Project Literature Review, Final Report
JaMiMaKa, topic: Analysis of Questions, Winter 2023

Kacper Grzymkowski	Jakub Fołtyn	Marceli Korbin	Mikołaj Malec	Anna Wróblewska
MSc student	MSc student	MSc student	MSc student	supervisor
WUT	WUT	WUT	WUT	lecturer at WUT
kacper.grzymkowski .stud@pw.edu.pl	01151388 @pw.edu.pl	01142124 @pw.edu.pl	01142129 @pw.edu.pl	anna.wroblewska1 @pw.edu.pl

Abstract

This document represents the final report for the project on the topic of *Analysis of Questions*. We will briefly discuss the project's topic and the related works introduced previously, but our focus in this document will be on presenting the results of our project. We will especially focus on describing the path that led to said results, as we believe that it best represents the exploratory nature of this project.

1 Introduction

There is no denying that one of the most fundamental parts of inter-human communication is asking questions. People inquire about lots of different things. When someone arrives late at a bus stop, usually the first thing that comes out of their mouth (apart from heavy breathing) is the question "has the bus left yet?". One may look for their lost pair of socks, and then questions such as "where are they?" come naturally. But these types of questions may be considered " simple ". They usually concern basic subjects, and serve as a tool to receive some certain knowledge. Questions, however, are not limited to obtaining just the most basic information. On the contrary, questions may be used to gain some more complex insight, while some questions may be intended not to be answered at all.

Questions, in general, are a complicated subject. That is why they have been a topic of interest for scientists for a long period of time now (Bromberger, 1966), (Schaeffer and Presser, 2003), (Dayal, 2016). Some scientists research how to properly answer certain types of questions (such as "why" questions) (Bromberger, 1966), while others tackle the very essence of question-asking and curiosity as a whole (Schaeffer and Presser, 2003). This proves that questions are indeed an important part of human cognition and that the notion of researching them in greater detail is valid. Having acknowledged this conclusion, in our project we would like to focus on one more potential aspect of questions. As already mentioned, more complicated questions may be an indicator of a process of greater understanding and cognitive complexity. Some questions may tackle subjects of science, mathematics, or even philosophy and induce one to perform a greater intellectual effort. We believe that some questions may be connected to the notion of *creativity* (Kaufman, 2016), and by researching them, one may be able to understand the whole *creative process*.

To properly conduct the research by combining questions with the idea of creativity, we need to focus on the questions' complexity. The more complex the question asked is, the more complicated and creative the idea needs to be. Sometimes, even the question itself may indicate that the person asking it has already conducted some creative process, e.g. by theorizing what other purposes may a certain object possess. However, as the formulation of the project's topic is quite broad, we needed to find a more concrete way of researching questions rather than checking their complexity in general. That is why we have decided to focus on clustering questions together into groups based on the questions' topics and then check if questions inside these topics possess some inherent complexity levels (in other words, if there are topics that agglomerate more complex questions). We believe that our project remains firmly in the exploratory domain, and as such, our goals are not as clearly defined as may be the case with other project groups. Nevertheless, we think that it has enormous scientific potential and, as such, do not consider the lack of a formal definition as a negative.

In the latter parts of this document, we will briefly discuss some related works regarding our project's topics. Then, we will present our methods and re-

sults, especially focusing on the exploratory path that led to said results. Finally, we will provide some discussion about the results as well as conclusions.

2 Related works

2.1 Other approaches and datasets

In this subsection we will describe some of the other works related to the task of question analysis. We will also provide a detailed description of the dataset that we have used in our project.

2.1.1 Bloom's taxonomy

In 1956, Benjamin Bloom and collaborators Max Englehart, Edward Furst, Walter Hill, and David Krathwohl introduced the Taxonomy of Educational Objectives, commonly known as Bloom's Taxonomy (Bloom et al., 1956). This framework, widely used by K-12 teachers and college instructors, categorizes educational goals into six major groups: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. These categories, presented as "skills and abilities" beyond Knowledge, emphasize that knowledge serves as a prerequisite for applying skills.

A revision in 2002 (Krathwohl, 2002) by cognitive psychologists, curriculum theorists, instructional researchers, and assessment specialists resulted in A Taxonomy for Teaching, Learning, and Assessment. This revised taxonomy shifts focus from static "educational objectives" to a more dynamic classification, using verbs and gerunds to describe cognitive processes. The six cognitive processes are Remember, Understand, Apply, Analyze, Evaluate, and Create.

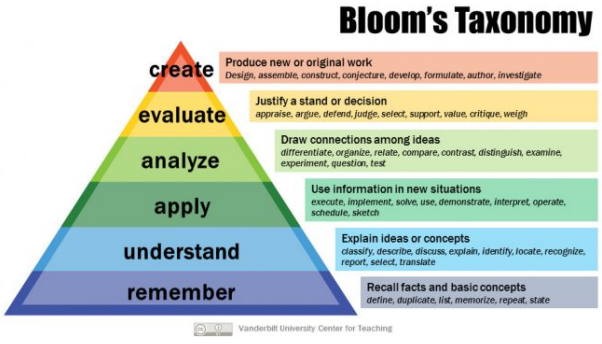

Figure 1: Blooms Taxonomy

In this revised taxonomy, knowledge underlies cognitive processes, and a separate taxonomy classifies types of knowledge used in cognition:

- Remember: ability to retrieve information from memory, demonstrating one's capacity to recognize and recollect previously learned facts or details
- Understand: Involves the cognitive processes of interpreting, exemplifying, classifying, summarizing, inferring, comparing, and explaining, showcasing the capacity to comprehend and articulate information in a meaningful way.
- Apply: Encompasses the actions of executing and implementing, demonstrating the practical application of acquired knowledge and skills in real-world scenarios.
- Analyze: Involves the cognitive actions of differentiating, organizing, and attributing, showcasing the ability to break down information, structure it, and assign relevant attributes for a deeper understanding.
- Evaluate: Encompasses the critical activities of checking and critiquing, while also involving creative processes such as generating, planning, and producing. This skill set highlights the ability to assess, appraise, and generate novel solutions or creations.
- Create: Involves the generative processes of planning and producing, showcasing a combination of creative activities such as generating novel ideas and implementing structured plans. This skill set underscores the ability to innovate, design, and bring forth new solutions or creations, emphasizing the synthesis of knowledge and skills in a constructive manner.

Using revised taxonomy allows establishing objectives in pedagogical interactions serves to clarify the purpose for both teachers and students. This clarity is further enhanced through the organization of objectives, which facilitates a structured approach. An organized set of objectives plays a crucial role in supporting teachers by aiding in the planning and delivery of appropriate instruction, the design of valid assessments, and ensuring alignment with overall teaching goals.

Utilizing Bloom's Taxonomy in the classification of questions within our paper provides a structured framework for assessing the cognitive complexity of each query. By categorizing ques-

tions according to the taxonomy's cognitive processes—ranging from simple recall to more advanced skills like analysis, synthesis, and evaluation—we can better understand the depth and complexity of information processing required for each question. This classification aids in tailoring our approach to different types of questions, ensuring a more nuanced and effective response strategy. Additionally, aligning our questions with the taxonomy assists in establishing clear learning objectives, facilitating organized teaching, and guiding the design of assessments that accurately reflect the intended educational outcomes.

2.1.2 Text complexity metrics

One of the most important parts of our project is to measure the complexity of questions – with the idea being that the more complex the question is, the more complex the answer it requires, and thus it may be considered to be more creative. There are several techniques to measure the complexity of a given text. We would like to introduce two categories of such metrics: the first one is *lexical diversity* (Malvern et al., 2004), which measures how diverse the text is in terms of words used as well as the quality of words used – meaning that texts with more "sophisticated" words usually score "higher". The second one is *text readability* (Zamanian and Heydari, 2012), which measures how accessible the text is in terms of how hard it is to understand – especially for non-native English speakers.

Some of the measures of text complexity may be quite simple – following this principle, we have started our investigation by utilizing some simple and easy-to-understand and compute metrics. Those included measuring the longest word in a question (with the idea being that questions including longer words may be more complicated) and measuring how many of the words that form the question are among the top 5,000 most used words in the English language.

Other lexical diversity metrics include *Type-Token Ratio* (*TTR*) (Templin, 1957), *Carroll's Corrected Type-Token Ratio* (*CTTR*) (Carroll, 1964) and *Simpson's D metric* (Simpson, 1949). The first one, TTR, is fairly simple:

$$TTR = \frac{V}{N}$$

Where V is the number of types in the text (types are just unique words), and N is the total number of words in a text. As one may expect, this metric measures the percentage of unique words in a text (with the idea being that the more unique words a text is comprised of, the more diverse and complex it is). Carroll then introduced a slight modification of TTR in 1964, and as such:

$$CTTR = \frac{V}{\sqrt{2N}}$$

Finally, Simpson's D measure is calculated as such:

$$D = \sum_{i=1}^{V} f_V(i, N) \frac{i}{N} \frac{i-1}{N-1}$$

where $f_V(i, N)$ is the number of types occurring i times in a text of length N. As the formula suggests, it is a measure of diversity that takes into account the prevalence of different types in the text. It is also widely used in biology to measure the biodiversity of a given habitat.

As is the case with lexical diversity, there are also multiple metrics used for measuring the readability of a given text. Some of the simpler ones include mean sentence length and mean syllables in a sentence – once again, with the idea that the longer the sentences (words) are, the more complicated the text is and the more difficult it will be for a reader to understand. Some more sophisticated metrics exist, however. One of them is *Flesch's metric* (Flesch, 1948). It is computed as such:

$$F = 206.835 - (1.015 \times ASL) - (84.6 \times \frac{n_{sy}}{n_w})$$

Where ASL is *Average Sentence Length*, defined as $\frac{\text{number of words}}{\text{number of sentences}}$, n_w is the number of words in a document and n_{sy} is the number of syllables in a document. It assumes values from 0 to 100, and the higher the score, the easier the text is to read. The readability easiness is expressed in the form of text-appropriate grades – meaning that, for example, a text with a score between 50 and 60 is suitable to be read by people at the level of 10th to 12th grade (in the American school system). The highest possible score corresponds to 5th grade, and the lowest – to professional-level texts.

Another metric of text readability is the *Automated Readability Index* (*ARI*) (Senter and Smith, 1967). It is computed as

$$ARI = 0.5ASL + 4.71AWL - 21.34$$

Where AWL stands for *Average word length*, defined as $AWL = \frac{\text{number of characters}}{\text{number of words}}$. This metric is usually rounded to integers and ranges from 1 to 14. It is in many ways analogous to Flesch's index in that it also expresses different text readability in terms of American grades – but here, the lower the metric, the easier the text is, starting from kindergarten level (ages 5-6), ranging all the way up to college student-level (ages 18 and up).

Finally, there is another text complexity metric that we would like to consider in our project. In fact, it is the most important metric and a metric that we would like to rely on in all our experiments, as this metric takes into account also the *creativity level* of a given text. This metric is the *Divergent semantic integration* (*DSI*) (Johnson et al., 2023).
DSI is yet another metric with the purpose of measuring text's diversity. This time, however, instead of utilizing properties such as word/sentence length, number of syllables etc., DSI aims to compare the meanings of words themselves. The idea is that the more disjoint and dissimilar the words in a sentence are, the more diverse and creative the text is. DSI is computed by comparing words inside a document among themselves – usually, it is done by computing cosine similarity between the embeddings of individual words – as this may give us a clear indication of the similarity of meanings between these words. The similarity is computed in a pairwise manner. Authors of (Johnson et al., 2023) used *BERT* (Devlin et al., 2018) to compute word embeddings in a text.

2.1.3 WTC-corpus

When thinking about the topic of questions in NLP, one may usually consider the task of question answering. While this is a perfectly valid task, in our project, we would like to go beyond it and focus on processing and acquiring information about the questions themselves. That is why works such as the article titled "What makes us curious? Analysis of a corpus of open domain questions" (Xu et al., 2021) was especially important in our research. There, the authors propose a dataset consisting of over 10,000 questions (8,000 after filtering) asked by various residents of Bristol, England. This dataset is also called the **WTC-corpus**. The goal of that article was to study the curiosity and capture the thinking processes of Bristolians. To achieve this, in addition to question answering, authors also considered the tasks of question topic classification and question equivalence/similarity detection. Authors' intention was to create singular model capable of performing all of these tasks. For this purpose, **BERT** (Devlin et al., 2018) model was utilized, or, more precisely – **S-BERT** (Reimers and Gurevych, 2019), a modification of BERT that captures sentence similarity and provides embedding for a given sentence. This model was then fine-tuned utilizing some additional datasets specialized in each of the abovementioned tasks. The resulting (after fine-tuning) model was called **QBERT**.

For question topic classification and question answering (in the form of choosing an appropriate answer from a set of given answers) authors used simple classification with the Softmax function applied to an element-wise difference of embeddings values. For the similarity detection task and question answering in the variant of retrieving possible answers from Wikipedia articles summaries, authors used cosine similarity. In general, the models created yielded promising results, although their performance depended heavily on the configuration of datasets used for fine-tuning the QBERT model. These results are promising and may give us some preliminary proposals on how to tackle the tasks of question clustering and topic classification. Another interesting and potentially useful action proposed by the authors was dividing questions into two categories: factual and counterfactual questions. The first group consists of so-called "WH-questions", meaning questions involving words "when", "who", "where", "why" etc. while the second group usually includes questions with the word "if". Questions from the counterfactual group, according to authors, usually require more complex answers and may give a better insight into a person's reasoning.

2.1.4 Topic modeling visualization

Topic modeling, while not being directly connected to questions, may still prove to be most beneficial for our project. It is an unsupervised technique used to identify natural topics in text (Blei and Lafferty, 2009). It may be especially helpful in the task of question clustering, as intuitively questions regarding similar topics should be clustered together.

One of the most popular methods of topic modeling is **latent Dirichlet allocation** (**LDA**)(Blei et al., 2003). It has been used in another arti-

cle that we would like to mention in this section: "Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach" (Buenano-Fernandez et al., 2020).

In this paper, as the title suggests, authors apply the LDA topic modeling method to a dataset of open-ended questions (and their answers). That dataset was created from the online surveys for self-assessment of teachers in an Ecuadorian university. The approach for this task may be considered quite straightforward: authors simply create a term-document matrix from a previously preprocessed database, and then apply LDA technique to retrieve various topics from documents and assign specific terms to them. In this way, it is possible to retrieve the frequency of specific topics for specific documents. An expert then assesses the relevance of found topics.

While the presented pipeline may be basic, the main strength of this article is the various visualizations, depicting the possible clusterings based on retrieved and assigned topics. They may serve as a possible inspiration for the latter stages of our project.

2.1.5 Stanford Question Answering Dataset

Stanford Question Answering Dataset (Rajpurkar et al., 2016), also known as **SQuAD** in short, is a reading comprehension dataset prepared by the Stanford University. It exists in two versions: 1.1 and 2.0; in the work, we are focusing on the former version. It contains over 100,000 questions, posed by crowdworkers on more than 500 articles from Wikipedia; each of them is provided at least one answer, based on the corresponding reading passage (if multiple answers to a question occur, they can be the same). The original dataset is distributed in a form of two JSON files, serving as the train and test split. Several question-answer pairs examples include:

- Q: When were the Normans in Normandy?
  A: 10th and 11th centuries
- Q: How long is the section of the Rhine near Chur?
  A: 86 km long
- Q: When many people are arrested, what is a common tactic negotiating?
  A: solidarity
- Q: What is an object's mass proportional to at the surface of the Earth?
  A: force of gravity

2.2 Tools and algorithms

In this subsection we will explore in more detail various tools and algorithms used during our project.

2.2.1 LDA

Having already mentioned LDA, we would like to explore it in more detail now. **Latent Dirichlet Allocation** (LDA) (Blei et al., 2003) is a powerful and widely used statistical model for understanding and analyzing collections of discrete data, particularly in the context of text corpora. It is considered a generative probabilistic model, meaning that it helps explain how the data might have been generated. At its core, LDA is a three-level hierarchical Bayesian model that seeks to uncover the latent structure hidden within a collection of documents. This structure revolves around the concepts of topics and the distribution of words within those topics. In LDA, a document is viewed as **a mixture of topics**. It assumes that a document is not about just one topic, but rather a combination of several topics. These topics are latent, meaning they are not directly observed but are inferred from the text. Each topic, in turn, is viewed as a distribution over words. This means that a topic is characterized by a set of words that are more likely to occur when discussing that particular topic. These word distributions are also hidden variables that LDA aims to uncover. The model assumes that documents are generated in a probabilistic manner. To be more precise, for each document, LDA assigns a distribution of **topic probabilities**. These probabilities indicate the likelihood of a particular document containing certain topics. For example, a news article about technology might have a high probability of containing topics related to "technology" and "innovation", but it may also have a smaller probability for other topics like "politics" or "health".

Finally, the actual words in the document are generated based on the topics. For each word in the document, LDA selects a topic from the distribution of topics specific to that document and then selects a word from the distribution of words for that topic. In essence, LDA seeks to reverse-engineer the topic structure that might have generated a given collection of documents. It does

this through **statistical inference**, trying to find the most likely set of topics and their associated word distributions that would have created the observed documents. The LDA model can be trained on a large text corpus to discover these latent topics and word distributions, providing valuable insights into the content and themes present in the data.

As already mentioned, LDA assumes that each document is comprised of several topics, and these topics are further comprised of sets of words. In fact, the topics in LDA may be treated as simple distributions of word frequencies. For that reason, found topics usually do not have a "proper" interpretation (they can be interpreted based on the most prevalent words in a topic, so words that have the highest probability of belonging to a topic). The LDA, at its' core, seeks to find those word distributions and thus create the documents' topics. It is important to note, however, that after finding these topics, it is possible to retrieve then the topics' assignment of each document (each document has a set distribution of topics it tackles). By choosing the topic with the highest probability, we are thus able to create a clustering using the LDA method.

LDA's generative nature allows it to be a versatile tool for various natural language processing tasks, such as document clustering, topic modeling, text summarization, and even recommendation systems. By understanding the underlying topics within a collection of documents, LDA enables researchers, data scientists, and analysts to uncover patterns, extract meaningful information, and gain a deeper understanding of the content within textual data.

On a final note, we would like to mention another version of LDA briefly. It was introduced by Jagarlamudi et al. in (Udupa, 2012). It is so-called seed-guided LDA. It is essentially an extension of a "regular" LDA algorithm that allows users to use some pre-defined words as topics. These words are often referred to as "seeds". The algorithm is then forced to include these words in topic distributions, thus enabling users to "guide" the LDA into a chosen path (by forcing it to include seeds in the topics word distribution calculations).

2.2.2 Large Language Models (LLM)

The **transformer architecture** (Vaswani et al., 2017) has driven notable advances in natural language processing. Tasks like text classification, machine translation, dialogue systems, and information retrieval achieve impressive results but come with significant computational costs (Singh and Mahmood, 2021). Training large models requires extensive data, resources, and expertise, making them inaccessible for many researchers. Efforts to enhance efficiency mostly focus on inference, less relevant to our goal of creating a new system.

Recent language model advancements prioritize larger models, evident in **GPT** successors: **GPT-2** (1.5 billion parameters) (Radford et al., 2019), **GPT-3** (175 billion parameters) (Brown et al., 2020), and rumored **GPT-4** (1.76 trillion parameters) (Schreiner, 2023). Large model size complicates even "light-weight" fine-tuning. Efforts like Low Rank Adaptation (**LoRA**) (Hu et al., 2021) aim to help, but fine-tuning remains data-intensive.

Addressing challenges and increasing model efficacy, "prompt engineering" (Brown et al., 2020) enables information priming without model weight updates. It includes zero-shot, one-shot, and few-shot learning for rapid, cost-effective specialized model creation with minimal expertise. This opens a way to very quickly and very cheaply create specialized models, and can be theoretically done with no expertise in natural language processing, deep learning or programming.

Considering our project, questions like "Is a question What is the meaning of life? difficult from 1 to 10?" can be used to evaluate the difficulties of the question. The outputs of such a query should be viewed as just an opinion of the model.

However, "prompt engineering" has limitations, notably the risk of "hallucinations" in which the model fabricates information. This issue is less pronounced in "common-sense" questions, often ambiguously defined.

2.2.3 BERT

BERT, or **Bidirectional Encoder Representations from Transformers** (Devlin et al., 2018), is a breakthrough in natural language processing and understanding that has revolutionized the way machines comprehend and generate human language. BERT is not a generative model like LDA but rather a pre-trained transformer-based model that excels at various language understanding tasks.

Unlike earlier models that processed text in a unidirectional manner, BERT introduced **bidirectional context**. It understands words in relation

to their entire context within a sentence. In other words, BERT considers both the words that come before and after a given word, allowing it to capture the full meaning and nuances of the language. This bidirectionality is crucial in understanding the context, making it highly effective in tasks like sentiment analysis, question answering, and language translation.

BERT is built upon the **transformer architecture**, which has proven to be highly effective for a wide range of natural language processing tasks. The transformer architecture is designed to handle sequential data, like text, and is based on the concept of attention mechanisms. BERT leverages this architecture to process input data through a stack of self-attention layers, which enables it to model the relationships between words in a sentence efficiently.

BERT's power comes from **pre-training** on massive amounts of text data. During pre-training, it learns to predict missing words in a sentence and understand the context in which each word appears. This pre-trained model is then **fine-tuned** for specific downstream tasks. Fine-tuning involves training the model on smaller, task-specific datasets, which makes it adaptable to various applications. This fine-tuning process enables BERT to excel in tasks such as text classification, named entity recognition, and machine translation.

BERT has been trained in multiple languages, making it a valuable resource for multilingual applications. It can understand and generate text in a wide range of languages, which is crucial for global businesses and organizations that need to process text in different languages.

BERT's deep bidirectional learning enables it to capture semantic relationships between words and phrases. This means it can understand not only the surface meaning of words but also their contextual significance. It's capable of recognizing synonyms, antonyms, and even nuances in sentiment, which is especially important in tasks like sentiment analysis and language generation.

BERT introduced the concept of transfer learning to NLP, which allows models to leverage knowledge learned from one task to perform better on other, related tasks. This has greatly reduced the amount of labeled training data needed for many NLP applications and accelerated progress in the field.

BERT has become a cornerstone of modern natural language processing, while its availability as a pre-trained model has lowered the barriers to entry for developing NLP applications. Researchers, developers, and organizations can take advantage of BERT's pre-trained representations and fine-tuning capabilities to quickly build and deploy powerful language understanding systems.

In summary, BERT represents a major advancement in natural language processing by bringing bidirectional context, transfer learning, and transformer architecture together. Its versatility and ability to understand the intricate details of language have made it an indispensable tool for a wide range of NLP tasks, from sentiment analysis to machine translation and beyond.

2.2.4 Sentence Embedding

Sentence Embedding is a crucial technique in natural language processing that plays a significant role in various text analysis tasks, including text classification, semantic similarity, and information retrieval. It aims to transform a sentence or a piece of text into a fixed-length vector representation, preserving its semantic meaning and context. Several approaches have been developed to create sentence embeddings; here, we will explore three prominent methods: Averaging Word Embeddings, Pre-trained Models like BERT (Devlin et al., 2018), and Neural Network-Based Approaches.

Averaging word embeddings is a simple yet effective technique to obtain sentence embeddings. In this method, each word in a sentence is represented as a word embedding vector, typically obtained from pre-trained word embeddings like Word2Vec or GloVe. The sentence embedding is then calculated by averaging the word embeddings of all words in the sentence. While straightforward, this method often captures the overall meaning of the sentence, making it useful for tasks where context may not be as critical. However, it may lose nuances and complex sentence structures.

Pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) have revolutionized the field of sentence embedding. BERT, as a contextualized language model, learns to represent words based on their surrounding context. To obtain sentence embeddings using BERT, one can simply feed the entire sentence to the model, and it returns contextualized embeddings for each word. These embeddings are then pooled

or aggregated to create a single sentence embedding. BERT embeddings are highly contextual and capture the meaning of the sentence with intricate details, making them suitable for a wide range of tasks, from sentiment analysis to question answering.

Neural network-based approaches involve training specific models to generate sentence embeddings. These models can range from simple feedforward neural networks to more complex architectures like Siamese networks. Siamese networks, for example, are trained to compare sentence pairs and produce embeddings that reflect the similarity or dissimilarity between sentences. Neural network-based approaches offer flexibility in designing models tailored to specific tasks and datasets, making them suitable for applications where fine-tuned control is needed over the embedding process.

Each of these approaches has its advantages and is applicable in different scenarios:

- Averaging Word Embeddings method is quick and easy, making it a practical choice for applications where computational resources are limited. It's particularly useful for tasks that require simplicity and speed.
- Pre-trained Models like BERT: BERT-based embeddings offer state-of-the-art performance in many NLP tasks. They excel at capturing context, nuances, and semantic meaning. These embeddings are highly recommended for applications where understanding the full context of a sentence is crucial.
- Neural Network-Based Approaches: These approaches provide a middle ground, offering a balance between simplicity and context preservation. They can be fine-tuned for specific tasks and are suitable when a certain level of customization is required.

In summary, sentence embedding is a fundamental technique in NLP, and the choice of method depends on the specific requirements of the task. Whether through simple averaging, leveraging pre-trained models like BERT, or utilizing neural network-based approaches, the goal remains the same: to convert text into meaningful numerical representations that can be used for a wide range of natural language processing applications.

3 Methods

Our analysis is more focused on exploration and trying to make sense of a difficult and poorly defined task. As such, we have tried approaching the project topics from different angles, some of them have ultimately been considered unsatisfactory.

3.1 LLM

Our first approach to tackling this project was to use an LLM, specifically the *mistrail-7b-AWQ* transformer to assign questions to specific levels of the *Bloom's taxonomy*. This has been done by utilizing simple *Prompt engineering*. We have also used *all-minilm-l6* sentence transformer to create sentence embeddings that were later used for question clustering purposes. The prompts crafted for extracting question labels were as follows:

```
    Bloom's Taxonomy is a hierarchical
    ordering of cognitive
    skills that can help
    teachers and students
    in the classroom.
    There are multiple levels
    of understanding:
1 - Knowledge
2 - Comprehension
3 - Application
4 - Analysis
5 - Synthesis
6 - Evaluation.
For example, the question '{}'
is considered to be of level
```

And the second, more verbose prompt:

```
    Bloom's Taxonomy is a
    hierarchical ordering of
    cognitive skills that
    can help teachers and
    students in the classroom.
    There are multiple
    levels of understanding:
1 - Knowledge, which relates to
identification and recall of
information,
2 - Comprehension, which relates
to organization and selection of
facts and ideas,
3 - Application, which relates
to the use of facts, rules and
principles,
4 - Analysis, which relates
```

```
to separation of a whole into
component parts,
5 - Synthesis, which relates
to the combination fo ideas
to form a new whole,
6 - Evaluation. which relates
to developing opinions, judgements
or decisions.

For example, the question '{}' is
considered to be of level
```

In general, the model was asked to assign each question a label through the use of the above-presented prompts. Labels ranged from 1 to 6, corresponding to the different levels of the Bloom's taxonomy. In general, the lower the label, the less creative the question may be considered. There were also two additional labels: -1 and -2, which corresponded to situations in which the model either did not analyze the question or failed to generate a label for a given question.

3.2 Clustering

We performed question clustering on the embeddings provided by the *all-minilm-l6* sentence transformer. We verified multiple clusterings by changing the parameter controlling the number of clusters. Then we choose the best parameter according to the silhouette score (Rousseeuw, 1987). That parameter was used for all the subsequent clusterings. We also sampled questions from different clusters, trying our best to match the questions to the Bloom's taxonomy and to isolate interesting clusters. We then created a single run of t-SNE (Van der Maaten and Hinton, 2008) for visualization. Found clusters, LDA, DSI and question words were added to the visualization by coloring the observation points. Afterwards, we combined the embeddings, DSI scores and the detected question words into a single data frame. We then performed another clustering, applied t-SNE visualization and sampled questions from interesting structures, in order to assess their validity.

3.3 LDA

For accessing *Latent Dirichlet Allocation*, we have used the *Gensim* package (Řehůřek and Sojka, 2010) for the Python language. From the very same package, we have also taken some simple preprocessing methods for adjusting our dataset for the use of LDA. Those methods included the *simple_preprocess* function and the *word corpus* creating function.

We have converted our dataset into a *bag-of-words* using the *gensim* (Řehůřek and Sojka, 2010) package. Following this, we have invoked the *LDAMulticore* function from the same package on the "raw" version of our dataset (so a dataset with no words removed). It is worth noting that the *LDAMulticore* function is the parallelized version of LDA computation, which allows us to speed up the computation process. We did not change the default parameters of the model, setting only the *random_state* for reproducibility.
We have set the number of topics to be found to 5. It was an arbitrary choice, as the dataset did not include any clearly defined topics.

We have also tried used the *seed-guided LDA*. It enabled us to use our own pre-defined words as topics to be found by the LDA algorithm. AS there were no obvious topics in the original dataset, we have decided to use the levels of the *Bloom's taxonomy* (Krathwohl, 2002) as our seeds. For *Seed-guided LDA*, we have used the *lda* Python package.

3.4 Text complexity

For text complexity measurement, we have used the *quanteda* package (Benoit et al., 2018) for the R programming language (R Core Team, 2021). This package included all the necessary functions for computing multiple text complexity measures, as well as some functions for text tokenization and creating a word corpus. Unfortunately, ultimately we did not utilize any of these methods for assessing our clustering results, as we decided that our dataset contains too short texts for such an analyzis.

We did, however, utilize the DSI ((Johnson et al., 2023)) metric to analyze the validity of our clustering. It has been provided by authors of (Johnson et al., 2023) in the form of a ready-to-use Python program.

3.5 Question words

Another approach to retrieving the type of question (in terms of complexity) has been the *question word analysis*. It has been proposed by the authors of (Xu et al., 2021). In general, the approach states that certain so-called "question words" belong to more complex (and, therefore, interesting) questions. Following this notion, we could potentially

cluster the questions based only on the question word they contain.
The selected words were: “how”, “what”, “when”, “where”,“which”, “who”, “why” and “if”. In case when none of these words were present, the question was labeled ”none”. The “if” questions were especially considered to be associated with greater question complexity, as argued by the authors of (Xu et al., 2021).

3.6 Dataset

In our work, we used the Stanford Question Answering Dataset as our training data. Due to computational load, for the training we use a subset of 1,000 questions from the dataset, basing on its preprocessed version prepared for and used in (Zhou et al., 2017).

4 Results

In this section we will present all the results of our project, emphasizing the path that brought us to them. The various subsections of this section will represent this path.

4.1 LLM

We generated labels for a subset of questions from the Standford dataset using prompts presented in the *methods* section (3).
We applied the LLM to our dataset. Unfortunately, the process soon proved to be very computationally demanding and slow. What is more, the vast majority of labels assigned corresponded to the lowest levels of the taxonomy (1). Any higher values were exceptions. Nevertheless, there have been some questions that were assigned a higher label. We tried to investigate them in detail. Example questions are presented in figure 2. As we can see, the questions are rather diverse, but there seems to be an underlying trend of questions regarding Beyoncé, a famous American singer. We consider these results to be rather unsatisfactory and as such refrained from continuing the tests with the LLM models.

4.2 LDA

LDA performed on the whole dataset has yielded unsatisfactory results – as all the found topics included only the most common stopwords (such as “the”, “what” etc.) To combat this, we have decided to remove all the stopwords from the dataset before retrying the LDA method. It is worth noting

246	how many schools have a similar men 's basketb...	1	4
254	to where are the loyal sons in `` notre dame f...	2	1
266	what areas did beyonce compete in when she was...	1	3
273	in what city did beyonce grow up ?\n	1	3
279	what was the first album beyoncé released as a...	1	3
282	what was beyoncé 's role in destiny 's child ?\n	1	2
285	after her second solo album , what other enter...	2	3
290	which album was darker in tone from her previo...	2	4
291	after what movie portraying etta james , did b...	2	1
297	in her music , what are some recurring element...	3	2
299	which magazine declared her the most dominant ...	1	2
302	how did beyonce describe herself as a feminist...	2	1

Figure 2: Example questions that have been assigned a label greater than 1. The first column is the question itself, the second and third column represent labels corresponding to two different prompt formulations. As we can see, there seems to be an abundance of questions regarding famous American singer Beyonce.

that removing the stopwords from questions shortened them significantly – as some questions now consisted of only a few words. Having removed the stopwords, newly found topics were noticeably more diverse – as shown in figure 3. The questions

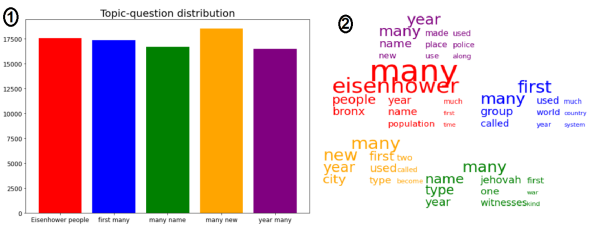

Figure 3: Results of LDA topic-based clustering. In image 1) we can see the size distribution among clusters – as we can see, clusters are fairly similar in size. In image 2) we have a visualization of the most popular words used in each cluster – as we can see, the word “many” is the most prevalent among all clusters.

were evenly distributed among clusters, and there appeared to be some original and distinguishable terms in some of the topics. Unfortunately, most of the clusters still were based on similar words, such as the prevalent “many” word. The inefficiency of this clustering method was later confirmed by visualizing it using the *TSNE* technique (Van der Maaten and Hinton, 2008) 4. Also a further investigation into the questions classified

into given clusters has proven that there are no noticeable "leading" topics in any of them. Having

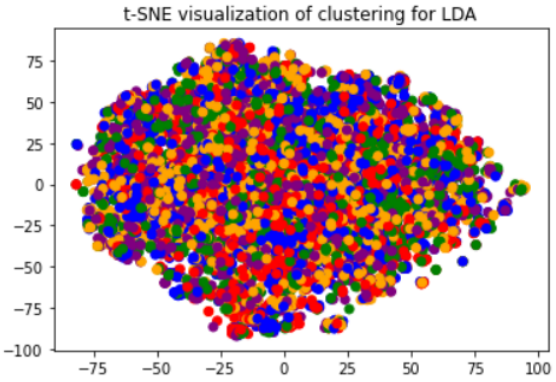

Figure 4: Clustering visualized on the dataset modified using the TSNE technique.

investigated the "normal" LDA model, we have tried fitting a *seed-guided LDA* model (Griffiths and Steyvers, 2004) to our data. The results of the new clustering are shown in figure 5. As we can see, the most numerous clusters are the clusters corresponding to the "remember" and "evaluate" levels. While the "remember" cluster is quite accurate, as most questions in our chosen dataset are factual questions (questions asking for a set piece of information), the sizes of different clusters are quite alarming. Unfortunately, the *DSI* metric measurements among clusters confirm that the clustering is indeed inaccurate – as there are no noticeable differences between the clusters (in a proper clustering, the metric should be decreasing with each level, as questions corresponding to different levels of *Bloom's taxonomy* should differ in their creativity).

4.3 Clustering

The t-SNE visualizations of both clusterings are present in 6 and 7. In the figure 6, several clusters can be spotted, but they often overlap and are overall not very concrete. The combined clustering, illustrated in 7, was more accurate, resulting in clusters easier to identify and distinguish, even though in every group visible from the t-SNE visualization alone there are at least two clusters.

4.4 DSI

The DSI metric was calculated for a sample of observations and visualized on the t-SNE, as can be seen on figure 8 With the first clustering, there does not seem to be a clear pattern of how the DSI

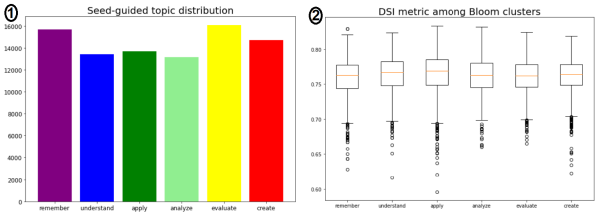

Figure 5: Results of the guided LDA question clustering. In image 1) we can see the distribution of a number of questions among chosen clusters – as we can see, the questions corresponding to the "remember" and "evaluate" clusters are the most numerous. In image 2) we can see the DSI metric distribution inside the clusters. Unfortunately, there are no noticeable differences among different clusters.

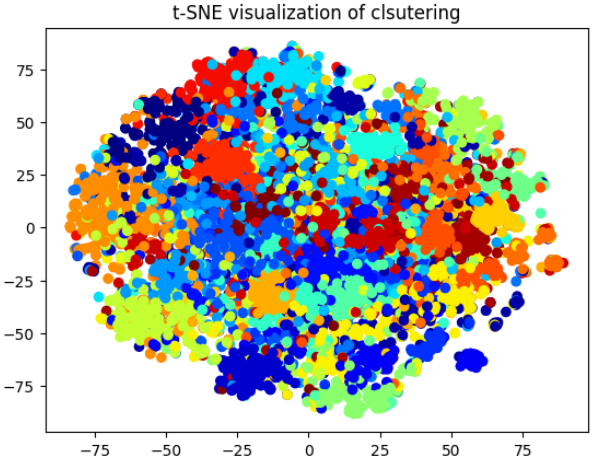

Figure 6: First clustering visualization.

metric is distributed. There are no clear hot-spots, or cold-spots. We then investigated the difference in DSI across clusters found in the clustering step, which can be seen on figure 9. Most observations fell in the 0.7 - 0.8 range of the metric, with some outliers reaching as high as 0.84 and as low as 0.6. However, there is no strong difference in DSI between clusters. In the second clustering, seen on figure 10, the DSI metric seemed to organize itself in a gradient-like arrangement, with one end of a cluster having a much higher score than the other end. This is quite expected, as the second clustering included the DSI metric as a predictor, but is still interesting. Comparing with the clusters seen in figure 7, this "gradient" behavior might be useful to separate the clusters depending on their complexity.

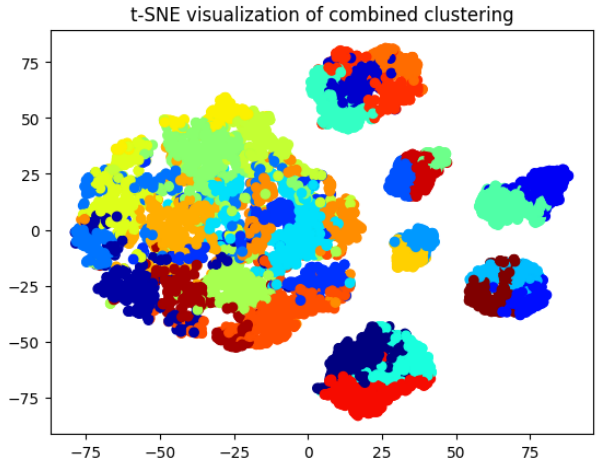

Figure 7: Second clustering visualization.

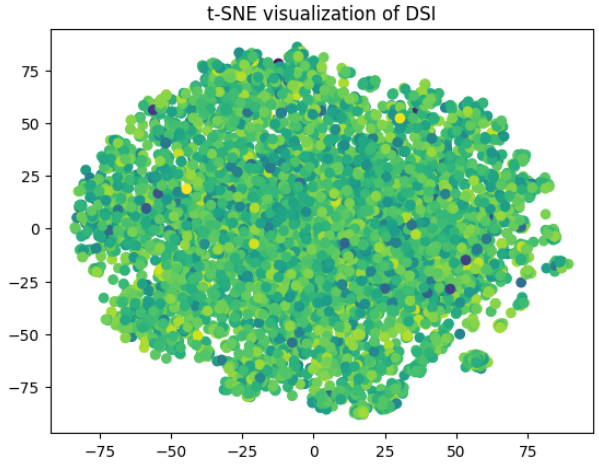

Figure 8: First clustering DSI visualization.

4.5 Question words

In the first clustering, there does not seem to be a pattern to the distribution of question words, as can be seen on figure 11. The main thing to notice is the large amount of "what" words, which can also be seen on the histogram in figure 12. However, an interesting pattern appears in the t-SNE visualization when visualizing the question words in the second clustering. It can be seen on figure 13. While most questions words formed their own "blobs", there is a mixed cluster in the middle. In terms of k-means clustering, this correlates to two clusters.

Sentences sampled from these interesting clusters were manually analysed. Some of the questions showed promise as they seemed more inquisitive and require a large understanding of the subject to answer. For example, "*why was there a large population of algonquian people in bermuda*

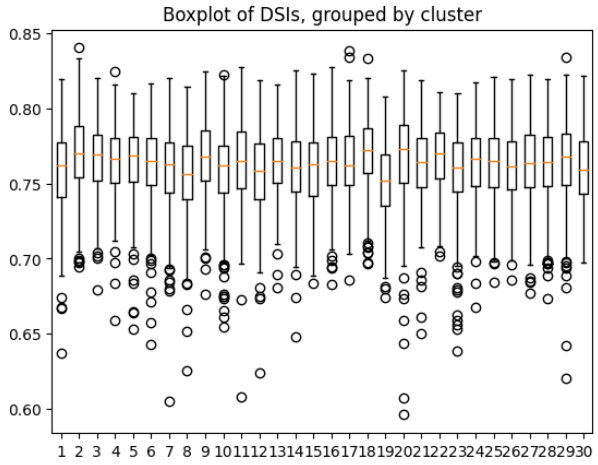

Figure 9: Box plot of DSI distribution across different clusters in the first clustering.

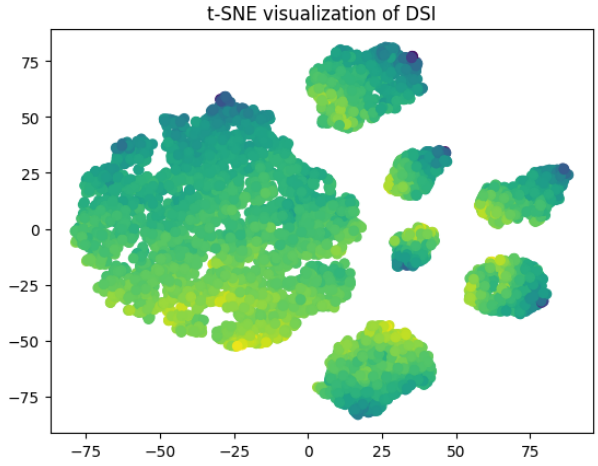

Figure 10: Second clustering DSI visualization.

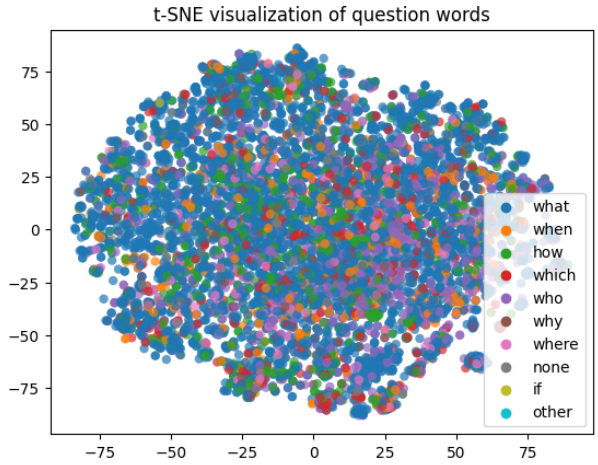

Figure 11: First clustering question words visualization

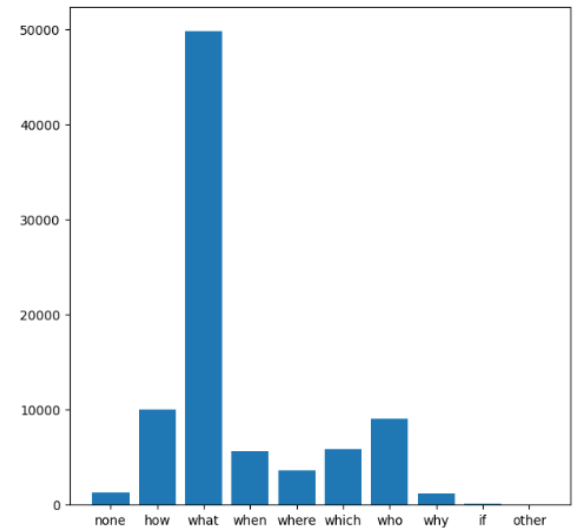

Figure 12: Distribution of question words in the dataset

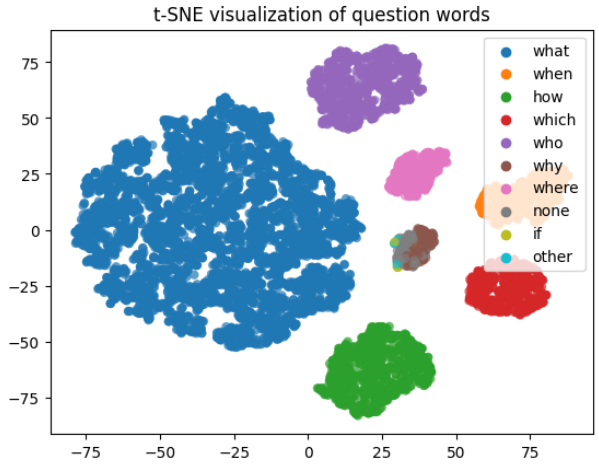

Figure 13: Second clustering question words visualization

?", "*why is it possible to distinguish utf-8 from other protocols ?*" and "*why are 99 % of pesticide related deaths in underdeveloped countries ?*" are examples of questions which we think that might be more complex than average in this dataset. However, also present in this cluster were very messy, no-context questions, such as: "*did they need parental consent ?*", "*may welsh clubs enter the competition ?*", "*are the ewell 's considered rich or poor ?*" and "*name a smaller newspaper ?*". It's hard to gauge the number of sentences belonging to each group in this cluster without going through each one individually, but we think the split is somewhat even.

5 Discussion

The topic in question is a really difficult one. With the exploratory nature of this task, it's hard even to declare the success or failure of the project. While our results are far from perfect, we see that there might be sparks of something interesting in our analysis. In particular, we believe that a simple extraction of question words might be a powerful tool in the study of question complexity. However, sentence embeddings provided a good starting point and allowed us to construct and explore the data, while LDA, DSI offered different perspectives on the problem

We believe a major limitation in our work is the dataset. While there are many prepared, cleaned and labelled datasets containing questions, many of them only contain simple queries. This is because these datasets were created with the intent of building better search engines and virtual assistants, like Google or ChatGPT. This is a big problem in our case however, as we try to classify the difficulty of the questions being asked, rather than try to answer questions.

6 Conclusion

We believe that it might be possible to build a more robust method of analyzing questions from the perspective of their difficulty. But it's likely going to require a much greater focus on the data acquisition and labelling.

7 Table of contributions

Present in table 7.

Name	Contribution
Kacper Grzymkowski	Transformer and LLM literature review
Kacper Grzymkowski	Conclusion section of literature review
Kacper Grzymkowski	Solution concept, architecture and proposal
Kacper Grzymkowski	Feasability rebuttal
Kacper Grzymkowski	LLM prompts and proof-of-concept
Kacper Grzymkowski	Sentence-embeddings generation
Kacper Grzymkowski	All-dataset clustering
Kacper Grzymkowski	Visualization for t-SNE, DSI and stop-words
Kacper Grzymkowski	Secondary clustering
Kacper Grzymkowski	Post-hoc analysis of secondary clustering
Kacper Grzymkowski	Final code deliverable debugging
Kacper Grzymkowski	DSI, question word results
Jakub Fołtyn	Report introductions
Jakub Fołtyn	Literature review WTC, topic modelling and text complexity metrics sections
Jakub Fołtyn	Literature review redaction and revisal
Jakub Fołtyn	LDA and GuidedLDA clustering
Jakub Fołtyn	Text complexity metrics computation
Jakub Fołtyn	DSI implementation research
Jakub Fołtyn	LDA methods, LDA and LLM results
Marceli Korbin	Dataset description and choice
Marceli Korbin	Explanatory data analysis on the Stanford Question Answering Dataset
Marceli Korbin	Report reorganisation
Marceli Korbin	Code notebook cleanup and reorganization
Marceli Korbin	Review #2 rebuttal (about manual labelling of clusters)
Mikołaj Malec	Bloom's taxonomy
Mikołaj Malec	Optimize number of clusters in the LDA task
Mikołaj Malec	Find best groups to present LDA task

Table 1: Our contributions

8 Acknowledgment

This project was realized in cooperation with Prof. Yoed Kenett from Israel Institute of Technology in Haifa, Israel.

References

[Benoit et al.2018] Kenneth Benoit, Kohei Watanabe, Haiyan Wang, Paul Nulty, Adam Obeng, Stefan Müller, and Akitaka Matsuo. 2018. quanteda: An r package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30):774.

[Blei and Lafferty2009] David M Blei and John D Lafferty. 2009. Topic models. In *Text mining*, pages 101–124. Chapman and Hall/CRC.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bloom et al.1956] Benjamin S Bloom, Max D Engelhart, Edward J Furst, and H Walker. 1956. Hill, and david r. krathwohl. *Taxonomy of Educational Objectives: The Classification of Educational Goals, Handbook I: Cognitive Domain*.

[Bromberger1966] Sylvain Bromberger. 1966. *Why-questions*. na.

[Brown et al.2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[Buenano-Fernandez et al.2020] Diego Buenano-Fernandez, Mario Gonzalez, David Gil, and Sergio Luján-Mora. 2020. Text mining of open-ended questions in self-assessment of university teachers: An lda topic modeling approach. *Ieee Access*, 8:35318–35330.

[Carroll1964] John B. Carroll. 1964. Chapter i: Linguistics and the psychology of language. *Review of Educational Research*, 34(2):119–126.

[Dayal2016] Veneeta Dayal. 2016. *Questions*, volume 4. Oxford University Press.

[Devlin et al.2018] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[Flesch1948] Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

[Griffiths and Steyvers2004] Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

[Hu et al.2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[Johnson et al.2023] Dan R Johnson, James C Kaufman, Brendan S Baker, John D Patterson, Baptiste Barbot, Adam E Green, Janet van Hell, Evan Kennedy, Grace F Sullivan, Christa L Taylor, et al. 2023. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior Research Methods*, 55(7):3726–3759.

[Kaufman2016] James C Kaufman. 2016. *Creativity 101*. Springer publishing company.

[Krathwohl2002] David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

[Malvern et al.2004] David Malvern, Brian Richards, Ngoni Chipere, and Pilar Durán. 2004. *Lexical diversity and language development*. Springer.

[R Core Team2021] R Core Team, 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Radford et al.2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[Rajpurkar et al.2016] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text.

[Řehůřek and Sojka2010] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA. `http://is.muni.cz/publication/884893/en`.

[Reimers and Gurevych2019] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

[Rousseeuw1987] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

[Schaeffer and Presser2003] Nora Cate Schaeffer and Stanley Presser. 2003. The science of asking questions. *Annual review of sociology*, 29(1):65–88.

[Schreiner2023] Maximilian Schreiner. 2023. Gpt-4 architecture, datasets, costs and more leaked. Blog post on the-decoder.com.

[Senter and Smith1967] RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Technical report, DTIC document.

[Simpson1949] Edward H Simpson. 1949. Measurement of diversity. *nature*, 163(4148):688–688.

[Singh and Mahmood2021] Sushant Singh and Ausif Mahmood. 2021. The nlp cookbook: Modern recipes for transformer based deep learning architectures. *IEEE Access*, 9:68675–68702.

[Templin1957] Mildred C Templin. 1957. Certain language skills in children; their development and interrelationships.

[Udupa2012] Jagadeesh Jagarlamudi; Hal Daumé III; Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In Walter Daelemans, editor, *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France, April. Association for Computational Linguistics.

[Van der Maaten and Hinton2008] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

[Xu et al.2021] Zhaozhen Xu, Amelia Howarth, Nicole Briggs, and Nello Cristianini. 2021. What makes us curious? analysis of a corpus of open-domain questions. *arXiv preprint arXiv:2110.15409*.

[Zamanian and Heydari2012] Mostafa Zamanian and Pooneh Heydari. 2012. Readability of texts: State of the art. *Theory & Practice in Language Studies*, 2(1).

[Zhou et al.2017] Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study.