to unfair or inappropriate outputs.
Bias mitigation requires careful data curation and the implementation of algorithmic fairness techniques.
They can be misused to generate misleading, harmful, or unethical content.
Verification and moderation mechanisms are needed to monitor and control output.
Training transformers effectively requires vast amounts of high-quality, domain-specific data.
Despite their responsiveness, generating real-time personalised content can demand high processing power, making scalability challenging.

Conclusions
Transformer NNs can capture long range dependencies and hence are most suited for chatbots to generate cohesive and contextually relevant content.
Transformer NNs have benefits over RNNs when it comes to learning mappings between sequences.
TNNs have solved the issues with long term memory (by means of attention) and computation speed (by removing recurrent segments).
TNN can generate highly tailored and personalized content enhancing user experience and satisfaction.
Hence RNNs may no longer be the first choice for generating language models. LSTMs are still being used today; one would need to find the best model for the job in hand.

Marks	Level descriptor
No marks	• No knowledge or understanding of the relevant issues and concepts. • No use of appropriate terminology.
Basic 1–3 marks	• Minimal knowledge and understanding of the relevant issues or concepts. • Minimal use of appropriate terminology. • The answer may be little more than a list. • No reference is made to the information in the case study or independent research.
Adequate 4–6 marks	• A descriptive response with limited knowledge and/or understanding of the relevant issues or concepts. • A limited use of appropriate terminology. • There is limited evidence of analysis. • There is evidence that limited research has been undertaken.
Competent 7–9 marks	• A response with knowledge and understanding of the related issues and/or concepts. • A response that uses terminology appropriately in places. • There is some evidence of analysis. • There is evidence that research has been undertaken.
Proficient 10–12 marks	• Demonstrates detailed understanding and accurate comparison of the relevant computer science concepts. • A response that uses terminology appropriately throughout. • There is competent and balanced analysis. • Conclusions are drawn that are linked to the analysis. • There is clear evidence that extensive research has been undertaken.