Int J Artif Intell Educ (2024) 34:1072–1101
https://doi.org/10.1007/s40593-023-00377-8

ARTICLE

Understanding Visitors' Curiosity in a Science Centre with Deep Question Processing Network

Zhaozhen Xu[1] · Amelia Howarth[2] · Nicole Briggs[3] · Nello Cristianini[4]

Accepted: 24 October 2023 / Published online: 20 November 2023

Abstract

Questions have a critical role in learning and teaching. People ask questions to obtain information and express interest in ideas. The Bristol scientific centre "We The Curious" launched "Project What If" in 2017 to inspire residents of Bristol to record their questions and pursue their curiosities. Researching these questions may help the museum better understand the curiosity of its audiences and create exhibitions or educational content that are more relevant to their interests and lives. The project managed to collect more than 10,000 questions on various topics, and more questions are being collected on a daily basis. With this large amount of data collected, it is time-consuming to process and analyse all the questions by humans. This research aims to apply artificial intelligence (AI) techniques and models in analysing these questions gathered by We The Curious. Meanwhile, in AI, there is a lack of tools that focus on processing and analysing the questions. Thus, we introduce a deep neural network called QBERT to process the questions for three tasks: question taxonomy, equivalent question detection, and question answering. Then we apply QBERT to provide an analysis of the questions collected by We The Curious, as well as comprehend Bristolians' curiosity. Then using QBERT, we categorise the We The Curious questions into 90 themes and 5,930 communities. Moreover, 436 questions are answered by one-sentence answers extracted from Wikipedia.

✉ Zhaozhen Xu
zhaozhen.xu@bristol.ac.uk

Amelia Howarth
arhowarth@live.co.uk

Nicole Briggs
Nicole.Briggs@wethecurious.org

Nello Cristianini
nc993@bath.ac.uk

1 MRC Integrative Epidemiology Unit, Bristol Medical School, University of Bristol, Bristol, UK

2 University of Edinburgh, South Bridge, Edinburgh, UK

3 We The Curious, Bristol, UK

4 Department of Computer Science, University of Bath, Bath, UK

Keywords Deep Learning · Natural language processing · Question answering · BERT

List of Abbreviations

AI	Artificial intelligence
WTC	We The Curious
BERT	Bidirectional encoder representations from Transformers
SBERT	Sentence BERT
QBERT	Question processing BERT
QT	Question topic classification
QE	Equivalent question recognition
QA	Question answering
YT	Yahoo! Answer topic classification corpus
YQA	Yahoo! Answer question answering corpus
QQP	Quora question pair

Introduction

Questions play a pivotal role in both learning and teaching. Numerous studies (Song, 2016; Ebersbach et al., 2020) have demonstrated that asking questions improved educational outcomes and promoted a further understanding of the learning material. People ask questions to obtain information and express interest in ideas. Moreover, encouraging learners to generate questions fosters critical thinking in the learning content. Research (Cuccio-Schirripa and Steiner, 2000) showed that questioning is one of the essential thinking processing skills for critical thinking, creative thinking, and problem-solving.

Meanwhile, with the development of online communities and smart voice assistants, many people are asking questions and searching for answers on a daily basis. Vast quantities of questions emerge every day. The idea of online question communities like Quora and Stack Overflow also encourages users to ask questions and connect with people who have the same question or have unique insights and quality answers.

The numerous queries produced lead to a challenge in artificial intelligence: Can we use the machine to understand and analyse these large-scale queries sourced from different contexts? Furthermore, how might such models be advantageous to education?

A typical question processing module includes query formulation and answer type detection (Jurafsky and Martin, 2014). The query formulation can include part-of-speech tagging and stopword removal. Answer type detection categorises the questions according to the anticipated answer type, such as numeric, location, entity, and so on. While the primary objective of the typical question processing module is to enhance the performance of a question answering system, the vast amount of data offers the potential for extracting other valuable information from the question itself.

Therefore, we report here on a multi-task processing question network we call QBERT (Q stands for question), which is built with a leading deep neural network

BERT (Devlin et al., 2018). QBERT is built as a generalist to solve three tasks we defined in the question domain: detecting the topic of questions, detecting equivalent questions with similar meaning but different wording, and locating potential answers to these same questions. The goal of having a generalist is that we can use one model to solve all the tasks, even if it does not excel in any task. These three tasks are helpful in discovering the information in the questions beyond facts.

Our approach is based on a fine-tuned language model sentence-BERT (SBERT) (Reimers and Gurevych, 2019), a Siamese BERT that projects the sentences into high-dimensional vector space. The embeddings of the sentences with similar semantic meanings are close to each other in the high-dimensional space. Notice that our intention is not to design a new model but to fine-tune SBERT in a multi-tasking way for processing questions. After fine-tuning SBERT with different tasks and loss functions, the embeddings generated from input data can be used for both classification and retrieval tasks. Moreover, our approach is time-efficient in finding the related sequence in a large corpus such as Wikipedia while combining with approximate nearest neighbour search (Johnson et al., 2019).

After training QBERT, we apply it to analyse a real-world dataset. In 2017, "Project What If" was started at the "We The Curious" science centre of Bristol (UK), with the stated intention of being the first exhibition all about "the curiosity of a city". Its aim was no less than to capture the curiosity of Bristolians (and visitors) by collecting all their questions. It was focused on the questions "of real people", and through these is aimed at understanding what Bristolians were curious about. In other words, it was not so much about the answers to individual questions as it was about understanding a Community from the questions it asks.

Despite the clear identity of "We The Curious" as a science centre, the organisers of this project were trying to gauge a broader set of interests, about culture and society, in a time of rapid change. A collection of the spontaneous questions of thousands of people was expected to provide deep insights about the people who asked them. It was expected that through this project, "We The Curious" could learn more about how the role of a science centre could evolve in addressing these questions in collaboration with its community. Over the following three years, the project gathered over 10,000 questions, both in their "museum" venue and in initiatives around the city. That list, taken together, contained many questions, worries, doubts, and ambitions of thousands of citizens.

A "We The Curious" corpus (WTC corpus) has been constructed from these questions. By analysing the queries from the WTC corpus, we present QBERT in a practical way, demonstrating its applicability for analysing question data from other sources.

With QBERT, we perform topic classification to identify the visitor's interest. On top of the topic classification task, we also include a type classification when analysing the questions. Besides, we introduce an external knowledge source as a candidate answer set when searching for answers to a given question. In such a way, we try to reveal more information contained in the question corpus. For example, from the question taxonomy, we can understand the people who asked the questions from their interests and comprehension level of the content; recognising equivalent questions can group the common questions and help locate identical and popular questions; and the question answering task can identify the questions that can be answered by an

external source, which potentially reduces the workload on answering queries. As a result, educational content providers can adjust their focus on the common questions or questions that a machine cannot answer based on the analysis.

In this article, we fine-tune a multi-task deep neural network to process questions. With this model, we aim to understand We The Curious visitors' curiosity by processing and analysing a new question corpus collected by We The Curious. The article describes the WTC corpus in Section 3, the algorithm in Section 4, the content-analysis of the corpus in Section 5, and the discussion of results in Section 6.

Background

Research (Cuccio-Schirripa and Steiner, 2000) showed that questioning is one of the essential thinking processing skills for critical thinking, creative thinking, and problem-solving. Learners' questions play a critical role in both learning and teaching (Chin and Osborne, 2008). Students' questions can help construct knowledge, self-evaluate, and motivate their interest in a topic during learning. On the other hand, teachers can diagnose students' understanding and evaluate their thinking through their questions.

Different kinds of questions can stimulate different extents of learning (Scardamalia and Bereiter, 1992). For example, knowledge-based questions generated from interest or to better understand and extend the knowledge have a higher order than text-based questions that are asked in response to given content. Scardamalia and Bereiter's research also showed that students tend to ask questions about basic information for less familiar topics but more wondering questions for familiar topics. Thus, categorising the questions can be beneficial to understanding the questioners and tailoring the learning content.

One of the most famous methods to categorise the questions for teaching is Bloom's taxonomy (Bloom, 1956). It divided the questions into knowledge, comprehension, application, analysis, synthesis, and evaluation. The taxonomy was generated in a way to help students learn.

In AI, question processing is considered a process in some early question answering systems (Li and Roth, 2002; Ferrucci, 2012). This process includes query formulation, answer type detection and keyword extraction. The question processing module intends to improve the performance of the question answering system. Nevertheless, questions are always a special type of text. In education, a question is used to reflect the student's ability and concerns (Chin and Osborne, 2008). Therefore, it is interesting to define question processing with regard to data analysis and find out what we can learn from it.

AI research pays attention to three types of question-related tasks: question classification, measuring question similarity, and question answering. Question classification is a specific case of text classification. Question similarity is similar to measuring text similarity. Moreover, question answering is a combination of text classification, similarity measures, and information retrieval.

Question classification is widely used to improve question-answering performance. The questions are classified based on the type of answer. Early work like Lehnert's

taxonomy (Lehnert, 1977) has 13 semantic classes for the questions. Recent work like TREC (Li and Roth, 2002) focuses on factual questions and creates a hierarchical taxonomy that separates the questions into 6 coarse classes (abbreviation, entity, description, human, location and numeric value) and 50 fine classes. However, to further analyse the comprehension content from the questions' complexity, the question type categorisation needs to be general, including factual as well as non-factual, regardless of the theme. According to the framework in (Mohasseb et al., 2018), the questions can be classified into 6 types based on grammar: confirmation, factoid, choice, hypothetical, causal, and list questions. In this paper, we did not train the model to predict the exact types as in (Mohasseb et al., 2018) because both the model and the dataset are not publicly accessed, and we are not able to label more data following the framework. Instead, we classified questions on the basis of which "interrogative word" they contain. An interrogative word is a function word for asking questions, such as what, which, when, where, who, whom, whose, why, whether and how.

Question similarity is a sub-field of measuring sentence similarity. This paper leverages a corpus-based approach that measures semantic similarity by utilising the information from large corpora (Chandrasekaran and Mago, 2021). Vector representations learned from the corpora are used to encode the text. This process is also known as embedding. As we discussed in the previous sections, the embedding methods used to capture sentence similarity have evolved in the past decades, from word embedding (Mikolov et al., 2013; Pennington et al., 2014) to sentence embedding (Arora et al., 2016; Conneau et al., 2017) to deep contextualised embedding (Reimers and Gurevych, 2019; Peters et al., 2018). This leads to the final method to measure semantic similarity, deep neural network-based methods. Among all the deep learning models, pre-trained language models (Devlin et al., 2018; Lan et al., 2019; Liu et al., 2019) created a top performance in capturing semantic similarity.

Question answering is always a challenging research task in natural language understanding. For question answering, we identified our research as open-domain and open-book answer retrieving. The system was designed to infer the correct answer from knowledge sources like Wikipedia in a concise sentence. Open-domain question answering is more challenging in using large-scale knowledge sources and machine comprehension than answering questions using a knowledge base. Previous research (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020; Chen et al., 2017; Yang et al., 2019) leveraged a retriever-reader or retriever-generator that retrieved the relevant passage from the knowledge source and extracted an answer span from the passage. The passage can be a document, paragraph, sentence or fixed-length segment. However, this two-stage system is computationally expensive. Inspired by DenSPI (Seo et al., 2019), we encode all the sentences in the knowledge base and search for the most relevant sentence with the query. In addition, we perform the approximate nearest neighbour search (Johnson et al., 2019) to reduce the search time.

Among all these tasks, the pre-trained language models had earned state-of-the-art results. The model we used, Bidirectional encoder representations from Transformers (BERT) (Devlin et al., 2018), and its modified models had leading performance in classification (McCreery et al., 2020; Sun et al., 2019) and question answering (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020). We used BERT to embed

the questions into high-dimensional vectors, and we used it to recognise the topic, equivalence and potential answers of a given question.

However, the complex architecture of the large language model requires more computational resources to train and utilise the model. Furthermore, it becomes more challenging to interpret the internal process of the network and understand the reasoning behind the network decision. As a result, we leverage multi-task learning (MTL) that trains all tasks jointly. It aims at improving generalisation by adapting domain information contained in the related training tasks (Caruana, 1997).

MTL is not a new idea. McCann et al. (2018) designed an MTL model that solved all the tasks simultaneously without identifying the primary tasks. In this paper, we call the single model that can perform multiple tasks a "generalist".

In natural language processing, MT-DNN (Liu et al., 2019) trains their multi-tasking model with the transformer encoder and task-specific layer so that it can apply to classification and regression tasks. To adapt to various tasks, some researchers reframe all the datasets into the same format. MQAN (McCann et al., 2018) formulates all the datasets into question answering over context. T5 (Raffel et al., 2020) creates a sequence-to-sequence format for the tasks. All these models focus on general language understanding tasks like GLUE (Wang et al., 2018), and decaNLP (McCann et al., 2018).

In contrast, we focus on a range of different tasks for processing questions. Furthermore, we aim to reduce the task-specific layers in the model by fine-tuning a general embedding model that can be utilised to process questions in different tasks.

WTC Corpus Overview

We will call our dataset of open-domain questions "the WTC corpus"[1], this section describes its origin and main features.

The dataset originated from a project run in Bristol (UK) by "We The Curious" (henceforth WTC), an educational charity and science centre.

Data Collection

Between January 2017 to October 2019, We The Curious collected over 10,000 open-domain questions from a diversity of sources: We The Curious venue in Bristol, offsite, and online. Offsite question gathering ensured questions were received from Bristol postcodes (BS1 - BS16). In the meanwhile, all Bristol postcodes were covered through general submissions. The data came from people of all ages and backgrounds. All these data collected started a digital database of questions, held by We The Curious. There are four main methods for question collection.

Project What If Cube This was a space set up on the venue floor whereby visitors could enter a large wire cube, write down their question on a piece of paper and then

[1] The anonymised and moderated WTC dataset is available from We The Curious on reasonable request for research purposes. Please contact information@wethecurious.org.

attach it to the cube. Questions were routinely taken down and stored, after which they were inputted by hand into the question database spreadsheet.

Curious Cube outings This was a portable cube that portrayed asked questions through a mirrored surface in LED lights. The cube could be connected to an iPad through which questions could be entered, stored, and displayed on the cube. The team collected questions at various events around Bristol.

Question gathering This was a series of events whereby We The Curious staff visited places of interest, such as schools and community centres, where they facilitated question input of participants using question cards. Questions were then inputted by hand into the question database spreadsheet.

Online input: An entry point was made available on the We The Curious website, whereby any user could enter a question digitally, and the question would be stored. When a question was entered through the website, no personal information was taken.

Apart from the online input, the questions were collected after visiting We The Curious venue or taking part in the events held by We The Curious. The participants were asked to leave any questions they were curious about without giving any further instruction related to the content or topic.

We The Curious created a digital database of questions, which is in We The Curious's possession. All the questions collected were represented verbatim. We The Curious is responsible and accountable for protecting the personal data of individuals submitting this information alongside their questions. All personal data is held by We The Curious in compliance with GDPR protocol (European Parliament, and Council of the European Union, 2016), and personal data is not shared with other parties, including the analysis team of this project. For the purpose of the present study, a smaller dataset was generated by removing all the personal data that was associated with the questions, and only this was shared with the analysts (Zhaozhen Xu and Nello Cristianini).

Data Pre-processing

Manual Curation of the WTC Corpus Questions were first moderated manually by We The Curious staffs. The questions in the database were also screened for any possible identifying information or potentially offensive or inappropriate language or content. These were removed from the database. After moderation, the resulting dataset contained 10,073 questions.

Automated Pre-processing The raw corpus also includes repeated questions, various types and topics, and other non-question sentences. Some simple pre-processing is performed before content analysis, such as removing exactly identical questions and questions shorter than three words. After these steps, the filtered WTC dataset contains 8,600 questions. This pre-processed, anonymised and moderated, textual dataset is

what we will call the WTC corpus in this paper. Table 1 shows the statistics of the WTC corpus. The distribution of the lengths of the questions is illustrated in Fig. 1.

The word cloud in Fig. 2 shows that the questions cover the universe and space, the human body, energy and climate change, animals and plants, chemistry and materials, the future and some other topics outside of the typical science categories listed before. The size of the word is proportional to its frequency in the corpus.

QBERT: a Question Processing BERT

In this section, we propose a multi-task approach to train BERT for processing short questions on a diverse set of NLP tasks, such as multi-class classification (question topic classification), pairwise classification (equivalent question recognition), and regression (similar question mining and question answering).

Question Taxonomy (QT) is a classification task that can identify the type and topic of a given question. The model categorises the questions into nine types and ten topics. The types are based on main question words like "what", "how", etc., and the topics depend on the content of the queries.

Equivalent Question Recognition (QE) can identify similar questions in the corpus by calculating the distance of the questions in a high dimensional vector space. This task helps analyse the questions in two ways: minimal the unique questions for further studying and understanding the most common questions.

Question Answering (QA) in this article is defined as an open-domain open-book task, as there is no limitation on the questions, and it can search for the answer from an external source. Through this task, we intend to understand what kind of questions can be answered with one sentence from a knowledge base like Wikipedia with confidence.

Datasets and Metrics

There are 8,600 questions in the WTC corpus, which is a relatively small amount of data for training a deep neural network. Furthermore, the datasets used to pre-train our model are real-world questions from users in different online communities and search engines, mostly without human curation or processing, similar to the WTC questions. As a result, before applying the model to the WTC corpus, we first trained and

Table 1 The statistics of the WTC corpus

Number of Questions	8,600
Vocabulary	5,732
Question Length	3 - 55 words
Average Question Length	7.15 words
Median Question Length	6 words

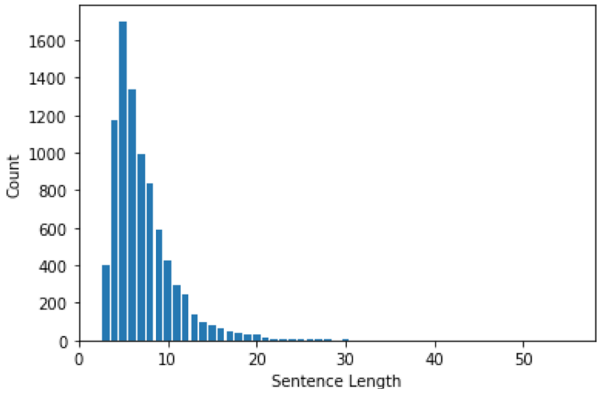


Fig. 1 The length distribution of the questions in the WTC corpus

fine-tuned the neural network on some open-accessed large-scale question datasets. After achieving a fair performance on the existing corpus, we adapted the model to the WTC dataset.

Based on the tasks we choose for processing questions, we train the model on Yahoo! Answer (Zhang et al., 2015), Quora Question Pair (Csernai, 2017), and WikiQA (Yang et al., 2015).

Yahoo! Answer generated by Zhang et al. is a corpus that includes queries, best answers and topics from the Yahoo! Answer website. There are 1,460,000 samples evenly distributed into ten topics. We utilised the Yahoo! Answer dataset for topic classification and question answering training. To separate the dataset for these two tasks, we named the dataset, which contains query-topic pairs Yahoo Topic (YT) and the dataset with question-answer pairs Yahoo Question-Answering (YQA).

Fig. 2 The word cloud is generated from the curated and filtered WTC corpus. The words were lemmatised before generating the graph

Quora Question Pair (QQP) is a question pair identification competition released in 2017. The dataset is widely applied for similar question classification. The questions are collected from user input on Quora, an online community for raising all kinds of questions. It includes 404,290 pairs of questions with annotations to identify whether the question pair is duplicated. There are 537,931 unique queries in the QQP dataset.

WikiQA is a dataset with factual query logs on Bing and candidate answers from Wikipedia Summary. Wikipedia Summary is the first paragraph of the Wikipedia article. The queries were carefully picked with a start of Wh-words such as "when" and "which". Furthermore, for any question, it required a minimum of 5 users to click on a Wikipedia page after searching. The dataset includes 3,047 questions and 26,154 candidate sentences with human labelling for correct answers.

One of the main reasons we chose these datasets is that the questions from these datasets are all user-generated and sourced from widely used communities and platforms such as Yahoo! Answer, Quora, and Bing. Moreover, Yahoo! Answer and QQP provide a broad and varied sample for training the model. On the other hand, WikiQA contains a one-sentence answer extracted from Wikipedia, which aligns with the QA task we defined. The summary of each dataset and the evaluation metrics are illustrated in Table 2.

Note that the answers gathered from Yahoo! Answer are not limited to short phrases or sentences. Sometimes, the answers could be an article. In this study, we intend to train the model to extract a one-sentence answer for a given query. Therefore, we will not evaluate the models on YQA.

Method

Recent research showed that pre-training language models achieved a good performance on language processing and understanding tasks. This article presents a fine-tuned question processing network based on a pre-training model called BERT (Devlin et al., 2018). BERT was initially trained on two tasks: the next sentence pre-

Table 2 Summary of the training datasets

Dataset	#Train	#Test	Name of Classes	Metrics
YT	1,400,000	60,000	Business & Finance, Computers and Internet, Education & Reference, Family & Relationships, Health, Politics & Government, Science & Mathematics, Society & Culture, Sports	Accuracy
QQP	283,001	121,286	is duplicate, not duplicate	Accuracy/F1
WikiQA	23,080	6,116	correct answer, incorrect answer	Accuracy/F1
YQA	14,000,000	600,000	answer	–

Note that we did not create our own training and test sets. The open-accessed dataset was separated into training and test sets in advance. The number of test data states here is the validation set that is publicly accessed

diction that understands the relationship between context and the masked language model that predicted the masked token in the content. The training was performed with BookCorpus (Zhu et al., 2015) and English Wikipedia (only the text part is included).

In the previous study (Xu et al., 2021), we applied a BERT-based model to process questions. One of the limitations is that the model used two network structures for different tasks. The model performed QT with a single BERT structure because QT requires a single question as the input and selects a label from multiple classes. However, in QE and QA, we used a Siamese BERT for taking pairwise inputs, such as (Question, Question) pair and (Question, Answer) pair. This reduces the consistency of the model and makes the model more complicated to be used in practice.

To improve upon this previous study by performing these three tasks with one generalist model, we consider the multiclass classification as a pairwise classification by taking the (Sequence, Class) as the pairwise input. We minimise the distance between sequence and class during training. On the other hand, we retrieve the closest class to a sequence instead of categorising the class for each sequence during inference.

Our method is based on SBERT (Reimers and Gurevych, 2019), which projects input sequence (sentence in this case) in high dimensional space with a Siamese BERT architecture. As a result, the sentences with similar meanings will be close to each other in the vector space. SBERT was trained based on BERT but with extra data (Bowman et al., 2015; Williams et al., 2018; Cer et al., 2017) to capture the sentence similarity. The idea of using the Siamese structure is that the model might capture the similarity between questions as well as the relationship between the question and its corresponding answer. Cosine similarity can be calculated between sentence representations produced by the model. Additionally, we can apply our model in binary classification by introducing a similarity threshold Θ, so that

$$Label = 1, \quad when \;\; Cosine\; Similarity > \Theta$$

By fine-tuning with the corpus described in Section 4.1, we develop a question-oriented SBERT, which we call QBERT. QBERT was first introduced by Xu et al. (Xu et al., 2021) in 2021. In this paper, we will improve QBERT based on our previous research by introducing a "generalist" that can achieve comparable results on multiple question processing tasks with a single model (Xu and Cristianini, 2023). Figure 3 illustrates the architecture of the model.

Input Layer $S = (s_1, ..., s_n)$ is an input sequence with n words. The sequence can be either a topic, question, sentence, or paragraph. The model takes a pairwise input (S, S') such as a question pair, question-topic pair, or question-answer pair. The pairwise input is then passed to two BERTs that share the parameters.

BERT Layer The shared embedding layer following the setup of $BERT_{base}$ which takes the sequence input as word tokens and generates an output for each token as well as a [CLS] token at the beginning of the output sequence. $BERT_{base}$ uses the an encoder containing 12 layers and 110M parameters and is pre-trained with two

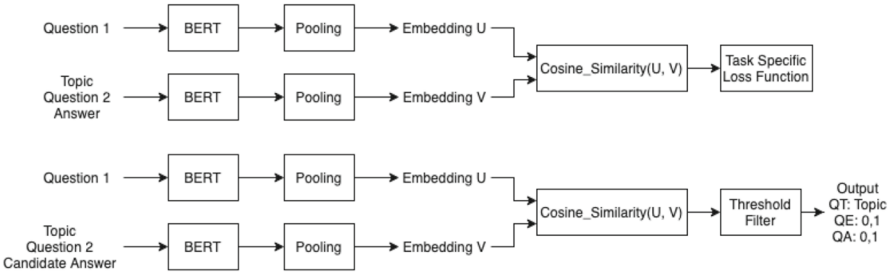


Fig. 3 Generalist QBERT architecture. The architecture is based on SBERT but trained to balance performance on various tasks. Top: Training as binary classification, Bottom: Inference by calculating the cosine similarity between input sequences. All BERTs share the same parameters. **U** and **V** are the embeddings of "Question 1" and "Topic / Question 2 / Answer", repectively

unsupervised tasks: masked language model and next sentence prediction. The output of the BERT layer is in $\mathbb{R}^d$ vector space, and according to BERT, $d = 768$.

Pooling Layer Similar to SBERT, the model leverages a mean pooling strategy that computes the mean of all output tokens (except [CLS]) of the sequence from BERT. After the pooling function, the model generates a pair of embedding U and embedding V as (1) , where $U \in \mathbb{R}^d$ and $V \in \mathbb{R}^d$.

$$Embedding = \frac{1}{n}\sum_{i=1}^{n} \Phi_{BERT}(s_i) \tag{1}$$

We apply two different loss functions for different types of data: online contrastive loss for binary classification tasks that have both positive and negative sample, and multiple negatives ranking loss for information retrieving datasets that does not contain positive nor negative label. Adam optimiser (Kingma and Ba, 2015) minimises the loss based on the cosine similarity $D_{cosine}(U, V)$.

Pairwise Classification Specific Loss QBERT introduces the contrastive loss (Hadsell et al., 2006) for pairwise classification. It aims to gather positive pairs in the vector space while separating negative pairs. For embedding U, V, the loss is calculated as follows.

$$L_{contrastive} = \frac{1}{2}\left\{Y\,(1 - D_{cosine})^2 + (1 - Y)\,[max\,(0, m - (1 - D_{cosine}))]^2\right\} \tag{2}$$

Where Y is the binary label. $Y = 1$ if U and V are similar. And the distance $D = 1 - D_{cosine}$ between U, V is minimised. When $Y = 0$, the distance increases between U, V until larger than the given margin m. In particular, we apply online contrastive loss that only computes the loss between hard positive and hard negative pairs.

Retrieval Specific Loss One of the advantages of applying multiple negative ranking loss is that the training dataset no longer requires either positive or negative labels. For a given positive sequence pair (S_i, S'_i), the function assumes that any (S_i, S'_j) is negative

when $i \neq j$. For example, in question answering, for question set $Q = \{q_1, ..., q_m\}$ and answer set $A = \{a_1, ..., a_m\}$, (q_i, a_i) is a positive pair given by the dataset, (q_i, a_j) is a negative pair randomly generated from the dataset. The cross-entropy loss of all the sequence pairs is calculated as follows.

$$L_{multiple\_negative} = -(Ylog\,(D_{cosine}) + (1 - Y)\,log(1 - D_{cosine})) \tag{3}$$

During inference, QBERT introduces a threshold filter. It calculates the $D_{cosine}(U, V)$, the cosine similarity between embeddings U and V, and applies different similarity thresholds for each task to determine if two sequences are related in terms of topic, equivalent question, or corresponding answer. The threshold of best performance is selected after training.

As shown in the previous research Xu et al. (2021), the training curriculum was critical for multi-task question processing. In Xu et al. (2021), the tasks were trained once at a time, from QE to QA to QT (QT was trained with different network architecture). However, the tasks learned in the earlier stage had a worse performance compared to the tasks learned in the later stage. To improve this, we train QBERT in a fixed-order round robin (RR) curriculum.

In this thesis, we refer to the training curriculum as the learning order for all the tasks. During training, the data in each dataset get divided into batches $Z = \{z_1, ..., z_n\}$. In each step, one batch z_i is selected randomly, and the model parameters are updated by stochastic gradient descent.

By applying the RR curriculum, QBERT trains all the tasks simultaneously in the RR curriculum. The data in each task-specific layer are built as mini-batches and divided into two task-specific loss functions. During each step, the model is trained and updated by batches with online contrastive loss and multiple negative ranking loss. QBERT-RR alternates between tasks during training, preventing the model from forgetting about the tasks learned at the beginning of the training.

The model applied a threshold filter instead of the task-specific loss function during inference in QE and QA tasks. The model only tried to minimise the distance between the related input sequences while training, regardless of identifying if the sequences were related. Thus, we introduced a distance threshold that can tell us if the question pair is similar or if the candidate sentence is the answer to the question. The distance threshold with the best performance was found during training.

Implementation Details

Before learning on question-related tasks, the BERT layer in the model was pre-trained following (Devlin et al., 2018) and (Reimers and Gurevych, 2019).

The length of the input sequence was limited to 35 tokens. Any sequence with a length of more than 35 tokens was truncated at the end. Compared to the original BERT, which has a maximum sequence length of 512 tokens (including special tokens), we reduced the sequence length because questions are mostly short sentences, and in the WTC corpus, 87.96% of the questions are within ten words. Besides, instead of

concatenating the sequence pair input into one sequence like BERT, QBERT applied two identical BERTs to read and process the sequence pair.

Usually, the machine takes a single sentence as input and performs multi-label classification for question topic classification. To adapt the QT task into the QBERT Siamese network architecture, we convert the topic classification into a topic retrieval task. QBERT embeds the questions as well as the topics, and labels all the question topic pairs as related. During training, the model minimises the multiple negative ranking loss between the topic and questions. During inference, the model calculates the cosine similarity between the question and all candidate topics. The candidate with the largest similarity is considered the topic of the question.

We train QBERT with the online contrastive loss for QE. We define the similarity threshold for QE based on the best accuracy on the training set. Then, we evaluate the model on both QE classification and retrieval tasks. The QE retrieval candidate corpus is constructed by sampled queries in the QQP test set.

For QA, we train WikiQA with the online contrastive loss and YQA and SQuAD with multiple negative ranking loss. This is because YQA and SQuAD only contain question answering pairs and do not come with negative samples. However, WikiQA has both positive and negative samples. In the meanwhile, there are questions with no answers in the dataset. Thus, a threshold is needed to identify whether the closest candidate to the question is the high-confidence answer. The threshold is defined as the one that creates the best precision in the WikiQA training set. Using the threshold that creates the best precision ensures the retrieving candidates have a low false positive rate. In other words, the answers selected by the model are more likely to be the correct answer.

The implementation of QBERT is based on PyTorch and SBERT. The training parameters are shown in Table 3. The margin for positive samples and negative samples is 0.5. We train the model for 5 epochs with a batch size of 32 and a learning rate of $2e-5$. 10% of the training data is used for warm-up.

We train QBERT with one GeForce GTX TITAN X GPU. Training specialised QBERT takes 7 hours on the QQP dataset, 0.5 hours on WikiQA, 9 hours on the YT, and 16.5 hours on YQA. On the other hand, training the generalist QBERT takes 93 hours. Even though training QBERT-DI is time-consuming, once trained, the model is much faster during inference. It takes 1.5ms, 5.44ms, 19.62ms, and 49.76ms per question in YQT, QQP, WikiQA, and SQuAD, respectively.

After training the network, we found the best threshold for the QE task by evaluating the performance of training data. In the QE task, the model was trained to classify

Table 3 Training parameters for QBERT

Epoch	5
Batch Size	32
Sequence Length	35
Learning Rate	2e-5
Warm-up Steps	10% of the training data
Margin for Online Contrastive Loss	0.5

similar questions. During inference, the model calculated the cosine similarity between the question pairs, where $similarity = 1 - D_{cosine}$. If the question embeddings had a larger similarity than the defined threshold, they were considered similar questions. In an ideal situation, all the question pairs with high similarity should be labelled as 1 (similar). We sorted the question pairs by similarity in descending order to find the best threshold. All the similarity scores calculated between the training question pairs were utilised as the threshold to evaluate the accuracy of the model. The threshold with the best accuracy was recorded for the QE task.

The implementation of QBERT was built with PyTorch and the SBERT library (https://github.com/UKPLab/sentence-transformers). All the training data is open accessed.

Performance

We evaluate QBERT with YT for QT classification, QQP for QE classification and QE retrieval, and WikiQA for QA.

For QE, we evaluate classification and retrieval task accuracy with the QQP dataset. If the question pair has a similarity larger than the threshold, it is categorised as equivalent in classification. To perform similar question mining, we create a question corpus based on QQP. First, all the relevant questions for the given query are included in the dataset, ensuring that there is always a relevant question in the corpus. Second, we fill the rest of the corpus with irrelevant questions. There are 104,033 samples in total. While mining the similar questions from the corpus, the candidate with the highest similarity larger than the threshold is defined as the duplicate question.

The performance of classification tasks like YT and QQP is evaluated with accuracy on the test set. In QE, the label was predicted by the threshold. On the other hand, in WikiQA, the question is not guaranteed to have an answer. Therefore, the model takes the sentence with the highest cosine similarity score in the candidate set for each question and compares it with the threshold. The prediction is correct if the similarity is above the threshold and the sentence is labelled as a correct answer. The results are illustrated in Table 4.

Table 4 The performance of QBERT-RR compares with the performance of single-task SBERT trained on QT, QA and QE

Models	YQT Acc.	QQP Acc.	WikiQA Acc./F1
SBERT	35.27±0.58	74.80±0.32	77.46±2.82/58.24±12.48
QT	72.44±0.39		
QE		89.79±0.23	
QA			79.05±5.89/72.50±11.08
E5	55.16±0.44	77.72±0.21	**82.54±3.55**/72.88±6.86
QBERT-RR	**73.77±0.58**	**90.13±0.19**	81.90±5.60/**73.73±8.12**

SBERT and E5 without training on any datasets mentioned in Section 4.1 are used as the baselineThe bold entries indicate the best performance for each column

SBERT was only trained on natural language inference datasets and semantic textual similarity datasets containing sentence pairs with labels. It, therefore, manages to detect similar question pairs, albeit with poor performance. However, SBERT was not trained to group sentences with the same topic, and it is unable to identify the question topic. Since SBERT achieves similar accuracy to other models on the WikiQA dataset, it has a worse F1 score compared to others.

In Table 4, model QT, QE, and QA represent single-task training. It leverages the same architecture as QBERT. However, for each task, it has a separate model. While fine-tuning the single-task model, we update the BERT layer to minimise the task-specific loss for each dataset. The results show that QBERT-RR achieves comparable or better performance on most question datasets compared to the single-task model.

We also compare QBERT-RR with E5 (Wang et al., 2022), a state-of-the-art general-purpose embedding model that achieved strong performance in classification, clustering, and retrieval. Compared to QBERT, E5 trained the shared encoder with a prefix identifier for the data. The results show that E5 obtains similar performance on QA. On the other hand, QBERT-RR outperforms E5 on QT and QE tasks.

In addition, we evaluate QA retrieval using WikiQA. We evaluate the performance of the retrieval task by accuracy@K, which calculated the accuracy of the answer within the closest K positions to the question. We create an answer corpus for each query in the WikiQA test set. First, all correct answers to the given query are included in the answer candidates corpus so that we made sure that there was always a correct answer in the corpus. Second, we filled the rest of the corpus with irrelevant sentences in the WikiQA dataset. The results of retrieving answers from WikiQA are shown in Fig. 4.

The results show that QBERT has a better performance in retrieving answers from a given corpus compared to SBERT and the single-task model. However, E5 achieves a leading performance in answer retrieval.

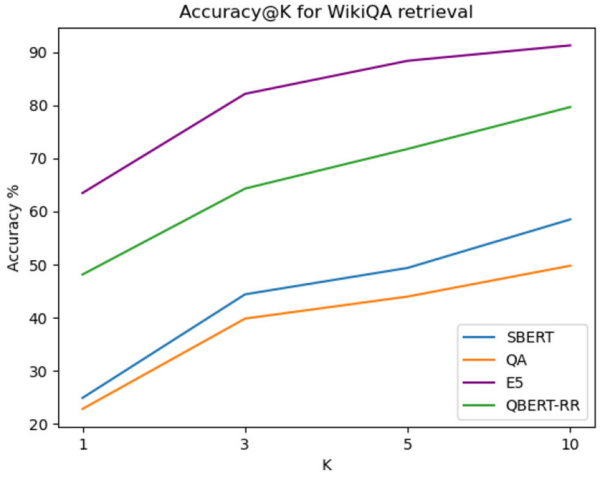


Fig. 4 Accuracy@K for different models on retrieving answers from WikiQA. There are 26k sentences in the candidate answers set

The corpus generated for QA retrieval has a size of 26k sentences, which is much smaller than the original Wikipedia Summary. However, we did not manage to find the same dump of Wikipedia as in WikiQA datasets. Thus, searching in a larger corpus might further affect the results.

WTC Corpus Analysis with QBERT

After training QBERT-RR, we apply the model to the WTC corpus for processing and understanding questions. In the WTC corpus, we have questions with various content from different people. The questions cover multiple topics and overlap in the content.

Through the analysis, we intend to understand the question content, like the types and topics, so that people can further understand the questioner's interest and understanding level. We can filter the duplicate questions to reduce the workload for further analysis when detecting similar questions. Furthermore, QBERT can link the new questions in the future with the existing data. The repeating questions also reveal the common doubt from the visitor that can be used to comprehend the questioner. In the end, QBERT discovered the questions that can be answered with high confidence by Wikipedia Summary. Under an educational scenario, people can focus more on the knowledge-based questions that the machine cannot answer.

Question Taxonomy (QT)

In the WTC corpus, there are confirmation questions such as "Are all babies born with blue eyes?"; factual questions such as "Who built the internet/electricity?"; and counterfactual questions such as "How long will the earth and humans last if we carry on damaging it and nothing changes?". Different types of questions can indicate the questioner's depth of thinking (Chin and Brown, 2000). Identifying the type and topic of the questions provides insight into the visitor's understanding and thinking (White and Gunstone, 2014).

When categorising the questions by type, we focus on classifying them based on the grammatical forms: interrogative words. The type of question is categorised into the following categories: WHAT, WHO, HOW, WHEN, WHERE, WHY, WHICH, IF, and OTHER, which will be further discussed below. Keyword matching is leveraged for type classification. On the other hand, for topic classification, we focus on the semantic topic of the questions, which includes Business & Finance, Computers and Internet, Education & Reference, Family & Relationships, Health, Politics & Government, Science & Mathematics, Society & Culture, Sports. The topic of a question is classified by QBERT-RR. We had nine types and ten topics, and therefore 90 question "Themes" to which we can allocate the over 8,000 distinct questions that have survived the various stages of filtering.

Identifying the interrogative words in a question helps identify grammar-based question categories used in linguistic research, such as confirmation, factoid, hypothetical, and casual.

Confirmation questions are also known as yes-no questions. A yes-no question is a polar question that contrasts with a non-polar question like the wh-question. And it expects the answer to be one of two choices. Since the answer to the confirmation question can be a simple "yes" or "no", the question is considered less complicated compared to others.

A factoid question is a question related to facts. Most of them contain question words like "what", "who", "which", "when" and "how". Usually, factoid questions can be answered with a concise sentence from a knowledge source. Both confirmation and factoid questions are mostly asked to help understand the conceptual knowledge of given content.

Causal questions usually begin with "why" or "how". This category of questions usually requires further explanation in the answer and can be more challenging to answer with one sentence.

A hypothetical or counterfactual question is defined as a question of the type: "what would happen if X was true". The understanding is that X is not a true fact, but the people who ask the question are considering the possible consequences of X being true. The counterfactual question takes its name from being "counter to the facts", is often used in defining the notion of causality (e.g. in Pearl, 2018), and indicates a mental process directed at understanding the mechanism behind observations. Causal and hypothetical questions are asked with further thinking, interest, and effort to make sense of the world.

Although there are overlaps between interrogative words and grammar-based question categories, classifying the questions into various types can still contribute to diagnosing people's understanding.

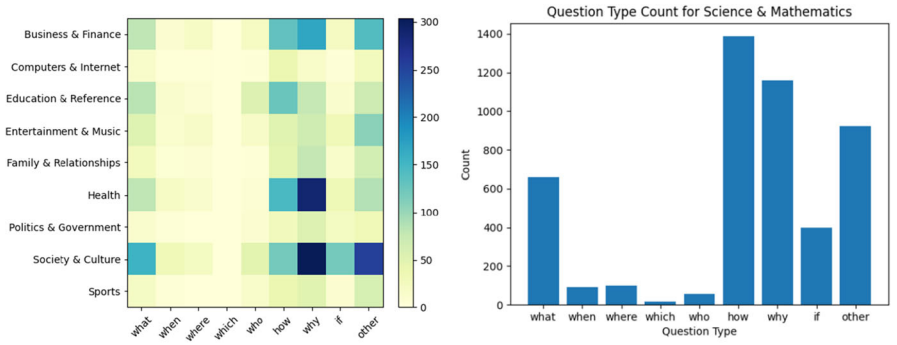
Given a question, we assign it to the type of the first keyword from our list that is found in it, with one notable exception for type IF. For example, the question "Why do we get butterflies when we like someone?" is categorised as WHY regardless of any other keywords that are in the query after "why". However, questions that contain the keyword "if", such as "what if" and "How ... if ..." are classified as IF questions regardless of the position of "if". The category OTHER includes yes/no questions or sequences that do not fit into other categories.

Define a further class of IF questions intends to find a simple way to approximate the counterfactual questions of the type "what if", which, however, are difficult to capture precisely by keyword matching but can well be approximated in this context by checking for the use of the word "if".

After categorising the type of queries, QBERT-RR identifies the topics for each question in the WTC corpus. Since QBERT-RR is trained with Yahoo! Answer dataset, it classifies the WTC corpus following the ten most popular question themes from Yahoo! Answer. It is essential to notice that there are many non-scientific questions in this list, which was part of the initial intent of the overall project: to assess the scope and breadth of the curiosity of an entire community.

QBERT-RR embeds the topics when classifying the topic and compares the cosine similarity between the question and all ten topics. The topic with the highest similarity with the question is considered correct.

Figure 5 shows the frequency distribution of the questions across types and topics. The most "asked" topics in the WTC corpus are science & mathematics and society



(a) Science & Mathematics Excluded (b) Science & Mathematics

Fig. 5 The number of topics and types in WTC. Left: All the types and topics excluding Science & Mathematics. Right: Type count for Science & Mathematics

& culture, which make up 67.94% of the corpus. Besides, 50.90% of the questions are HOW and WHY questions. The theme " HOW + Science & Mathematics" contains 1,387 questions, which is the highest among all 90 themes.

Equivalent Question Recognition (QE)

We identify the equivalent questions in the corpus by mapping the questions into the 768-dimensional embedding space with QBERT-RR. These representations can be compared using cosine similarity in the embedding space. Two questions are deemed to be equivalent if their cosine similarity is above the given threshold.

To better understand the performance of QBERT-RR in finding equivalent questions, we label some question pairs in the WTC corpus. To generate question pairs, we first randomly sample 1,000 questions. The questions were embedded by the SBERT that trained only with the NLI (Williams et al., 2018) dataset. For each question, we select ten questions in the corpus that have the largest cosine similarity with the given questions. These 1000×10 question pairs are selected as similar question pair candidates. We only keep one for duplicate question pairs such as [Q1, Q2] and [Q2, Q1] for annotation.

We have two annotators labelling 5,022 candidate question pairs, hence, 2,511 pairs for each annotator. On top of that, each annotator labels another 30 questions for measuring inter-annotator agreement (IAA). The questions that can be answered with the same answer are considered equivalent. The question pairs were labelled with 0 or 1, where 0 represents different and 1 for similar. From the labelled question pairs, 728 pairs are equivalent, and 4,294 pairs are different. The Cohen's Kappa (Cohen, 1960) between two annotators is 0.75. Table 5 provides some examples of the data we label.

When the cosine similarity threshold is 0.809, QBERT-RR obtains 90.76% accuracy on the sampling WTC data, which is a similar performance to QBERT-RR on the QQP dataset (90.13% accuracy). This proves that QBERT-RR can be applied to a different corpus of unseen questions. We also compare the classification results with SBERT and SBERT-QE. The results are shown in Table 6.

Table 5 Examples from labelled WTC question pairs and results

	Question 1	Question 2	Label	QBERT-RR
1	How does earth spin around?	What way does the Earth spin?	same	correct
2	Who made us?	Who invented people?	same	correct
3	How are mirrors made?	How are crystals made?	different	correct
4	How do you make glass?	How is glass made?	same	wrong
5	What will happen if the world ends?	If the world stops spinning, what would happen?	different	wrong
6	When did the humans come alive?	How did humans first exist?	different	wrong

Another experiment we perform to recognise equivalent questions is applying a "graph community detection" method (Clauset et al., 2004; Hagberg et al., 2008) in order to group similar questions. A graph is built with question nodes to cluster the questions using the cosine similarity matrix. An edge exists between nodes if the cosine similarity between a pair of questions is larger than the threshold. There are 5,930 communities found in the WTC corpus, which represents 5,930 different questions in the corpus. Of these, 5,337 questions do not have any similar question groups in the corpus.

Question Answering (QA)

To find out if the model can answer the WTC questions with high confidence, QBERT-RR retrieves a sentence as the answer from an unstructured knowledge base, Wikipedia Summary (Scheepers, 2017).

Wikipedia Summary includes the title and the first paragraph as the summary for each Wikipedia article extracted in September 2017. The raw texts of Wikipedia have 116M sentences initially. Of these, 22M are in the summaries. After embedding with QBERT-RR, there are around 21M distinct sentences left. The summary of Wikipedia provides the article's primary information. In the meanwhile, it reduces about 80% of the sentences from the original Wikipedia.

While retrieving answers from a large-scale knowledge Source, it is important to ensure the answer is the closest to the question and with high confidence. We define confidence score with the cosine similarity threshold. Suppose the similarity between the question and the candidate sentence is larger than the threshold. In that case, the candidate is considered as the answer to the given question. We leverage the threshold

Table 6 Results of equivalent question recognition for the WTC corpus

	Threshold	Accuracy	F1
SBERT	0.814	85.07%	53.12%
QE	0.833	90.08%	60.22%
QBERT-RR	0.809	90.76%	60.27%

calculated in Section 4.4 from the WikiQA dataset as the confidence threshold for the WTC. The best-scored sentence from Wikipedia Summary with a higher cosine similarity than 0.795 is believed to be the correct answer for a given question.

Seven GeForce GTX TITAN X GPUs are used to embed all the sentences from the Wikipedia Summary with QBERT-RR. It takes 1.5 hours to encode all the 21M sentences. Due to the scale of the dataset, it is time-consuming to perform an exhaustive search among all candidates. Instead, we locate the answers to questions by using an approximate approach: approximate nearest neighbour (ANN), which trades off accuracy for searching speed. When searching with ANN, it takes a few hours to build the index for the candidates at the beginning. Nevertheless, it takes less than a second per query to search.

The Wikipedia Summary index is trained and built using the inverted file with exact post-verification for 4 hours (Johnson et al., 2019). After building the index, it takes 3 minutes to search the candidate answers for all 8,600 WTC questions with one GPU, an average of 0.02s per question.

After filtering the high-confidence answers, there are 463 questions in the corpus that can find an answer within the Wikipedia Summary using QBERT-RR. The number of answered questions in different types and topics are shown in Fig. 6.

Result Discussion of the WTC Corpus

"Project What If" was launched in 2017 across Bristol and involved thousands of people. It aimed to explore the questions of Bristolians rather than the answers, to see what they said about the local community. It aimed to observe similarities and differences in people's curiosity regardless of their age, gender, and geography. By

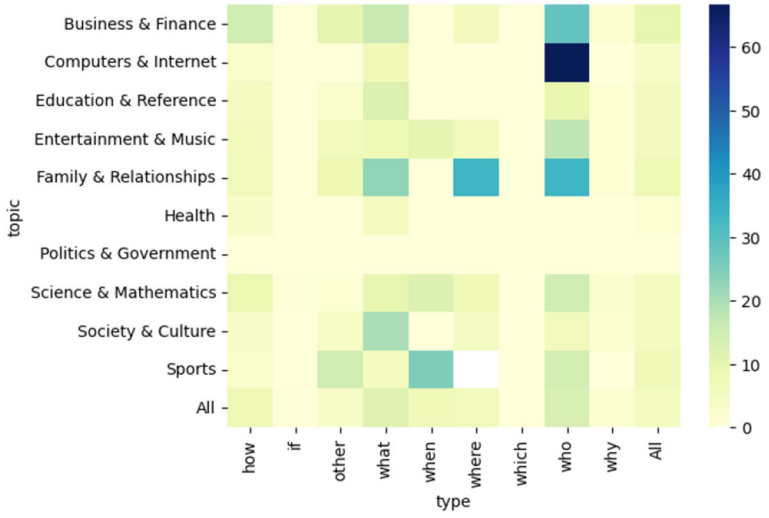


Fig. 6 Number of questions in the WTC corpus that can be answered with high confidence over the size of the groups

not setting any rules or prescribing what topics to explore, it was hoped the science centre's exhibitions and the educational content might better reflect the interests of its community. In this section, we try to discuss the result produced by QBERT-RR and present further analysis of the data.

The automated QT analysis of the corpus revealed that more than half of the questions are in the domain of Science & Mathematics (55.66%), followed by Society & Culture (12.28%), and then by Health (7.77%). The most frequently asked type of question is the type WHY (26.34%) followed by HOW (24.56%).

It is expected that most of the questions were collected under the Science & Mathematics domain as many of the questions collected were in the science centre setting or during experiences facilitated by WTC staff. Therefore, people are already in that frame of mind. Apart from science-related subjects, the questioners are most interested in Society & Culture. The results show that the questioners are least interested in Computer & Internet related subjects.

The word cloud in Fig. 7 illustrates the high-frequency words for each topic. We notice that words like "people" and "human" have a high frequency in most topics. Humans are naturally related to all these topics. Therefore, in order to better understand other information unique to each topic, we remove "human/people" as well as the stop words while visualising the topics. The size of the words is proportional to their frequency in the corpus.

Word clouds give us an insight into the key information held in different topics by showing high-frequency keywords. In Computer & Internet, the keywords include "computer", "internet", "game", and "Fortnite", which are not spotted in other topics. In Family & Relationships, the most frequent word is "love". Besides, it includes many words about feelings and relations. Since Science & Mathematics is a broad topic, the keywords illustrated in the figure also contain different varieties such as "earth", "animal", and "planet" etc. These observations again show that QBERT-RR is able to distinguish the topics in the questions.

On the other hand, the type classification also reveals some interesting patterns. More than 50% of the questions are type WHY and HOW, which are causal questions, showing that the questioners went through further thinking while asking the questions. For factoid questions, WHAT questions are much more than other types like WHO, WHEN, and WHERE.

Since the questions are collected under the "Project What If", it inspires people to ask IF questions. By navigating the IF question, we observed that most of the questions are counterfactual, such as "What if we never went to sleep?" and "If you could hear in space, how loud would the Sun be?". However, the corpus contains some factual questions in type IF as well, as would be expected in a science centre setting. For example, "I'd like to know if atoms are made up of other atoms.". Factual questions in type IF can be further filtered in future work.

Furthermore, we calculated the $P(type, topic)$ and $P(type) * P(topic)$ to understand the associations between type and topic. Figure 8 illustrates the associations. The question-type WHO is strongly associated with Education & Reference and with Sport because the $P(Who, Sport)$ is 3 times larger than the probability of $P(Who) * P(Sport)$. The topic Education & References is strongly associated with types: HOW, WHAT, WHEN, WHO. The type IF associates strongly with Politics

Fig. 7 High frequency words in each topic

& Government, but not with Education & Reference. We also observe that people expected more explanations on the Health and Family & Relationships topic because type WHY associates more with these two topics rather than others.

One limitation of using QBERT-RR for topic classification is that it only takes the top 10 topics from Yahoo! Answer regardless of all other possible topics. Nevertheless,

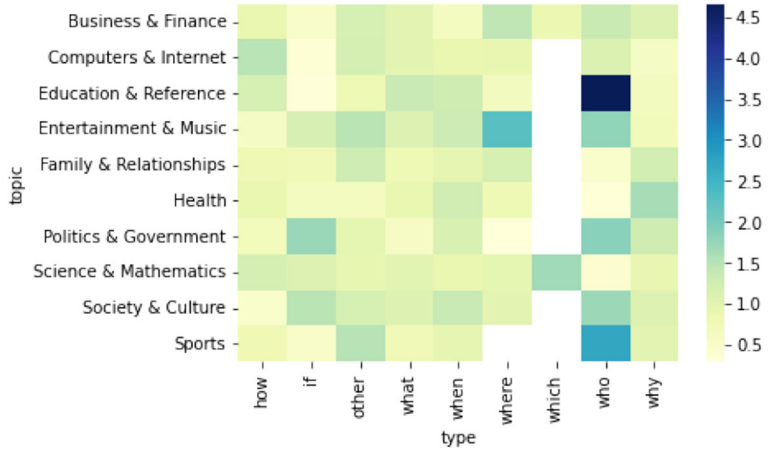


Fig. 8 The associations between types and topics for the WTC corpus. The association was calculated with $P(type, topic)/(P(type) * P(topic))$

we notice that many questions do not belong to any of the groups in the Yahoo! Answer dataset. This can be improved by training the model with more question topics or introducing a threshold to determine the questions that are not close to any topic.

When clustering the similar questions by distance with graph community detection, we find that one of the groups has 858 questions, which, by definition, are supposed to share similar meanings within the same graph. This happens because queries are connected with an edge when they have a cosine similarity larger than the threshold; however, the questions were not fully connected within the group, which means not all the questions within the subgraph are equivalent. For example, a group contains a set of questions $[Q_1, Q_2, Q_3]$. (Q_1, Q_2) and (Q_2, Q_3) are connected with an edge, respectively, but the similarity between Q_1 and Q_3 is smaller than the threshold, which means that they are not an equivalent pair and are not connected. In this section, we analyse similar questions regardless of this particular group that included over 800 queries.

The most popular questions in the WTC corpus are related to life outside of Earth. There are 144 of them in the corpus, such as "When will we find intelligent life in the universe?" and " Is there any life on any other planets or solar systems in the universe?". Following that, there are 143 questions related to the end of life/world, such as "What would happen if all humans were extinct and the world stopped spinning?" and "What does it feel like on the moment you die?". Overall, 26 groups of questions have more than 10 similar queries within the datasets. 3,263 questions in the WTC corpus have at least one similar query. Thus, applying similar question detection can help reduce the corpus size for further processing for the data collector.

With QBERT-RR, we find answers to 463 out of 8,600 questions in the WTC corpus. However, we notice that answers with high confidence (over 0.92 cosine similarity) are equivalent questions that QBERT-RR finds in the knowledge source. For example, QBERT-RR is tricked by a sentence in Wikipedia Summary, "How did the universe come about?", and considers it the answer to the question "How did the universe begin?". We consider the questions in Wikipedia as false positive answers.

To improve the answer retrieval, we evaluate the second closest sentence instead of the closest candidate that contains "?" at the end of the text. After filtering the false positive answers, there are 300 questions that QBERT-RR can answer. The percentage and number of the questions that can be answered from each type and topic are illustrated in Fig. 9.

Compared to IF questions and confirmation questions (included in type OTHER), WH questions, such as WHICH, WHAT, HOW, WHERE, and WHO, are more likely to be answered by Wikipedia Summary. Due to the QBERT mechanism, the answer is supposed to be one sentence from Wikipedia. In this case, factoid questions, which can be answered with facts expressed in a short and concise sentence, are more likely to be answered. This explains the reason WH questions have higher answer rates. Furthermore, non-factoid questions, like some of the WHY or IF questions, that require more explanation in the answer are harder to match with a one-sentence answer from a knowledge source.

The questions under the topic of Education & Reference, Science & Mathematics, Sports, and Business & Finance are more likely to be answered confidently by Wikipedia Summary. On the other hand, QBERT-RR can only answer around 0.95%, 1.23%, 1.20%, and 1.65% of the questions in Society & Culture, Politics & Government, Health, and Family & Friendships, respectively.

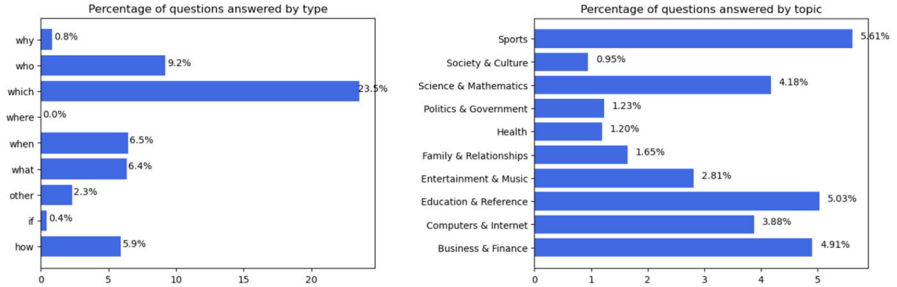
Here are some examples of the answers given by QBERT-RR. The questions are from the WTC corpus, and the answers are extracted from the Wikipedia Summary. For comparison, we also include the top 3 candidate answers.

- **Q1:** Hello is a invented word but what does it mean?

  A1-1: Hello is a greeting in the English language. (Score: 0.895)
  A1-2: Hello is a salutation or greeting in the English language. (Score: 0.867)
  A1-3: It is similar in meaning to the English word Hello!. (Score: 0.845)

- **Q2:** How are clouds made?

  A2-1: On Earth, clouds are formed as a result of saturation of the air when it is cooled to its dew point, or when it gains sufficient moisture (usually in the form of water vapor) from an adjacent source to raise the dew point to the ambient temperature. (Score: 0.873)



(a) Type (b) Topic

Fig. 9 Percentage of the questions answered by QBERT-RR

A2-2: The precise mechanics of how a cloud forms and grows is not completely understood, but scientists have developed theories explaining the structure of clouds by studying the microphysics of individual droplets. (Score: 0.806)
A2-3: A cloud is a visible mass of condensed droplets or frozen crystals suspended in the atmosphere. (Score: 0.778)

- **Q3:** What is a black hole?

  A3-1: A stellar black hole (or stellar-mass black hole) is a black hole formed by the gravitational collapse of a massive star. (Score: 0.920)
  A3-2: Black hole may also refer to: (Score: 0.846)
  A3-3: A black hole is a region of spacetime exhibiting such strong gravitational effects that nothing, not even particles and electromagnetic radiation such as light, can escape from inside it. (Score: 0.837)

- **Q4:** Who discovered pluto?

  A4-1: It was discovered by a team of astronomers from the Institute for Astronomy of the University of Hawaii led by David Jewitt and Scott S. Sheppard and Jan Kleyna in 2001, and given the temporary designation S/2001 J 3. (Score: 0.901)
  A4-2: It was discovered by Nicholas Miller. (Score: 0.897)
  A4-3: It was discovered by Scott S. Sheppard, et al. (Score: 0.895)

From the question-answer pairs retrieved by QBERT-RR, we notice that the retrieved answer is adapted to a specific attribute rather than a general situation as human understanding. From Q3 in the examples, the answer adapts to "stellar black hole" instead of giving an answer to the general black hole.

QBERT-RR is only able to retrieve ONE most relevant sentence as the answer. Hence, the answer is not always included in one sentence. For example, for (Q4, A4), the answer itself might look promising. Nonetheless, the original summary contains A4 is "Hermippe, or Jupiter XXX, is a natural satellite of Jupiter. It was discovered concurrently with Eurydome …and given the temporary designation S/2001 J 3.". In this case, QBERT-RR fails to capture the entity "Hermippe" related to A4, which is also essential for answering the question.

While performing question answering, we observe one particular case. For the question "What is two plus two?", QBERT-RR gives an answer as "2 + 2 = ?". Even though it is not the correct answer to the question, it is still interesting to see that QBERT-RR "translate" the text into notations. This can be further investigated in future research.

Conclusion

People always have curiosity and doubts. Questions have been collected from various sources during the past ten years. Access to information is easy and comprehensible through digital and online channels in our modern world. Educational institutions like schools and museums are the places that come up with most of the curiosity and

questions and have therefore been challenged to adapt to this changing environment. How many things can we do with the questions, and what can we learn from these questions?

With these queries, we introduced QBERT for question content analysis, which can perform question taxonomy, equivalent question recognition, and question answering. The model can be applied to analyse a new question corpus collected by the WTC Science Centre. We observed that the generalist model QBERT-RR turned out to perform similarly to the specialists in QT, QE, and QA.

By applying QBERT, we managed to identify the common questions and answer 463 questions in the WTC corpus. The model also revealed the curiosity of the question contributors by categorising them into 90 themes.

In the results, we see that the contributors to "Project What If" were very interested in Science, Society, and Health; and asked many questions about the WHY and HOW types. This is not surprising within the setting of WTC as a science institution. But the next finding revealed a lot more: questions of the type IF tend to relate to Politics & Government topics and not to Education & Reference topics. Curiosity about Society & Culture is also very revealing. This is an emerging theme in the sector of science centres, where there is an ongoing discussion about expanding from Science Centres to Science & Cultural Centres. More generally, there is a movement in the sector currently to explore society and culture alongside traditional science such as Biology, Engineering, Chemistry etc. This seems to be reflected in the kind of questions Bristolians have been asking.

We believe that question processing specific AI models can benefit educational institutes in the future. For teaching, the teacher can understand the students' comprehension level through the difficulty of the questions. Teachers can also minimise their workload by letting the machine answer some of the questions and figure out the most common and essential problem to work on. With AI software to identify the difficulty of the questions, the students can also use the different levels of quizzes to improve learning, from conceptual questions to questions that require more comprehension of the content, even questions that expand the knowledge beyond the learning content.

This study demonstrated that BERT-based models can comprehend contextual information and contribute to understanding questions. More complicated models, such as GPT (Radford et al., 2018; Brown et al., 2020) and LLaMA (Touvron et al., 2023), can be leveraged in future studies to further analyse questions. Meanwhile, the nature of large language models also makes understanding how these models arrive at specific conclusions or responses more challenging. Future research can also focus on reasoning the decisions made by the models, especially in critical applications.

Acknowledgements We acknowledge Anna Starkey for her central role in "Project What If" and being part of the team who initiated this research, and David May for bringing us all together and providing his insight along the way.

Author Contributions ZX did programming, data analysis, and write up. NC assisted with design of experiments and methods and discussion. AH and NB supported the data collection, assisted with discussion and writing the introduction and conclusion.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Availability of Data The anonymised and moderated WTC dataset is available from We The Curious on reasonable request for research purposes. Please contact information@wethecurious.org.

Declarations

Competing Interests The authors declare that no relevant financial or non-financial interests to disclose.

References

Arora, S., Liang, Y., & Ma, T. (2016). *A simple but tough-to-beat baseline for sentence embeddings*.

Bloom, B.S. (1956). *Taxonomy of educational objectives: The classification of educational goals. Cognitive domain*.

Bowman, S., Angeli, G., Potts, C., & Manning, C.D. (2015). A large annotated corpus for learning natural language inference. *Proceedings of the 2015 conference on empirical methods in natural language processing* pp. 632–642.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & . others,. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877–1901.

Caruana, R. (1997). *Multitask learning. Machine learning, 28*(1), 41–75.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., & Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th international workshop on semantic evaluation (semeval-2017)* pp. 1–14.

Chandrasekaran, D., & Mago, V. (2021). Evolution of semantic similarity-a survey. *ACM Computing Surveys (CSUR), 54*(2), 1–37.

Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). *Reading wikipedia to answer open-domain questions*. arXiv:1704.00051 .

Chin, C., & Brown, D. E. (2000). Learning in science: A comparison of deep and surface approaches. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 37*(2), 109–138.

Chin, C., & Osborne, J. (2008). Students' questions: a potential resource for teaching and learning science. *Studies in science education, 44*(1), 1–39.

Clauset, A., Newman, M. E., & Moore, C. (2004). Finding community structure in very large networks. *Physical review E, 70*(6), 066111.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement, 20*(1), 37–46.

Conneau, A., Kiela, D., Schwenk, H., Barrault, L., & Bordes, A. (2017). *Supervised learning of universal sentence representations from naturallanguage inference data*. arXiv:1705.02364 .

Csernai, K. (2017). *Quora question pairs*. https://www.quora.com/q/quoradata/First-Quora-Dataset-Release-Question-Pairs

Cuccio-Schirripa, S., & Steiner, H. E. (2000). Enhancement and analysis of science question level for middle school students. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching, 37*(2), 210–224.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805

Ebersbach, M., Feierabend, M., & Nazari, K. B. B. (2020). Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning. *Applied Cognitive Psychology, 34*(3), 724–736.

European Parliament, & Council of the European Union. (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council*. Retrieved 13-April-2023, from https://data.europa.eu/eli/reg/2016/679/oj

Ferrucci, D. A. (2012). Introduction to "this is watson". *IBM Journal of Research and Development, 56*(3.4), 1–1.

Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). *Realm: Retrieval-augmented language model pre-training*. arXiv:2002.08909.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. 2006 IEEE computer society conference on computer vision and pattern recognition (cvpr'06) Vol. 2, pp. 1735–1742.

Hagberg, A., Swart, P., & S Chult, D. (2008). *Exploring network structure, dynamics, and function using networkx (Tech. Rep.)*. Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). US: Prentice Hall.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., & . . . Yih, W.-t. (2020). *Dense passage retrieval for open-domain question answering*. arXiv:2004.04906.

Kingma, D.P., & Ba, J. (2015). *Adam: A method for stochastic optimization*. (In ICLR.)

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). *Albert: A lite bert for self-supervised learning of language representations*. arXiv:1909.11942 .

Lee, K., Chang, M.-W., & Toutanova, K. (2019). *Latent retrieval for weakly supervised open domain question answering*. arXiv:1906.00300.

Lehnert, W.G. (1977). A conceptual theory of question answering. *Proceedings of the 5th international joint conference on artificial intelligence* volume 1 pp. 158–164.

Li, X., & Roth, D. (2002). Learning question classifiers. *Coling 2002: The 19th international conference on computational linguistics*.

Liu, X., He, P., Chen, W., & Gao, J. (2019). Multi-task deep neural networks for natural language understanding. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4487–4496.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., & . . . Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv:1907.11692 .

McCann, B., Keskar, N.S., Xiong, C., & Socher, R. (2018). *The natural language decathlon: Multitask learning as question answering*. arXiv:1806.08730

McCreery, C.H., Katariya, N., Kannan, A., Chablani, M., & Amatriain, X. (2020). Effective transfer learning for identifying similar questions: Matching user questions to covid-19 faqs. *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining* pp. 3458–3465.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* pp. 3111–3119.

Mohasseb, A., Bader-El-Den, M., & Cocea, M. (2018). Question categorization and classification using grammar based approach. *Information Processing & Management, 54*(6), 1228–1243.

Pearl, J. (2018). Causal and counterfactual inference. *The Handbook of Rationality*, 1–41.

Pennington, J., Socher, R., & Manning, C.D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* pp. 1532–1543.

Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies*, volume 1 (long papers) pp. 2227–2237. New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N18-120210.18653/v1/N18-1202

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., & Liu, P.J. (2020). Exploring the limits of transfer learning with a unified textto-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Scardamalia, M., & Bereiter, C. (1992). Text-based and knowledge based questioning by children. *Cognition and instruction, 9*(3), 177–199.

Scheepers, T. (2017). *Improving the compositionality of word embeddings (Unpublished master' thesis)*. Universiteit van Amsterdam, Science Park 904, Amsterdam, Netherlands.

Seo, M., Lee, J., Kwiatkowski, T., Parikh, A.P., Farhadi, A., & Hajishirzi, H. (2019). *Real-time open-domain question answering with dense-sparse phrase index*. arXiv:1906.05807 .

Song, D. (2016). Student-generated questioning and quality questions: A. *Research Journal of Educational Studies and Review, 2*(5), 58–70.

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). *How to fine-tune bert for text classification? China national conference on chinese computational linguistics* pp. 194–206.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., & . . . others (2023). *Llama: Open and efficient foundation language models*. arXiv:2302.13971.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *Proceedings of the 2018 emnlp workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* pp. 353–355.

Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., & . . . Wei, F. (2022). *Text embeddings by weakly-supervised contrastive pre-training*. arXiv:2212.03533 .

White, R., & Gunstone, R. (2014). *Probing understanding*. Routledge.

Williams, A., Nangia, N., & Bowman, S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: Human language technologies*, volume 1 (long papers) pp. 1112–1122.

Xu, Z., & Cristianini, N. (2023). Qbert: Generalist model for processing questions. *International symposium on intelligent data analysis* pp. 472–483.

Xu, Z., Howarth, A., Briggs, N., & Cristianini, N., et al. (2021). *What makes us curious? analysis of a corpus of open-domain questions. Cs & it conference proceedings* Vol. 11.

Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., & . . . Lin, J. (2019). *Endto-end open-domain question answering with bertserini*. arXiv:1902.01718.

Yang, Y., Yih, W.-t., & Meek, C. (2015). Wikiqa: A challenge dataset for open-domain question answering. *Proceedings of the 2015 conference on empirical methods in natural language processing* pp. 2013–2018.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *In Advances in neural information processing systems*

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the ieee international conference on computer vision* pp. 19–27.