UNIVERSITY OF CAPE TOWN

DEPARTMENT OF COMPUTER SCIENCE

CS/IT Honours Project Final Paper 2023

Title: Generating Adaptive Learning Materials

Author: Chiranjeev Keshav Nathoo

Project Abbreviation: GALMAT

Supervisor(s): Toky Hajatiana Raboanary, Assoc. Prof. Maria Keet

Category	*Min*	*Max*	Chosen
Requirement Analysis and Design	*0*	*20*	10
Theoretical Analysis	*0*	*25*	0
Experiment Design and Execution	*0*	*20*	15
System Development and Implementation	*0*	*20*	10
Results, Findings and Conclusions	*10*	*20*	15
Aim Formulation and Background Work	*10*	*15*	10
Quality of Paper Writing and Presentation	*10*		10
Quality of Deliverables	*10*		10
Overall General Project Evaluation (*this section allowed only with motivation letter from supervisor*)	*0*	*10*	0
Total marks		**80**	**80**

Knowledge graph-to-text generation using prompting

Chiranjeev Keshav Nathoo
NTHCHI002
University of Cape Town
Cape Town, South Africa
NTHCHI002@myuct.ac.za

ABSTRACT

Generating natural language text from knowledge graphs (KG) requires integrating information from KGs into semantically and syntactically correct sentences. Recent advancements in this domain have hinged on leveraging pre-trained language models (PLMs) due to their established prowess in handling natural language. The generated sentences are, however, just a basic verbalisation of the KG and with no added details for context. In this paper, we propose the use of Large Language Models (LLMs), through prompting, for the KG-to-text generation task. We focus on producing more detailed sentences so that when used in our proof-of-concept adaptive learning system, more informative documents can be produced for the learner. For this task, we experiment with 3 LLMs with varying parameter counts and show that as the parameter count goes up, the LLM produces more reliable outputs while requiring less post-processing. Hence, we contribute to making the learning process easier and improving the quality of education in general.

CCS CONCEPTS

• Natural language processing → Text generation

• Language models → Large language models

KEYWORDS

Knowledge graphs, text generation, large language models, natural language processing

1 INTRODUCTION

Although there is no precise definition of a knowledge graph, it is agreed that it represents entities and their interrelations as a way to store information on one or various topical domains [2]. Most uses of knowledge graphs require some sort of verbalization of the content into a natural language, which mostly happens to be English, so that it can be made more comprehensible for people [1]. This is what the KG-to-text generation task is all about. What makes this task challenging is that the generated text should be accurate and coherent. This means that the text should reflect the information captured within the knowledge graph and should also make sense as a whole. To this end, recent research showed that pre-trained language models (PLMs) can be used to automatically generate text when fed a knowledge graph. The idea of model pre-training is to train a model using a massive corpus of unannotated data so that the knowledge learned can be applied to new tasks [5]. A PLM then uses this history of unannotated texts to predict the next word. The generalisability of PLMs is what made them popular; however, they cannot be used out of the box for any task. PLMs must first be fine-tuned, meaning that they must be additionally trained on some task-specific data. The task-specific data must be manually prepared, which requires knowledge of the inner workings of a PLM and a lot of time to create enough data. Even after all this effort, the output is a simple verbalisation of the KG, often resembling some reference human-written text, with no context on the content.

The research aim of this paper is to find a method to create more detailed and human-like descriptions of knowledge graphs, represented in the form of triples[1], while maintaining a focus on the content. Despite requiring a lot of resources and time to fine-tune a PLM, only slight improvements are noticed and there is visibly no increase in the number of details of the generated text [1]. In search of a better method for the KG-to-text generation task, we will investigate the use of LLMs through prompting. LLMs are nothing but scaled-up PLMs that can be used without any fine-tuning [5]. We show how we can harness that extra knowledge to create more detailed descriptions of triples. The objective behind adding more details is to make the sentences self-explanatory so that the reader does not require any further clarifications on the facts described. A prompt or a prompting function is a text fed to the LLM that contains the user inputs in a specific format and an answer slot which is later filled by the LLM in its output [4]. Properly constructing the prompt allows the LLM to understand the right context, hence producing the right output. Prompting can be called a cheaper and simpler way of getting the LLM to perform a downstream task, compared to fine-tuning. While fine-tuning requires partly re-training the model, prompting is just us inferring from the model.

In this paper we demonstrate the usage of a LLM for the KG-to-text generation task through a proof-of-concept project for generating adaptive learning materials. Adaptive learning systems are an alternative to the traditional 'one size fits all' approach for creating educational materials. They try to personalise the learning experience by taking into consideration the learner's expectations, needs and other traits. They do so through machine

[1] A triple is in the form of <subject, predicate, object>, which is a way to express graph data. The predicate normally represents an edge while the other two are nodes.

learning algorithms and by applying item response theories [3,6]. The project is divided into 3 distinct tasks, see Figure 1. Each task is responsible of one aspect of the adaptive learning system. Task 1 handles automatic question generation (AQG) [7] and each question is dynamically generated based on the learner's answers to previous questions. Those answers are passed on to Task 2 which creates a knowledge graph in the form of triples, representing the concepts that the learner did not understand. Additionally, Task 2 uses Item Response Theory (IRT) [8], which widely used in educational testing and psychological evaluations, to quantify the knowledge of the learner on each concept, essentially adding a weight to each edge of the knowledge graph, which we call the learner ability. See Figure 2.

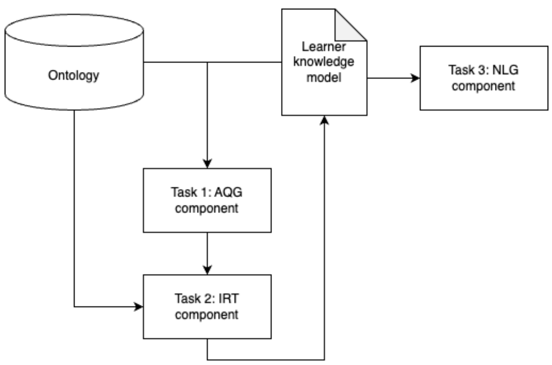


Figure 1: High level architecture of the project. The arrows show the flow of information between the components.

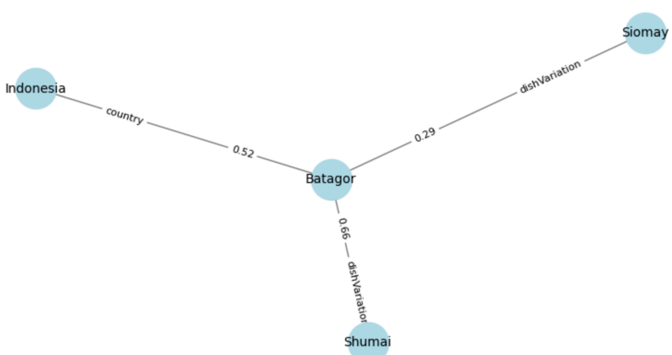


Figure 2: Example of a learner knowledge model. The nodes represent subjects and objects while the edges are the predicates. The weight of each edge is also shown which we know as the learner ability.

Together, the knowledge graph and learner ability form what we call the learner knowledge model. The learner knowledge model is used by Task 3, the LLM-based KG-to-text generator, to create an informative document which describes each concept, based on the value of the learner ability, see Figure 3. This is the task that we are concerned with and where we will demonstrate the usefulness of an LLM.

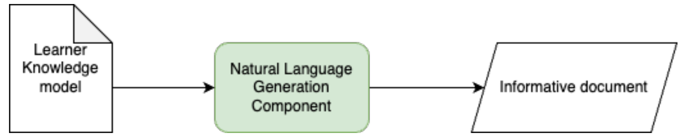


Figure 3: Architecture of Task 3 which takes the learner knowledge model as input and produces an informative document.

The rest of this paper is structured into 5 sections. Section 2 provides an overview of the current research that has been done regarding the use of PLMs for the KG-to-text generation task and prompting. Section 3 is about the methodology we employed to experiment with LLMs and how the evaluation of our method is done. Section 4 shows experimental results and their analysis to show the usefulness of LLMs in the task in question. This is followed by section 5, which draws conclusion from our findings and states what we intend to focus on in the future.

2 RELATED WORK

2.1 KG-to-text generation

Prior to major advancements in machine learning and language models, most methods pertaining to the KG-to-text generation task were centered around the use of templates and rules [9,10]. These methods can produce high quality texts as almost all their components are hand-engineered. Having a high dependence on human intervention is also their biggest disadvantage if we want the text structure and length to be dynamic. Later works explored the use of neural networks. Li et al. [11] used pointer-generator networks to basically automate the creation of templates with entities of the KG slotted in which means that the structure of the text is not flexible. Furthermore, pointer-generator networks are known to suffer from the vanishing gradient problem which hinders the performance of this method. Koncel-Kedziorski et al. [12] focused on encoding the graph structure of knowledge graphs using a transformer encoder-decoder [13] to model structural information. This method addresses the issue of incorporating structure of the KG in the text; however, as the KG gets larger, it might have issues encoding all that data. These methods are not aware of the grammatical structure of sentences since they lack an understanding of the natural language, which leads us to PLMs.

2.2 Pre-trained Language Models

PLMs, unlike the early works, have a basic inner representation of the natural language. This allows for some flexibility in the structure and semantics of the generated text. Guo et al. [14] uses a relational graph convolutional networks (R-GCN) based planner to construct an input with entities of triples placed at optimum positions. Since the PLM is a text-to-text generator and the input is not exactly in natural language, it must be given some sort of context on the task at hand to output relevant text. The PLM is therefore fine-tuned on several such plans. This is essentially guiding the PLM to fill in the blanks between the entities. Li et al.

[15] tries to incorporate the structure of the KG using a graph neural network-based KG encoder; however, fine-tuning is still required so that the PLM learns how to use that knowledge. These methods showed that PLMs can be used to generate textual descriptions of KGs; however, no extra detail is added in addition to verbalizing the triples. Furthermore, fine-tuning requires knowledge about the architecture of the PLM and a considerable amount of computational power which limits the use of methods involving this process to a smaller group. Our work primarily aims at producing more detailed descriptions, but it also has the advantage of being easier to use.

2.3 Prompt engineering

Along with the advent of LLMs, prompt engineering also came to exist and is currently being extensively researched. Prompt engineering can be defined as crafting the most appropriate prompt so that the LLM can produce a task-relevant output [4]. Research shows that there are three distinct types of prompting techniques. Zero-shot prompting [16] is providing the LLM with textual instructions on the task. This type of prompting produces good results if the LLM has a very high parameter count meaning that its knowledge of the world is vast. Where zero-shot prompting does not produce the optimum results, few-shot prompting is beneficial to enhance the performance of the LLM. With few-shot prompting, the LLM is provided with some examples of the output expected from it in addition to the prompt. The LLM is then able to build on the given examples to get a better grasp of the task [17]. The third type of prompting called chain-of-thought prompting allows the LLM to perform a complex task which is decomposed into intermediate reasoning steps [26]. In this paper, we only demonstrate the use of zero-shot and few-shot prompting for the KG-to-text generation task. Prompt engineering also deals with the shape of the prompts of which there are two kinds: cloze prompts and prefix prompts. Cloze prompts are to fill in the blanks of a textual string and prefix prompts, continuing a string prefix [4]. In our work, we will use prefix prompts since we need a description to follow the prompt containing the triples.

3 METHODOLOGY

In this section, we introduce the LLMs used and describe how they were configured for our experiments to test the hypothesis that LLMs can be used for generating detailed descriptions of KGs in the form of triples. We also talk about how evaluation was done to obtain concrete proof supporting our hypothesis. An automatic evaluation using metrics is done to show that larger LLMs can produce more details followed by a human evaluation to validate those details.

3.1 Models

In this research, we evaluate GPT-2 [18], BLOOM-6B [19] and GPT-3 [20] for the KG-to-text generation task. The first two models were obtained through the Hugging Face Transformers library and GPT-3 was accessed through OpenAI's API. Since to infer from these models, some kind of Graphical Processing Unit (GPU) is required, Google Collaboratory was the preferred IDE to use.

Prompting was used to infer from each of the models. Despite the prompts looking slightly different, the high-level structure was kept the same. The triple to be verbalised was embedded in the prompt along with a human-readable textual instruction for the model to describe it. In the case of few-shot prompting, the same steps were followed to provide the examples within the prompt.

3.1.1 Generative Pretrained Transformer - 2 (Fine-tuned)

GPT-2 is an LLM trained with a causal language modelling objective which means that it predicts the next token in a provided sequence of tokens. A token is the smallest unit of text that the model reads. It is to be noted that the use of zero-shot prompts yielded output that, albeit syntactically correct, was irrelevant to the content being represented by the triples, see Table 1.

Table 1: Output from base model GPT-2

Prompt	"\'Indonesia, Country, Batagor\' Description:
Response	'Indonesia, Country, Batagor' Description: The island of Batagor is a small island in the middle of the Indian Ocean. It is located in the middle of the Indian Ocean, and is the only island in the world that has a large population of people. The island is inhabited by a large number of people, and is known as the "Indonesian Islands". The island is inhabited by a large number of people, and is known as the "Indonesian

For our research we experimented with the medium version of GPT-2 which has a parameter count of around 345 million parameters. This makes it the LLM with the smallest parameter count, out of the three investigated models, implying that it is also the easiest to load. This allowed us to investigate what kind of output it produced when fine-tuned. To fine-tune the model, a training dataset had to be prepared. The dataset was structured as follows; each triple was slotted in a question template and that was immediately followed by its reference text description as the answer. A snippet of the training dataset can be found in the Appendix A. The LLM is then trained on this dataset for 50 epochs.

Table 2: Output from fine-tuned GPT-2

Prompt	"[Q] How to describe ('Basko', 'hasIngredient', 'Noodle')?"
Response	A dish that contains Noodle is Bakso.

When inferring from the fine-tuned LLM, the parameters do_sample, max_length, top_k and top_p were set to true, 50, 50 and 0.95 respectively. Configuring these parameters allowed for an effective use of sampling which introduced some randomness.

Each generation was different for the same set of triples. This allowed us do retries of the generation if the text did not contain entities from the triple. We allowed a maximum of 5 retries. Since the LLM produced text in the same format as in the training dataset, some post-processing clean-up was required. Table 2 shows an example of the model's cleaned up output.

3.1.2 BigScience Large Open-science Open-access Multilingual Language Model – 6B

BLOOM is a LLM that was specifically designed for reading comprehension tasks which means that when presented with a prompt, it does a better job in understanding the task at hand and can produce a relevant output. We experiment with the 6 billion parameter count version of the model since this is the largest one that could be loaded in Google Collaboratory. The prompt here is phrased as an instruction rather than a question. Both zero-shot and few-shot prompting were tested, see Table 3. Note that only results from the few-shot prompting were evaluated using metrics since they produced relevant results.

Table 3: Output from BLOOM-6B

Zero-shot prompting	
Prompt	"Describe the triple ('Basko', 'hasIngredient', 'Noodle') in a sentence:"
Response	I have a recipe for baking noodles. I have to make a recipe for baking noodles.
Few-shot prompting	
Prompt	"Describe the triple (Ajoblanco, country, Spain) in a sentence:\n Ajoblanco is from Spain.\nDescribe the triple (Asam_pedas, region, Malay_Peninsula) in a sentence:\n Asam pedas is from the Malay Peninsula region.\nDescribe the triple ('Basko', 'hasIngredient', 'Noodle') in a sentence:"
Response	Bakso is made of noodle.

When inferring from the LLM, the parameters max_new_tokens, temperature, top_k and top_p were set to 50, 2.0, 2000 and 0.1 respectively. Like GPT-2, this made use of sampling and a maximum of 5 retries were allowed. We found that this model was prone to producing duplicate texts, hence some post-processing had to be done to get rid of those.

3.1.3 Generative Pretrained Transformer – 3

GPT-3 is significantly different from GPT-2 although they are just one version apart. GPT-3 has a parameter count of 175 billion and has shown remarkable performance in multiple tasks without any fine-tuning. The size of GPT-3 is another reason why it is preferably accessed through openAI's API instead of loading it in a local system. Unlike the other two models, only zero-shot prompting was tested on GPT-3 as we did not want to constrain the number of details the model would add in the sentences by referring to examples given to it. This LLM was found to sometimes produce incomplete sentences when cut-off due to the imposed token limit imposed by the max_tokens parameter. This required some post-processing to be removed from the final output. Table 4 below shows an example of the model's final output.

Table 4: Output from GPT-3

Prompt	"Triples:\n- 'Basko', 'hasIngredient', 'Noodle'\n\nDescription:"
Response	Bakso is an Indonesian type of meatball or meat paste made from beef and tapioca flour. It usually contains noodles as an ingredient and is usually served in a rich soup.

Since this LLM produced the most detailed texts with the least errors for each triple in the KG, it was chosen to be part of the proof-of-concept adaptive learning system. As input, it received the learner knowledge model as a CSV file from Task 2, see Appendix B. The CSV file is read, and its contents cleaned and placed in a dictionary data structure. The algorithm then loops over each triple and based on its learner ability, a value for the max_tokens parameter is calculated. This allows us to control how much detail the model can add for a particular triple. A snippet of this code is provided in the Appendix F.

3.2 Dataset

All testing that involved the use of KGs in this paper and even in the proof-of-concept project used the food domain of the WebNLG (version 2.1) [21] dataset. More specifically, it is used to create the training data of the fine-tuned GPT-2 model and a food-based ontology that is an input of Task 1 and 2 of the project. The dataset that was specifically designed for training and evaluating systems that perform verbalization of KGs. Most of the data in it has been extracted from Wikipedia and DBpedia which contributes to the variety of domains. The food domain was selected for our uses because it is one of the largest domains in the dataset and contains data about things that more people have a better chance of having a prior knowledge on, hence facilitating our evaluations.

3.3 Evaluation

3.3.1 Automatic evaluation

Since the KG-to-text generation is a field of interest for a lot of researchers, a few evaluation metrics were developed over the years to automate the testing process. Out of them, the Bilingual Evaluation Understudy or BLEU metric, the Recall-Oriented Understudy for Gisting Evaluation or ROUGE metric and the Metric for Evaluation of Translation with Explicit Ordering or METEOR are the most used ones. All three of them evaluate the machine translation text by comparing it to one or more human-written reference texts, with METEOR being slightly different as it requires the sentences to be tokenized beforehand. These metrics however focus more on the structure of the generated sentences and how closely they resemble to the reference sentences rather than on the amount of details added. Hence, as more detail is seen in the generated sentences, a decrease in their similarity with the reference sentences should be observed, thus lowering the metric scores. Therefore, in our case, the lower the metric scores, the better, as this shows us that the model is adding more details to the sentences.

In our paper, we use SacreBLEU [22], ROUGE-L [23] and METEOR [25]. The main reason for using SacreBLEU instead of the original BLEU metric is that tokenization of the generated and reference texts is consistent which makes it more reliable. ROUGE-L evaluates the longest common subsequence between the generated and reference texts. This means that it can capture the overall content and structure of the text. The score is then an indication of how much of the knowledge represented by the KG has been verbalised in the generation. Unlike SacreBLEU and ROUGE-L, METEOR has a stemming algorithm and synonym matching to account for variations in word forms and synonyms. This means that this metric will be the least affected by the addition of details, but a decrease is still expected.

To make the evaluation of each model comparable to each other, all of them were made to produce text descriptions given the learner knowledge model. The description of each triple was then evaluated against its reference text using each metric and an average of the scores were calculated, see Appendix H. It is to be noted that for the ROUGE-L score, we only use the f-measure provided in the 'mid' confidence interval. The f-measure takes both the precision and recall into account and the 'mid' confidence interval provide reliable results while allowing for some minor fluctuations in the score.

3.3.2 Human evaluation

After correctly identifying the model producing the most amount of details for each triple, we use human evaluation to analyse the generated sentences and verify that the added details are valid and useful. This second evaluation is necessary since the metrics only confirm the addition of details and not their validity, which is important for us as the details need to be related to the triples and provide context on the facts. An LLM producing such outputs is required in the project to produce useful informative documents. For this purpose, two surveys were designed. A sample of both is provided in the Appendix I. Both ask questions about the reference texts, the learner knowledge model and the generated descriptions which are provided to the evaluator. They differ in two main aspects: one survey asks general questions while the other contains sentence-specific questions and secondly, in terms of the criteria they evaluate. The general survey evaluates the *quality, relevance, focus and human-likeness* of the generated descriptions while the sentence-specific survey evaluates the *information content* in addition to the correlation between the number of details and the learner ability and its necessity in conveying the facts represented by the triples. As per Howcroft et al. [24], a lot of papers in the NLG field of study that do conduct surveys, do not have a clue of what they are evaluating because they do not have properly defined criteria. To mitigate these issues, we based our definitions on Howcroft et al. [24]'s defined criteria. Here are the criteria, their definitions, and why we specifically chose these:

3.3.2.1 Quality of generated descriptions

Here by quality, we mean whether the sentences are grammatically correct or include false information or not (credibility). Having a sentence of higher quality is a desirable feature in any NLG task.

3.3.2.2 Relevance and focus

Relevant and focused sentences should have information about the triples in question. As with any typical KG-to-text generation task, this should be the primary objective.

3.3.2.3 Human-like quality

A human reader should not be able to distinguish between the machine-generated sentence and an actual human-written sentence. Showing that the generated sentences closely resemble a human-written text would mean that they feel natural to the reader and therefore more enjoyable.

3.3.2.4 Information content of the generated descriptions

This is about how much detailed the generated sentence is and whether it is too much or too little. Although having detailed generated descriptions is our aim, we also want to know how much is too much.

These criteria will allow us to verify that the added details are valid and assess to which extent our research aim of producing more detailed and human-like KG descriptions has been achieved with the use of LLMs.

4 RESULTS AND DISCUSSION

In this section, we present the results of our study on the use of LLMs for the generation of detailed text descriptions of knowledge graphs and discuss their implications concerning our research aim.

Table 5: Average of metrics per model

Models	SacreBLEU	ROUGE-L	METEOR
Fine-tuned GPT-2	20.48	0.461	0.597
BLOOM-6B (Few-shot)	26.05	0.479	0.557
GPT-3	**6.43**	**0.238**	**0.460**

*** Keeping in mind our aim that is to produce more detailed sentences, lower means better here**

Table 5 compares the performance of the models in KG-to-text generation, given the learner knowledge model CSV file as input.

The scores shown here are averages of the scores obtained by each generated sentence. Appendix H shows how these averages were calculated for each model. The results depict the GPT-3 model as having the worst performance despite being the largest model. This shows that it produces sentences with the most amount of details, however, still unverified. The added infromation act as noise for the metrics thus lowering the scores.

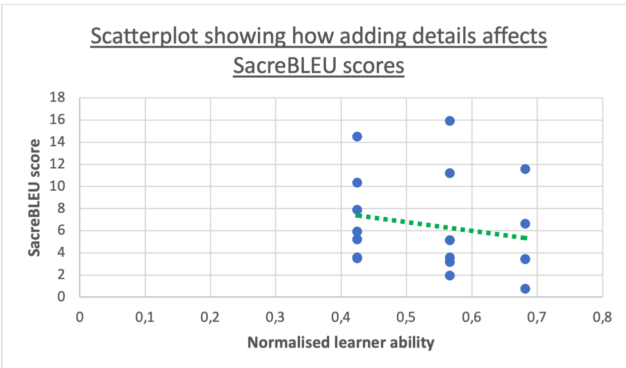


Figure 4: Scatterplot showing how SacreBLEU scores decrease when the generated sentences have more details.

Figure 4 shows a scatterplot of the GPT-3 model for normalised learner ability values of triples plotted against the SacreBLEU score obtained by their corresponding generated sentences. The raw data used to make this scatterplot is in Appendix G and H.3. The learner ability was normalised between a value of 0 and 1 in such a way that when the learner ability is high, the normalised value is low and vise-versa:

```
normalized_ability = (2 - learner_ability) / 4
```

The green trendline in Figure 4 shows us that as the learner ability goes down and the model adds more information, causing the SacreBLEU score to go down.

When comparing the fine-tuned GPT-2 and BLOOM-6B, their ROUGE-L and METEOR scores are almost the same and approximately 0.5, indicating that both produce sentences that encompasses the facts represented by the triple well which can be accepted as a valid verbalisation most of the times. The higher SacreBLEU score of BLOOM-6B indicates that its generated sentences resemble the closest to the reference sentences out of the two models. A visual inspection of the reference sentences in Appendix G confirms that. This result also demonstrates how powerful the use of few-shot prompting with LLMs can be. The model was able to understand the task at hand with just a few provided examples while if fine-tuning was used, training would have been required.

Table 3 shows how few-shot prompting provided the required context that the model needed to properly execute the triple verbalisation. If the objective was to solely get a near perfect verbalisation of the triples, then BLOOM-6B would be the obvious choice. Additionally, since BLOOM-6B is a model trained on multiple languages, it can potentially be used to make those generated sentences available in other languages, something that can prove to be useful in later stages of the proof-of-concept project. Another interesting observation from the comparison of the models' output, is that as the parameter count goes up, the model needs less input to understand the task and can also produce more detailed sentences, see Appendix C, D, E. This is in line with Radford et al.'s observation [18]. This strengthens our hypothesis that LLMs are suitable for generating detailed descriptions of KGs.

As the automatic evaluation metrics can only identify whether the model is injecting details or not and do not provide any information on the usefulness of the said details, the inclusion of human evaluation was a must. In our situation, this was in the form of anonymous surveys. This was conducted on the output of the GPT-3 model as it obtained the lowest metrics score out of the three models suggesting that it injects the most details. The surveys had a total of 5 participants. Both surveys contain only two types of questions; multiple choice (MCQ) and rating scale-based as both types offers the evaluator answer options which provides us with more consistent responses compared to open-ended questions. The questions were created according to Fink et al. [27]'s guidelines to further mitigate confusion for the evaluator. Appendix J contains a summary of the responses where the answers for the MCQs are colour-coded and numbered while for the rating scale based there is a bar chart along with the average rating obtained.

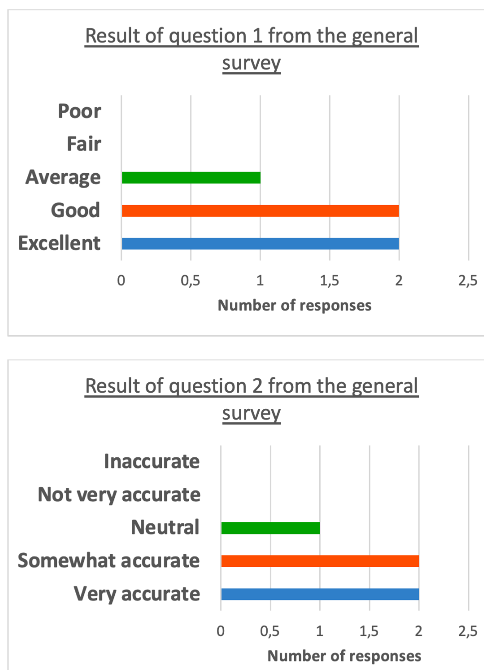


Figure 5: Results for questions 1 and 2 from the general survey, evaluating grammaticality and credibility of detail respectively.

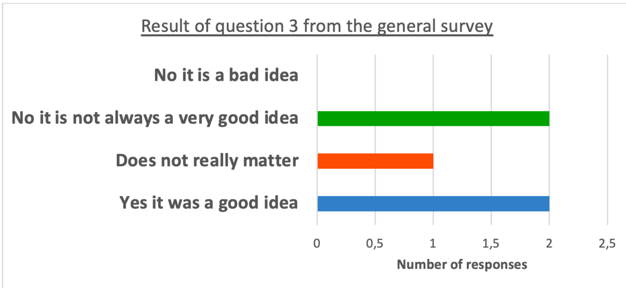


Figure 6: Result for question 3 from the general survey which evaluates whether adding extra details when learner ability is low, a good idea or not.

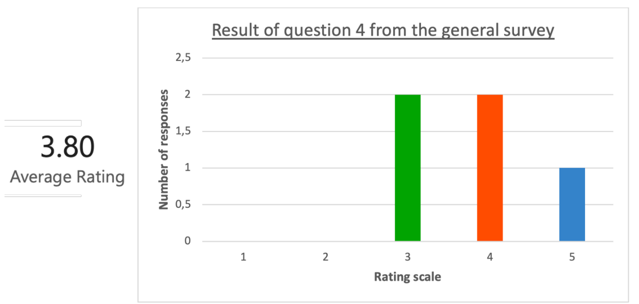


Figure 7: Result for question 4 from the general survey which evaluates how human-like the generated sentences are.

The general survey revealed that participants found the sentences to be grammatically and factually correct (see Figure 5); however, are not very human-like, given an average score of 3.8 out of 5 in question 4, see Figure 7. This suggests that the structure of sentences should not be let entirely on the model to decide. The mixed response received for question 3 suggests that adding extra details when the learner ability is low is mostly desirable; however, further research on when to add more details is needed, see Figure 6. A comment on question 3 also pointed out that a slow learner might not be able to digest all the added details and lose sight of the main concept that the triple represents.

From the sentence-specific survey, the sentences were found to be very focused on the content of the triples with the details being mostly related to them. The exception to this observation were the highly detailed sentences which further points towards the need of a better method to decide where and when to insert details, see Figure 8.

Reference sentence	Bacon Explosion is eaten in the Kansas City metropolitan area.			
Triple and associated learner ability	Subject BaconExplosion	Predicate isEatenInRegion	Object KansasCityMetropolitanArea	Learner ability -0.268
Generated sentence	The BaconExplosion is a dish popularized in Kansas City that consists of bacon woven into two layers around a core of sausage and seasonings, and is typically smoked or grilled. It is often eaten in the Kansas City metropolitan area.			

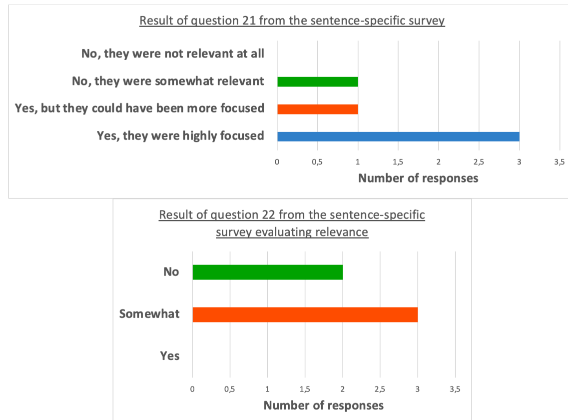


Figure 8: Results for questions evaluating focus and relevance of a highly detailed sentence in the sentence-specific survey. The generated sentence and the triple are shown along with the results.

Table 6: Average rating received by sentences to quantify their correlation between learner ability and level of detail. The question number of every occurrence of this question is also provided here.

Question (See Figure 9)	Average rating (out of 5)
3	4.20
7	4.00
10	4.00
13	3.40
16	4.00
20	4.20
23	4.40
Average	4.03

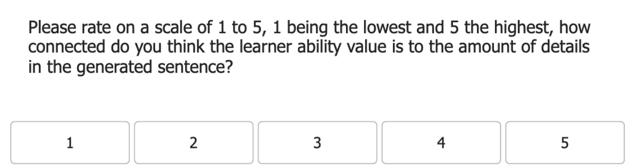


Figure 9: Question in the sentence-specific survey asking about correlation between learner ability and level of detail. This question was asked for each sentence.

Table 6 shows the average rating obtained by each generated sentences presented to the evaluators. See Appendix I.2 to view the generated sentences that were evaluated. The same question is asked for each of the sentences to quantify how related the learner ability value is to the number of details in reference sentence, see Figure 9. The average of the ratings for each question is 4.03 out of 5, showing that the correlation between the two is very good and the model can properly use the learner ability value. Coupling this result with the positive response we got for how necessary the details are to understand the concept suggests that adding details based on the learner ability is appreciated, see Figure 10.

Although the results show that the use of LLMs are very promising for the KG-to-text generation, there are still some limitations. As the parameter count of the LLM goes up and the quality of the output improves, so does the amount of resources required to load up the model. We therefore need to find a balance between what LLM is practically useable, given the resources available to us and the quality of the output we want to achieve. The knowledge that the models have is also not uniform on all topics. This is because the data on which they are trained were mostly web-scrapped, which is known to be skewed. This leads to some descriptions being inherently more detailed than others.

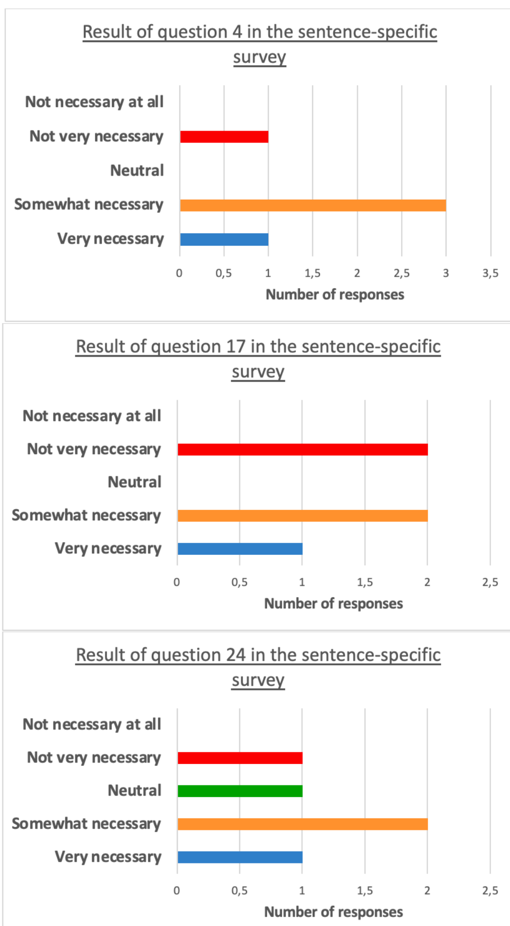


Figure 10: Results of questions in the sentence-specific survey evaluating how necessary the details were. These questions were only asked for highly detailed sentences.

5 CONCLUSIONS AND FUTURE WORK

In this paper we investigated the generation of detailed and human-like descriptions of KGs in the form of triples using LLMs. Through our experiments, we demonstrated that an LLM with a high parameter count such as the GPT-3 model we tested, can be used to generate detailed verbalisation of triples. Our human evaluation showed that those generated sentences have good syntax and semantics, and the added details were relevant to the concepts represented by the triples. Those details provide useful context for the reader to easily understand the concepts. The surveys also showed that there is a need to improve the human-likeness of the generated sentences and to find a method that determines when details must be added and not just using the learner ability. Through the project, we showed that a model like GPT-3 can be successfully used to create adaptive informative documents for learners and contribute to improving the quality of education.

The results of the work done in this research also present several opportunities for further investigation. This study was restricted to only three models and one domain from the WebNLG dataset. This can be extended to test more models with other datasets to get a sense of how much the model can generalise. The performance of the BLOOM-6B model when used with few-shot prompting is noteworthy; however, the examples of the task given to it contained only simple verbalisations of the triples which are sourced from the WebNLG dataset. The use of a dataset that contains more detailed reference texts can potentially allow BLOOM-6B to also produce detailed descriptions of the triples. Currently, the LLMs verbalise only a triple at a time. They can be fed several triples at once to test whether they are able to produce sentences that maintain the structure of the KGs as this is also an important aspect of the KG-to-text generation task. All the generated sentences in this study were in English; however, learning materials are used all around the world. Investigating how the sentences can be translated into other natural languages will contribute to making adaptive learning materials more accessible and a larger version of BLOOM can be used as a starting point since it is a multilanguage model.

ACKNOWLEDGMENTS

We would like to acknowledge Toky Hajatiana Raboanary and Assoc. Prof. Maria Keet for their continued guidance, feedback, and mentorship, as supervisors, throughout the research project. We also extend our heartfelt thanks to our families and friends for their unwavering support during this endeavor.

REFERENCES

[1] Anthony Colas, Mehrdad Alvandipour, and Daisy Zhe Wang. 2022. GAP: A Graph-aware Language Model Framework for Knowledge Graph-to-Text Generation. April 2022. Retrieved from http://arxiv.org/abs/2204.06674

[2] Ehrlinger, L., & Wöß, W. 2016. Towards a definition of knowledge graphs. SEMANTiCS (Posters, Demos, SuCCESS), 48(1-4), 2. Retrieved from http://www.semantic-web-journal.net/content/

[3] Soukaina Ennouamani and Zouhir Mahani. 2017. "An overview of adaptive e-learning systems", 2017 Eighth International Conference on Intelligent Computing and Information Systems: (ICICIS 2017): proceedings. Cairo, Egypt. DOI: https://doi.org/10.1109/INTELCIS.2017.8260060

[4] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. (July 2021). Retrieved from http://arxiv.org/abs/2107.13586

[5] Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2022. Pre-Trained Language Models and Their Applications. Engineering (September 2022). DOI: https://doi.org/10.1016/j.eng.2022.04.024

[6] Shuai Wang, Claire Christensen, Wei Cui, Richard Tong, Louise Yarnall, Linda Shear, and Mingyu Feng. 2023. When adaptive learning is effective learning: comparison of an adaptive learning system to teacher-led instruction. Interactive Learning Environments 31, 2 (2023), 793–803. DOI: https://doi.org/10.1080/10494820.2020.1808794

[7] Kurdi, G., Leo, J., Parsia, B., Sattler, U., & Al-Emari, S. 2020. A systematic review of automatic question generation for educational purposes. International Journal of Artificial Intelligence in Education, 30, 121-204.

[8] Embretson, S. E., & Reise, S. P. 2013. Item response theory. Psychology Press.

[9] Flanigan, J., Dyer, C., Smith, N. A., & Carbonell, J. G. 2016, June. Generation from abstract meaning representation using tree transducers. In Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Human language technologies (pp. 731-739).

[10] Konstas, I., & Lapata, M. 2013, October. Inducing document plans for concept-to-text generation. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1503-1514).

[11] Li, W., Peng, R., Wang, Y., & Yan, Z. 2020. Knowledge graph based natural language generation with adapted pointer-generator networks. Neurocomputing, 382, 174-187.

[12] Koncel-Kedziorski, R., Bekal, D., Luan, Y., Lapata, M., & Hajishirzi, H. 2019. Text generation from knowledge graphs with graph transformers.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. 2017. Attention is all you need. Advances in neural information processing systems, 30.

[14] Guo, Q., Jin, Z., Dai, N., Qiu, X., Xue, X., Wipf, D., & Zhang, Z. 2020. $\mathcal{P}2$: A Plan-and-Pretrain Approach for Knowledge Graph-to-Text Generation: A Plan-and-Pretrain Approach for Knowledge Graph-to-Text Generation. In Proceedings of the 3rd International Workshop on Natural Language Generation from the Semantic Web (WebNLG+) (pp. 100-106).

[15] Li, J., Tang, T., Zhao, W. X., Wei, Z., Yuan, N. J., & Wen, J. R. 2021. Few-shot knowledge graph-to-text generation with pretrained language models.

[16] van de Kar, M., Xia, M., Chen, D., & Artetxe, M. 2022. Don't Prompt, Search! Mining-based Zero-Shot Learning with Language Models.

[17] Lazaridou, A., Gribovskaya, E., Stokowiec, W., & Grigorev, N. 2022. Internet-augmented language models through few-shot prompting for open-domain question answering.

[18] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. 2019. Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.

[19] Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. 2022. Bloom: A 176b-parameter open-access multilingual language model.

[20] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

[21] Gardent, C., Shimorina, A., Narayan, S., & Perez-Beltrachini, L. 2017, September. The WebNLG challenge: Generating text from RDF data. In Proceedings of the 10th International Conference on Natural Language Generation (pp. 124-133).

[22] Post, M. 2018. A call for clarity in reporting BLEU scores.

[23] Lin, C. Y. 2004, July. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

[24] Howcroft, D., Belz, A., Clinciu, M., Gkatzia, D., Hasan, S., Mahamood, S., Mille, S., Van Miltenburg, E., Santhanam, S., & Rieser, V. 2020. Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In 13th International Conference on Natural Language Generation 2020 (pp. 169–182).

[25] Banerjee, S., & Lavie, A. 2005, June. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (pp. 65-72).

[26] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q., Zhou, D., & others. 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837.

[27] Fink, A. 2003. How to ask survey questions (Vol. 1). Sage.

APPENDIX

A GPT-2 training dataset

```
[Q] How to describe (Ajoblanco, country, Spain) ?
[A] Ajoblanco originates from the country of Spain.
[Q] How to describe (Ajoblanco, ingredient, Almond) ?
[A] Ajoblanco has almond as one of its ingredients.
[Q] How to describe (Ajoblanco, ingredient, Bread) ?
[A] Bread is an ingredient of Ajoblanco.
[Q] How to describe (Ajoblanco, ingredient, Garlic) ?
[A] An ingredient of Ajoblanco is garlic.
[Q] How to describe (Ajoblanco, ingredient, Olive_oil) ?
[A] Olive oil is an ingredient in Ajoblanco.
[Q] How to describe (Ajoblanco, ingredient, Water) ?
[A] Ajoblanco contains water.
[Q] How to describe (Ajoblanco, region, Andalusia) ?
[A] Ajoblanco is a food found in Andalusia.
[Q] How to describe (Almond, division, Flowering_plant) ?
[A] Almond is classed as a flowering plant.
[Q] How to describe (Almond, family, Rosaceae) ?
[A] Almond is part of the Rosaceae family.
[Q] How to describe (Almond, order, Rosales) ?
[A] Almond is one of the members of the Rosales order.
[Q] How to describe (Almond, order, Rosids) ?
[A] Almond is part of the order of Rosids.
[Q] How to describe (Amatriciana_sauce, country, Italy) ?
[A] Italy is the country Amatriciana sauce comes from.
[Q] How to describe (Amatriciana_sauce, course, Italian_meal_structure) ?
[A] Amatriciana sauce can be found in an Italian meal.
```

Figure 11: Example of training data used for fine-tuning GPT-2

B Learner Knowledge model

```
Subject,Predicate,Object,Learner Ability
Bakso,hasIngredient,Noodle,0.3
Bakso,hasIngredient,Vermicelli,0.3
Bakso,hasIngredient,Celery,0.3
Bakso,isEatenInCountry,NationwideInIndonesia,0.3
Bakso,isEatenInCountry,ChineseCuisine,0.3
Bakso,isEatenInRegion,NeighboringSoutheastAsianCountries,0.3
Bakso,isEatenInRegion,NationwideInIndonesia,0.3
BarnyCakes,hasIngredient,SpongeCake,-0.732
BarnyCakes,isEatenInCountry,France,-0.732
BarnyCakes,isCreatedBy,MondelezInternational,-0.732
BarnyCakes,hasDishVariation,Chocolate,-0.732
BarnyCakes,hasDishVariation,Apple,-0.732
BaconExplosion,hasIngredient,Sausage,-0.268
BaconExplosion,hasIngredient,Bacon,-0.268
BaconExplosion,isEatenInCountry,UnitedStates,-0.268
BaconExplosion,hasMainIngredients,Sausage,-0.268
BaconExplosion,hasMainIngredients,Bacon,-0.268
BaconExplosion,isEatenInRegion,KansasCityMetropolitanArea,-0.268
BaconExplosion,course,MainCourse,-0.268
```

Figure 12: Example of the CSV file storing the learner knowledge model

C Output of fine-tuned GPT-2

Table 7: Output of fine-tuned GPT-2 for given triples

Prompt		"[Q] How to describe ({subject}, {predicate}, {obj})?"
Triple		Response
1	'Bakso', 'hasIngredient', 'Noodle'	A dish that contains Noodle is Bakso.
2	'Bakso', 'hasIngredient', 'Vermicelli'	Vermicelli is an ingredient in Bakso.
3	'Bakso', 'hasIngredient', 'Celery'	Celery is a member of the genus Bakso.
4	'Bakso', 'isEatenInCountry', 'NationwideInIndonesia'	Bakso is eaten nationwide in Indonesia, and also in Malaysia and Singapore.
5	'Bakso', 'isEatenInCountry', 'ChineseCuisine'	Bakso is a dish from the Country of China.
6	'Bakso', 'isEatenInRegion', 'NeighboringSoutheastAsianCountries'	Bakso is a dish from the region of Singapore and the neighbouring Southeast Asian countries.
7	'Bakso', 'isEatenInRegion', 'NationwideInIndonesia'	Bakso is made from noodles, cuttlefish or squid, and has an ingredient widely found in Indonesia.
8	'BarnyCakes', 'hasIngredient', 'SpongeCake'	Sponge cakes contain an ingredient that a lemon belongs to.

D Output of BLOOM-6B

D.1 Zero-shot prompting

Table 8: Output of BLOOM-6B for given triples in a zero-shot setting

Prompt		"Describe the triple ({subject}, {predicate}, {obj}) in a sentence:"
Triple		Response
1	'Bakso', 'hasIngredient', 'Noodle'	I have a recipe for baking noodles. I have to make a recipe for baking noodles
2	'Bakso', 'hasIngredient', 'Vermicelli'	I have a recipe for baking bread with a mixture of flour, yeast, sugar, salt, and water. I also have a recipe for making a sauce with tomato paste, onion, garlic, and red pepper. I also have
3	'Bakso', 'hasIngredient', 'Celery'	I have a recipe for baking a cake with a mixture of flour, sugar, eggs, and butter. I have a recipe for
4	'Bakso', 'isEatenInCountry', 'NationwideInIndonesia'	The Bakso is eaten in Indonesia
5	'Bakso', 'isEatenInCountry', 'ChineseCuisine'	The Bakso is a meatball dish that is popular in Indonesia. It is made from pork meat, soy sauce, and spices. The Bakso is usually served with rice and vegetables. The

		Bakso is also popular in the Philippines,
6	'Bakso', 'isEatenInRegion', 'NeighboringSoutheastAsianCountries'	The triple (Bakso, isEatenInRegion, NeighboringSoutheastAsianCountries) describes the food Bakso is eaten in the region of Southeast Asia. The food Bakso is eaten in the region of Southeast Asia is
7	'Bakso', 'isEatenInRegion', 'NationwideInIndonesia'	The Bakso is eaten in the region of Indonesia
8	'BarnyCakes', 'hasIngredient', 'SpongeCake'	BarnyCakes is a cake that has a sponge cake as an ingredient

D.2 Few-shot prompting

Table 9: Output of BLOOM-6B for given triples in a few-shot setting

Prompt		"Describe the triple (Ajoblanco, country, Spain) in a sentence:\n Ajoblanco is from Spain.\nDescribe the triple (Asam_pedas, region, Malay_Peninsula) in a sentence:\n Asam pedas is from the Malay Peninsula region.\nDescribe the triple ({subject}, {predicate}, {obj}) in a sentence:"
Triple		Response
1	'Bakso', 'hasIngredient', 'Noodle'	Bakso is made of noodle
2	'Bakso', 'hasIngredient', 'Vermicelli'	Bakso is made with Vermicelli
3	'Bakso', 'hasIngredient', 'Celery'	Bakso is made with celery
4	'Bakso', 'isEatenInCountry', 'NationwideInIndonesia'	Bakso is eaten in Indonesia nationwide
5	'Bakso', 'isEatenInCountry', 'ChineseCuisine'	Bakso is eaten in China
6	'Bakso', 'isEatenInRegion', 'NeighboringSoutheastAsianCountries'	Bakso is eaten in neighboring Southeast Asian countries
7	'Bakso', 'isEatenInRegion', 'NationwideInIndonesia'	Bakso is eaten in Indonesia nationwide
8	'BarnyCakes', 'hasIngredient', 'SpongeCake'	BarnyCakes has sponge cake as its main ingredient

E Output of GPT-3

Table 10: Output of GPT-3 for given triples in a zero-shot setting

Prompt		"Triples:\n-{subject}, {predicate}, {obj}\n\nDescription:"
Triple		Response
1	'Bakso', 'hasIngredient', 'Noodle'	Bakso is an Indonesian type of meatball or meat paste made from beef and tapioca flour. It usually contains noodles as an ingredient and is usually served in a rich soup.
2	'Bakso', 'hasIngredient', 'Vermicelli'	Bakso is an Indonesian-style meatball or meatball-like food that typically contains Vermicelli, a type of thin noodle, as an ingredient.
3	'Bakso', 'hasIngredient', 'Celery'	Bakso is a popular Indonesian meatball dish that contains celery as one of its ingredients.
4	'Bakso', 'isEatenInCountry', 'NationwideInIndonesia'	Bakso is a type of meatball-based Indonesian food that is eaten in Indonesia nationwide.
5	'Bakso', 'isEatenInCountry', 'ChineseCuisine'	Bakso is a type of soup that is eaten in Chinese cuisine.
6	'Bakso', 'isEatenInRegion', 'NeighboringSoutheastAsianCountries'	A type of Indonesian meatball dish, Bakso is typically served as soup and is popular in Neighboring Southeast Asian countries such as Malaysia, Singapore, Brunei, and Cambodia.
7	'Bakso', 'isEatenInRegion', 'NationwideInIndonesia'	Bakso is a type of Indonesian meatball that is widely eaten across Indonesia's various regions.
8	'BarnyCakes', 'hasIngredient', 'SpongeCake'	BarnyCakes is a special recipe that contains the ingredient Sponge Cake.

F GPT-3 prompting code snippet

```python
# Method 2 to calculate a value for max_tokens
def calculate_max_tokens(learner_ability, min_maxTokens):
    normalized_ability = (2 - learner_ability) / 4
    max_tokens = min_maxTokens[0] + normalized_ability * (min_maxTokens[1] - min_maxTokens[0])
    return int(max_tokens)
```

```python
import openai

# Set up your OpenAI API key
openai.api_key = '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'

# Paramenters for the model
maxTokens = 0    # Initialise variable
min_maxTokens = (25, 100) # State range of acceptable values for maxTokens

# Create a formatted prompt with the triples and generate their respective descriptions

# Array to store all generated descriptions
sentences = []

# Construct the prompt for each triple
for triple in triples:
  prompt = "Triples:"
  subject, predicate, obj, learner_ability = triple

  """
  # Method 1 to select a value for maxTokens

  # Learner ability less than -1, learner barely understands the concept, maxTokens is increased
  if (float(learner_ability) < -1):
    maxTokens = 100
  # Learner ability is more than 1, learner understands the concept quite well, maxTokens is decreased
  elif (float(learner_ability) > 1):
    maxTokens = 25
  else:
    maxtokens = 75
  """

  # Uses previously defined function to determine a value for maxTokens - method 2
  maxTokens = calculate_max_tokens(float(learner_ability), min_maxTokens)

  # If the learner ability is 2, the student fully understands the concept hence that triple is skipped
  if (float(learner_ability) == 2):
    continue

  # Craft an appropriate prompt text for our task
  prompt += f"\n- {subject}, {predicate}, {obj}"
  prompt += "\n\nDescription:"

  # Generate a description using a version of GPT 3
  response = openai.Completion.create(
      engine="text-davinci-003",
      prompt=prompt,
      max_tokens=maxTokens
  )

  # Extract the generated description from the response
  description = response.choices[0].text.strip()

  # Store description in array
  sentences.append(description)
```

Figure 13: Code snippet showing how the GPT-3 model was prompted

G Reference sentences

Table 11: Human-written reference sentences along with their learner abilities used for evaluation

No.	Reference Description	Learner Ability
1	Noodle is an ingredient of Bakso.	0.3
2	Vermiceli is an ingredient of Bakso.	0.3
3	Celery is an ingredient of Bakso.	0.3
4	Bakso is eaten all over Indonesia.	0.3
5	Bakso is from the Chinese cuisine.	0.3
6	Bakso is eaten in the neighbouring Southeast Asian countries.	0.3
7	Bakso is a food eaten in Indonesia.	0.3
8	Barny cakes are made with sponge cake.	-0.732
9	Barny cakes can be found in France.	-0.732
10	Barny cakes is created by Mondelez International.	-0.732
11	Barny Cakes can be chocolate flavoured.	-0.732
12	Barny cakes can be made with apple.	-0.732
13	Sausage is an ingredient of Bacon Explosion.	-0.268
14	Bacon is an ingredient of Bacon Explosion.	-0.268
15	Bacon Explosion is eaten in the United States.	-0.268
16	Sausage is a main ingredient of Bacon Explosion.	-0.268
17	Bacon is a main ingredient of Bacon Explosion.	-0.268
18	Bacon Explosion is eaten in the Kansas City metropolitan area.	-0.268
19	Bacon Explosion is a main course.	-0.268

H Metrics evaluation results

H.1 Fine-tuned GPT-2

Table 12: Results of the metrics evaluation of each sentence generated by the fine-tuned GPT-2 model

Results for fine-tuned GPT-2			
Sentence Number	SacreBLEU	ROUGE-L	METEOR
1	15.62	0.429	0.528
2	30.74	0.615	0.691
3	14.25	0.533	0.619
4	12.01	0.421	0.579
5	12.55	0.500	0.611
6	33.89	0.560	0.749
7	9.48	0.333	0.532
8	4.45	0.111	0.181
9	29.00	0.666	0.810
10	13.91	0.286	0.654
11	13.89	0.714	0.894
12	100.00	0.933	0.999
13	2.28	0.160	0.163
14	4.77	0.250	0.316
15	36.28	0.625	0.662
16	7.35	0.190	0.454
17	10.55	0.353	0.417
18	4.25	0.333	0.559
19	33.93	0.750	0.935
Average	20.48	0.461	0.597

H.2 BLOOM-6B (Few-shot)

Table 13: Results of the metrics evaluation of each sentence generated by the BLOOM-6B model under a few-shot setting

Results for BLOOM-6B (few-shot)			
Sentence Number	SacreBLEU	ROUGE-L	METEOR

1	9.72	0.333	0.362
2	9.04	0.167	0.217
3	9.04	0.167	0.290
4	27.78	0.615	0.637
5	15.21	0.333	0.370
6	53.42	0.778	0.904
7	27.89	0.714	0.712
8	10.55	0.250	0.316
9	21.65	0.308	0.481
10	70.14	0.714	0.758
11	6.57	0.143	0.141
12	11.04	0.133	0.234
13	13.13	0.400	0.639
14	8.05	0.429	0.316
15	75.06	0.941	0.999
16	6.50	0.267	0.581
17	6.89	0.500	0.713
18	48.96	0.900	0.914
19	64.35	1.000	0.999
Average	26.05	0.479	0.557

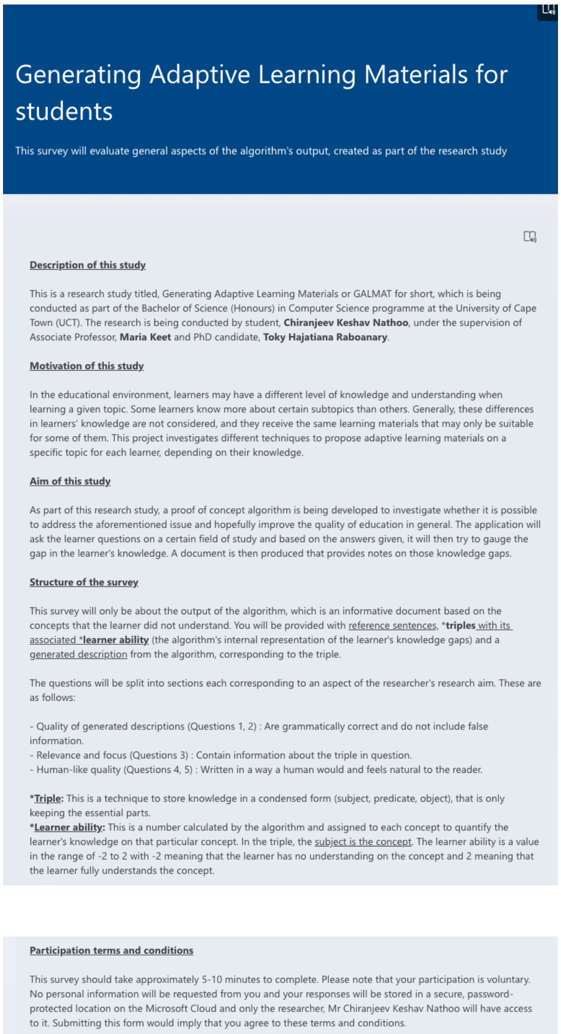
H.3 GPT-3

Table 14: Results of the metrics evaluation of each sentence generated by the GPT-3 model

Results for GPT-3			
Sentence Number	**SacreBLEU**	**ROUGE-L**	**METEOR**
1	3.52	0.158	0.500
2	5.15	0.207	0.496
3	3.46	0.182	0.380
4	7.86	0.364	0.477
5	14.46	0.421	0.587
6	5.91	0.324	0.576
7	10.34	0.417	0.337
8	3.39	0.211	0.351
9	11.50	0.200	0.525
10	6.59	0.222	0.469
11	3.38	0.100	0.260
12	0.71	0.031	0.188
13	5.09	0.231	0.469
14	1.93	0.163	0.437
15	11.18	0.292	0.633
16	3.58	0.258	0.452
17	3.12	0.195	0.535
18	15.89	0.353	0.695
19	5.08	0.196	0.374
Average	6.43	0.238	0.460

I Surveys

I.1 General



Generating Adaptive Learning Materials for students

This survey will evaluate general aspects of the algorithm's output, created as part of the research study

Description of this study

This is a research study titled, Generating Adaptive Learning Materials or GALMAT for short, which is being conducted as part of the Bachelor of Science (Honours) in Computer Science programme at the University of Cape Town (UCT). The research is being conducted by student, **Chiranjeev Keshav Nathoo**, under the supervision of Associate Professor, **Maria Keet** and PhD candidate, **Toky Hajatiana Raboanary**.

Motivation of this study

In the educational environment, learners may have a different level of knowledge and understanding when learning a given topic. Some learners know more about certain subtopics than others. Generally, these differences in learners' knowledge are not considered, and they receive the same learning materials that may only be suitable for some of them. This project investigates different techniques to propose adaptive learning materials on a specific topic for each learner, depending on their knowledge.

Aim of this study

As part of this research study, a proof of concept algorithm is being developed to investigate whether it is possible to address the aforementioned issue and hopefully improve the quality of education in general. The application will ask the learner questions on a certain field of study and based on the answers given, it will then try to gauge the gap in the learner's knowledge. A document is then produced that provides notes on those knowledge gaps.

Structure of the survey

This survey will only be about the output of the algorithm, which is an informative document based on the concepts that the learner did not understand. You will be provided with reference sentences, ***triples** with its associated ***learner ability** (the algorithm's internal representation of the learner's knowledge gaps) and a generated description from the algorithm, corresponding to the triple.

The questions will be split into sections each corresponding to an aspect of the researcher's research aim. These are as follows:

- Quality of generated descriptions (Questions 1, 2) : Are grammatically correct and do not include false information.
- Relevance and focus (Questions 3) : Contain information about the triple in question.
- Human-like quality (Questions 4, 5) : Written in a way a human would and feels natural to the reader.

***Triple:** This is a technique to store knowledge in a condensed form (subject, predicate, object), that is only keeping the essential parts.
***Learner ability:** This is a number calculated by the algorithm and assigned to each concept to quantify the learner's knowledge on that particular concept. In the triple, the subject is the concept. The learner ability is a value in the range of -2 to 2 with -2 meaning that the learner has no understanding on the concept and 2 meaning that the learner fully understands the concept.

Participation terms and conditions

This survey should take approximately 5-10 minutes to complete. Please note that your participation is voluntary. No personal information will be requested from you and your responses will be stored in a secure, password-protected location on the Microsoft Cloud and only the researcher, Mr Chiranjeev Keshav Nathoo will have access to it. Submitting this form would imply that you agree to these terms and conditions.

Reference sentences:

1 Vermiceli is an ingredient of Bakso.
2 Celery is an ingredient of Bakso.
3 Bakso is eaten all over Indonesia.
4 Barny cakes is created by Mondelez International.
5 Barny Cakes can be chocolate flavoured.
6 Bacon Explosion is eaten in the United States.
7 Sausage is a main ingredient of Bacon Explosion.
8 Bacon Explosion is eaten in the Kansas City metropolitan area.

Triples:

	Subject	Predicate	Object	Learner Ability
1	Bakso	hasIngredient	Vermicelli	0.3
2	Bakso	hasIngredient	Celery	0.3
3	Bakso	isEatenInCountry	NationwideInIndonesia	0.3
4	BarnyCakes	isCreatedBy	MondelezInternational	-0.732
5	BarnyCakes	hasDishVariation	Chocolate	-0.732
6	BaconExplosion	isEatenInCountry	UnitedStates	-0.268
7	BaconExplosion	hasMainIngredients	Sausage	-0.268
8	BaconExplosion	isEatenInRegion	KansasCityMetropolitanArea	-0.268

Generated descriptions:

1	Bakso is an Indonesian-style meatball or meatball-like food that typically contains Vermicelli, a type of thin noodle, as an ingredient.
2	Bakso is a popular Indonesian meatball dish that contains celery as one of its ingredients.
3	Bakso is a type of meatball-based Indonesian food that is eaten in Indonesia nationwide.
4	BarnyCakes are a type of cake made and distributed by Mondelez International, a global confectionary, food, and beverage company.
5	BarnyCakes is a type of cake that has a variation made with chocolate.
6	The BaconExplosion is a dish that consists of bacon woven into a lattice pattern with sausage and various seasonings, wrapped in bacon strips and smoked, typically over a wood-burning grill. It is typically eaten in the United States.
7	A BaconExplosion is a type of food dish which primarily consists of ingredients such as bacon, sausage, and other herbs and spices.
8	The BaconExplosion is a dish popularized in Kansas City that consists of bacon woven into two layers around a core of sausage and seasonings, and is typically smoked or grilled. It is often eaten in the Kansas City metropolitan area.

1

Please rate the grammar of the generated descriptions:

- Excellent
- Good
- Average
- Poor
- Fair

2

How accurate were the factual details presented in the generated descriptions?

- Very accurate
- Somewhat accurate
- Neutral
- Not very accurate
- Inaccurate

3

The algorithm adds extra details if the learner ability is low. Do you think this is a good idea?

(As a guide, a learner ability value of
-2 to -0.5 : Low
-0.5 to 1 : Neutral
1 to 2 : High)

- Yes it is a good idea
- Does not really matter
- No it is not always a very good idea
- No it is a bad idea

4

To what extent the generated descriptions could have been produced by a human? Please provide a rating ranging between 1 to 5, 1 being the lowest and 5 being the highest.

1	2	3	4	5

5

For this question, consider the **reference sentence, triple and generated sentence 8**:

Supposing you had no knowledge that the generated sentence was not written by a human, would you have categorised that sentence as a human-written one to describe the triple?

- Yes
- Maybe
- No

6

Please add any comments you might have pertaining to the answered questions here. Make note of the question number:

Enter your answer

You can print a copy of your answer after you submit

I.2 Sentence-specific

Generating Adaptive Learning Materials (GALMAT) for students

This survey will evaluate specific aspects of the algorithm's output, created as part of the research study

Description of this study

This is a research study titled, Generating Adaptive Learning Materials or GALMAT for short, which is being conducted as part of the Bachelor of Science (Honours) in Computer Science programme at the University of Cape Town (UCT). The research is being conducted by student, **Chiranjeev Keshav Nathoo**, under the supervision of Associate Professor, **Maria Keet** and PhD candidate, **Toky Hajatiana Raboanary**.

Motivation of this study

In the educational environment, learners may have a different level of knowledge and understanding when learning a given topic. Some learners know more about certain subtopics than others. Generally, these differences in learners' knowledge are not considered, and they receive the same learning materials that may only be suitable for some of them. This project investigates different techniques to propose adaptive learning materials on a specific topic for each learner, depending on their knowledge.

Aim of this study

As part of this research study, a proof of concept algorithm is being developed to investigate whether it is possible to address the aforementioned issue and hopefully improve the quality of education in general. The application will ask the learner questions on a certain field of study and based on the answers given, it will then try to gauge the gap in the learner's knowledge. A document is then produced that provides notes on those knowledge gaps.

Structure of the survey

This survey will only be about the output of the algorithm, which is an informative document based on the concepts that the learner did not understand. You will be provided with reference sentences, ***triples** with its associated ***learner ability** (the algorithm's internal representation of the learner's knowledge gaps) and a generated description from the algorithm, corresponding to the triple.

For each reference sentence, triple and generated description, you will be asked to answer questions to evaluate;

the focus of the generated description: Is the content of the generated description true relative to the triple or not.
the information content of the generated description: Is the amount of information conveyed by the generated description too much or too little.

***Triple:** This is a technique to store knowledge in a condensed form (subject, predicate, object), that is only keeping the essential parts.
***Learner ability:** This is a number calculated by the algorithm and assigned to each concept to quantify the learner's knowledge on that particular concept. In the triple, the subject is the concept. The learner ability is a value in the range of -2 to 2 with -2 meaning that the learner has no understanding on the concept and 2 meaning that the learner fully understands the concept.

Participation terms and conditions

This survey should take approximately 5-10 minutes to complete. Please note that your participation is voluntary. No personal information will be requested from you and your responses will be stored in a secure, password-protected location on the Microsoft Cloud and only the researcher, Mr Chiranjeev Keshav Nathoo will have access to it. Submitting this form would imply that you agree to these terms and conditions.

Reference sentence	Vermicelli is an ingredient of Bakso.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	Bakso	hasIngredient	Vermicelli	0.3
Generated sentence	Bakso is an Indonesian-style meatball or meatball-like food that typically contains Vermicelli, a type of thin noodle, as an ingredient.			

1

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

2

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

3

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

4

How necessary do you think the added details, in the generated sentence, that explain what is Basko is to describe the triple?

- Very necessary
- Somewhat necessary
- Neutral
- Not very necessary
- Not necessary at all

Reference sentence	Celery is an ingredient of Bakso.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	Bakso	hasIngredient	Celery	0.3
Generated sentence	Bakso is a popular Indonesian meatball dish that contains celery as one of its ingredients.			

5

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

6

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

7

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

Reference sentence	Barny cakes is created by Mondelez International.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	BarnyCakes	isCreatedBy	MondelezInternational	-0.732
Generated sentence	BarnyCakes are a type of cake made and distributed by Mondelez International, a global confectionary, food, and beverage company.			

8

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

9

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

10

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

Reference sentence	Barny Cakes can be chocolate flavoured.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	BarnyCakes	hasDishVariation	Chocolate	-0.732
Generated sentence	BarnyCakes is a type of cake that has a variation made with chocolate.			

11

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

12

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

13

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

Reference sentence	Bacon Explosion is eaten in the United States.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	BaconExplosion	isEatenInCountry	UnitedStates	-0.268
Generated sentence	The BaconExplosion is a dish that consists of bacon woven into a lattice pattern with sausage and various seasonings, wrapped in bacon strips and smoked, typically over a wood-burning grill. It is typically eaten in the United States.			

14

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

15

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

16

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

17

How necessary do you think the added details, in the generated sentence, that explain what is Bacon Explosion is to describe the triple?

- Very necessary
- Somewhat necessary
- Neutral
- Not very necessary
- Not necessary at all

Reference sentence	Sausage is a main ingredient of Bacon Explosion.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	BaconExplosion	hasMainIngredients	Sausage	-0.268
Generated sentence	A BaconExplosion is a type of food dish which primarily consists of ingredients such as bacon, sausage, and other herbs and spices.			

18

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

19

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

20

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

Reference sentence	Bacon Explosion is eaten in the Kansas City metropolitan area.			
Triple (and associated learner ability)	**Subject**	**Predicate**	**Object**	**Learner Ability**
	BaconExplosion	isEatenInRegion	KansasCityMetropolitanArea	-0.268
Generated sentence	The BaconExplosion is a dish popularized in Kansas City that consists of bacon woven into two layers around a core of sausage and seasonings, and is typically smoked or grilled. It is often eaten in the Kansas City metropolitan area.			

21

Did the generated sentence focus on the triple?

- Yes, they were highly focused
- Yes, but they could have been more focused
- No, they were somewhat relevant
- No, they were not relevant at all

22

When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

- Yes
- Somewhat
- No

23

Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

1 2 3 4 5

24

How necessary do you think the added details, in the generated sentence, that explain what is Bacon Explosion is to describe the triple?

- Very necessary
- Somewhat necessary
- Neutral
- Not very necessary
- Not necessary at all

You can print a copy of your answer after you submit

J Result of surveys

J.1 General

1. Please rate the grammar of the generated descriptions:

More Details

Excellent	2
Good	2
Average	1
Poor	0
Fair	0

2. How accurate were the factual details presented in the generated descriptions?

More Details

Very accurate	2
Somewhat accurate	2
Neutral	1
Not very accurate	0
Inaccurate	0

3. The algorithm adds extra details if the learner ability is low. Do you think this is a good idea?

(As a guide, a learner ability value of
-2 to -0.5 : Low
-0.5 to 1 : Neutral
1 to 2 : High)

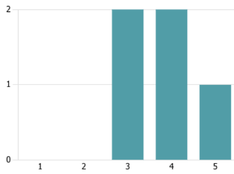
More Details

Yes it is a good idea	2
Does not really matter	1
No it is not always a very good i...	2
No it is a bad idea	0

4. To what extent the generated descriptions could have been produced by a human? Please provide a rating ranging between 1 to 5, 1 being the lowest and 5 being the highest.

More Details

3.80
Average Rating



5. For this question, consider the **reference sentence, triple and generated sentence 8**:

Supposing you had no knowledge that the generated sentence was not written by a human, would you have categorised that sentence as a human-written one to describe the triple?

More Details

Yes	2
Maybe	3
No	0

6. Please add any comments you might have pertaining to the answered questions here. Make note of the question number:

1 Responses

ID ↑	Name	Responses
1	anonymous	Question number 3: It is not always a good idea, because a slow learner may find it difficult to digest a lot of details, all at once, together with the basic idea.

J.2 Sentence-specific

1. Did the generated sentence focus on the triple?

More Details

Yes, they were highly focused	5
Yes, but they could have been m...	0
No, they were somewhat relevant	0
No, they were not relevant at all	0

2. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

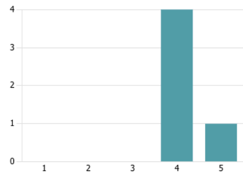
More Details

Yes	0
Somewhat	0
No	5

3. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

More Details

4.20
Average Rating



4. How necessary do you think the added details, in the generated sentence, that explain what is Basko is to describe the triple?

More Details

Very necessary	1
Somewhat necessary	3
Neutral	0
Not very necessary	1
Not necessary at all	0

5. Did the generated sentence focus on the triple?

More Details

Yes, they were highly focused	4
Yes, but they could have been m...	1
No, they were somewhat relevant	0
No, they were not relevant at all	0

6. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?

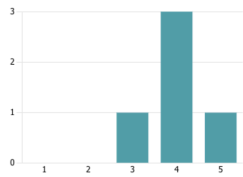
More Details

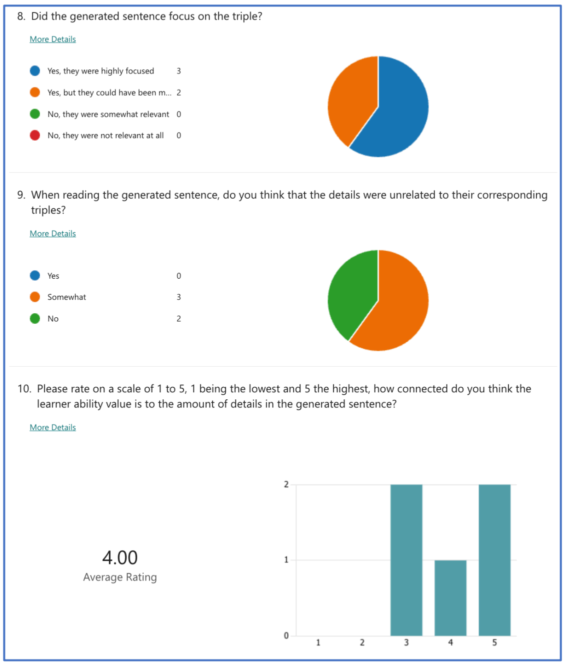
Yes	0
Somewhat	1
No	4

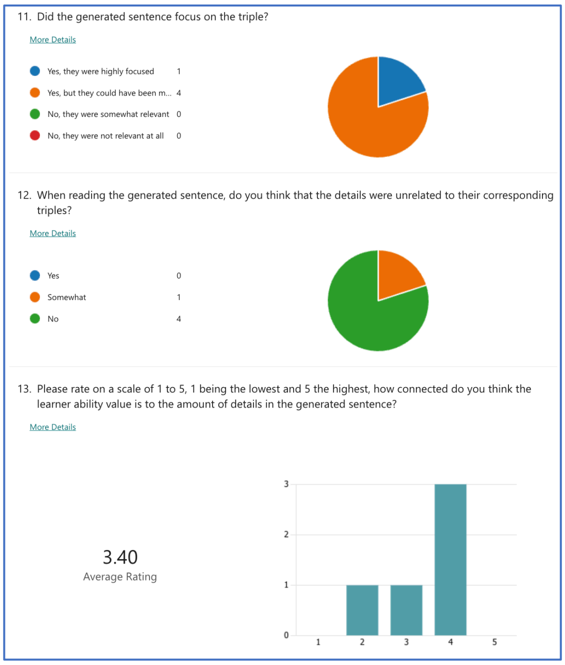
7. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?

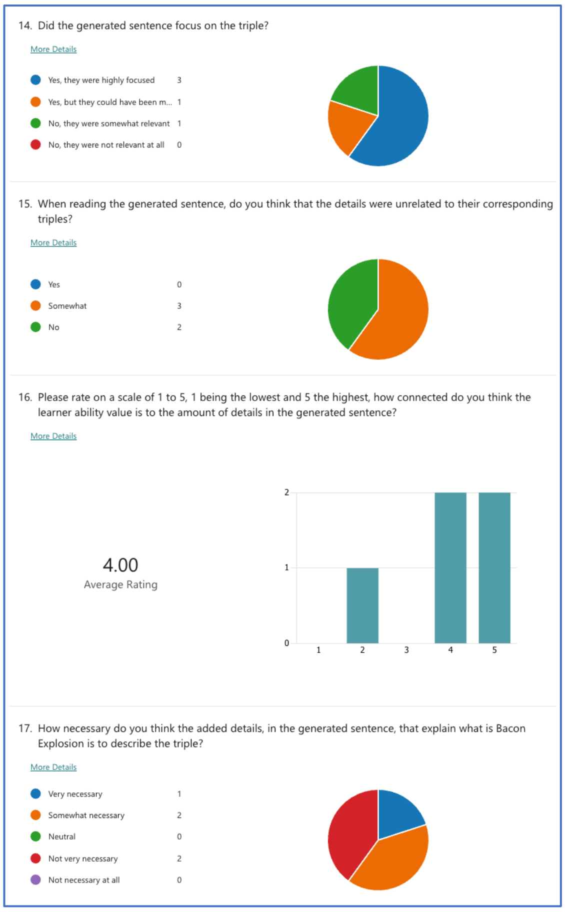
More Details

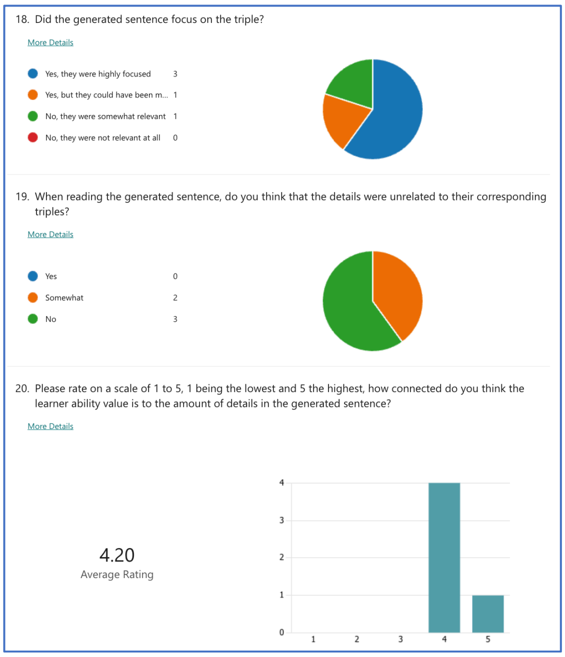
4.00
Average Rating

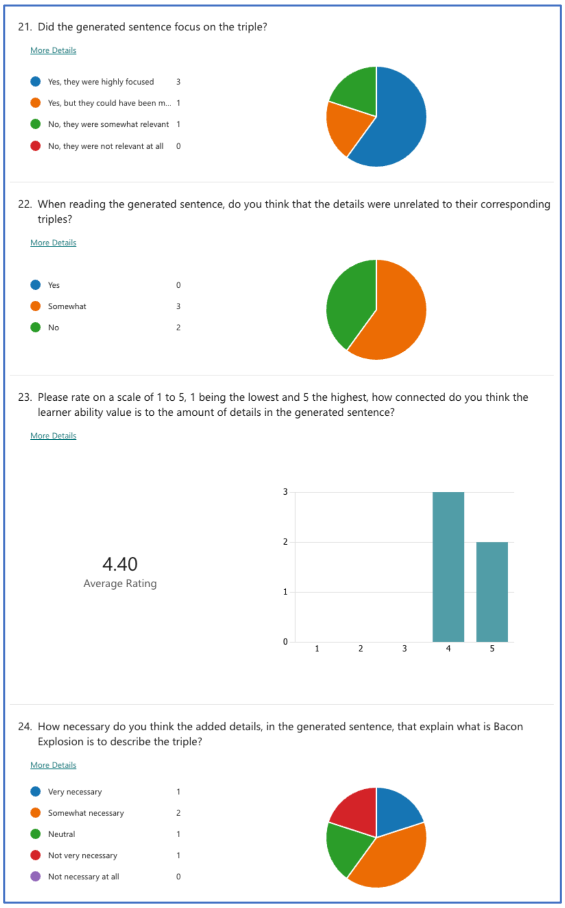



8. Did the generated sentence focus on the triple?
More Details
Yes, they were highly focused 3
Yes, but they could have been m... 2
No, they were somewhat relevant 0
No, they were not relevant at all 0
9. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?
More Details
Yes 0
Somewhat 3
No 2
10. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?
More Details
4.00
Average Rating


11. Did the generated sentence focus on the triple?
More Details
Yes, they were highly focused 1
Yes, but they could have been m... 4
No, they were somewhat relevant 0
No, they were not relevant at all 0
12. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?
More Details
Yes 0
Somewhat 1
No 4
13. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?
More Details
3.40
Average Rating


14. Did the generated sentence focus on the triple?
More Details
Yes, they were highly focused 3
Yes, but they could have been m... 1
No, they were somewhat relevant 1
No, they were not relevant at all 0
15. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?
More Details
Yes 0
Somewhat 3
No 2
16. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?
More Details
4.00
Average Rating
17. How necessary do you think the added details, in the generated sentence, that explain what is Bacon Explosion is to describe the triple?
More Details
Very necessary 1
Somewhat necessary 2
Neutral 0
Not very necessary 2
Not necessary at all 0


18. Did the generated sentence focus on the triple?
More Details
Yes, they were highly focused 3
Yes, but they could have been m... 1
No, they were somewhat relevant 1
No, they were not relevant at all 0
19. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?
More Details
Yes 0
Somewhat 2
No 3
20. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?
More Details
4.20
Average Rating


21. Did the generated sentence focus on the triple?
More Details
Yes, they were highly focused 3
Yes, but they could have been m... 1
No, they were somewhat relevant 1
No, they were not relevant at all 0
22. When reading the generated sentence, do you think that the details were unrelated to their corresponding triples?
More Details
Yes 0
Somewhat 3
No 2
23. Please rate on a scale of 1 to 5, 1 being the lowest and 5 the highest, how connected do you think the learner ability value is to the amount of details in the generated sentence?
More Details
4.40
Average Rating
3
2
1
0
1
2
3
4
5
24. How necessary do you think the added details, in the generated sentence, that explain what is Bacon Explosion is to describe the triple?
More Details
Very necessary 1
Somewhat necessary 2
Neutral 1
Not very necessary 1
Not necessary at all 0