# Generating Dynamic Reports on Competency Clusters Using Text Generation and Paraphrasing Approaches

Alfan Aris Setiawan and Masayu Leylia Khodra

School of Electrical Engineering and Informatics
Bandung Institute of Technology
Bandung, Indonesia
alfanaris@gmail.com, masayu@staff.stei.itb.ac.id

*Abstract*: An assessment center is an assessment tool that is carried out using multi-simulation, multi-assessors, and an assessment aggregation process that are all mandatory in the implementation. Currently, the competency cluster dynamics report, one of the components of the assessment report, is manually made by an assessor. This study proposes the use of Natural Language Generation (NLG) as an alternative solution for creating the competency cluster dynamics report. This study uses three approaches to create the NLG model. The template-based model can create paragraphs directly from tabular data, where the model is created by defining a sentence frame and then filling the frame with relevant information. Then, the data-to-text approach is carried out by transforming tabular data into a flat string format (linearization) as a model input. Finally, the text-to-text approach, a paraphrasing model, trains a pre-trained language model from text data input using template-based output as model input. Following quantitative and qualitative evaluations, the text-to-text model was found to be the model with the best report output. In terms of fluency, the model had the same score as the report by humans. On the other hand, in terms of faithfulness and coherence, the model had higher evaluation scores than the report by humans. Meanwhile, template-based model outperformed reports by humans and text-to-text models in the faithfulness. Data-to-text approach had the lowest evaluation score of all models.

*Keywords*: Assessment center; data-to-text; Natural Language Generation; Pre-Trained Language Model; template-based; text-to-text

## 1. Introduction

An assessment center is an assessment tool that uses various evaluation approaches to measure several behaviors that reflected the ability needed to succeed in a scope of a job. In its implementation, the use of multi-simulation, multi-assessor, or the assessment aggregation process in the assessor meeting is mandatory and cannot be eliminated [1]. Nonetheless, generating a report can be supported by technology.

The competency cluster dynamics report is prepared based on the competencies required in the job demonstrated during the simulation process with an assessor as an observer. The report explicitly describes the participant's strengths through key behaviors that can be demonstrated by the participant during the simulation. Key behavior itself is an indicator of behavior in mastering related competencies. In addition to describing the participant's strengths, the report must include areas of future development of the participant with a narrative that complements the existing key behavior.

Natural Language Generation (NLG) can be used as an alternative solution for creating competency cluster dynamics reports, which takes up a lot of assessor work time. NLG is a task for creating natural language text from structured or semi-structured data [2]. The assessment results are tabular data in the form of scores for each key behavior with a boolean value indicating whether the key behavior appeared during the assessment or not, making this it an NLG a data-to-text subtask.

In this study, an NLG method is proposed to generate competency cluster dynamics report, which are still manually prepared by assessors. This NLG model is proposed to replace the role

Received: February 2nd, 2025. Accepted: September 23rd, 2025
DOI: 10.15676/ijeei.2025.17.3.12

of assessors in making the reports so that assessors can be more productive and focus on other key aspects in of the implementation of the assessment center which is also an obligation in the assessment implementation. This study provides scientific contributions in the form of: 1). Application of generative AI in the domain of making assessment center reports, 2). Application of generative AI in making Indonesian language reports, and 3). Application of classical, neural, and hybrid methods in generating reports from tabular data.

In this paper, section 2 discusses the literature review related to the assessment center and the data-to-text approach. Section 3 discusses the methods carried out in this research while section 4 presents the results of the research. Lastly, section 5 conveys research conclusion and discusses future works.

## 2. Literature Review

This section consists of three parts. First, it discusses the assessment center and its reporting. Next, the section focuses the NLG approach on the data-to-text task.

### *A. Assessment Center*

An assessment center is a standardized evaluation method used to measure behavior based on multiple inputs. The assessment center is scored from a series of behavioral simulations by trained assessors, who observe, record, and classify behavior according to the relevant behavioral constructs. After the assessment process is conducted according to regulations, an overall assessment rating (OAR), which is a scoring process conducted through assessor consensus or statistical aggregation, is performed. The resulting score indicates a person's position on the behavioral construct [1].

The assessment center has ten important elements [1], namely:

1. Systematic analysis to determine job-relevant behavioral constructs, which focuses on identifying specific, observable, and job-relevant behaviors.
2. Behavioral classification, which discusses the categorization of behavior according to predetermined behavioral constructs.
3. Multiple assessment center components. Each assessment center must have various assessment components to ensure that the behavioral information provided is reliable, objective, and relevant.
4. Linkages between behavioral constructs and assessment center components, which is a matrix that describes the behavioral constructs assessed in each assessment center component.
5. Simulation exercises explained in the assessment center should consist of various simulations so that there is an opportunity to see behavior that is relevant to the behavioral construct being assessed.
6. Assessors are needed to observe and evaluate each participant. The assessment center must use more than one assessor who are diverse in demographics and experience.
7. Assessor training ensures that assessors are comprehensively trained and demonstrate performance that meets the established criteria.
8. Recording and scoring of behaviors. In accordance with the established behavioral construct, the assessor must use systematic procedures to accurately record and score behavioral observations.
9. Data integration combines observations and assessments of each participant's behavior based on discussions of combined observations and assessments from multiple assessors or using statistics according to professional standards.
10. Standardization is necessary to ensure that all assessed participants have an equal opportunity to demonstrate behavior relevant to the behavioral construct.

### *B. Competency Cluster Dynamics Report*

In the assessment report, the assessment center must pay attention to two factors. The first factor is the content report, which consists of the description of each competency obtained from

the behavior of the assessee during the assessment. The second factor is the style of report writing, where the report must be written in a language that is easily understood by lay people; not specifically for those experienced in the field of human resources or psychologists [3].

In general, the report contains a profile of the assessee along with recommendations for assessment results for a position. Details of the assessment results contain assessment rating scores along with descriptions of each competency. In addition, there is an executive summary section that contains the competency cluster dynamics report. This report provides an assessment summary describing the participant's strengths and areas for future development based on key behaviors observed during the job simulation. The assessee's strengths and weaknesses are presented concisely within a limited number of characters.

Competency clusters are a collection of competencies that are related to each other in a scope of work or role. Each cluster has a unique name to indicate the scope of the framework. Competency is a combination of employee knowledge, skills, and behavior that are observed, measured, and applied in their work, as well as the values, motivation, initiative, and self-control involved in creating added value to produce competitive advantage of a company. The nature of competencies must be observed and measured. Each competency contains several key behaviors or so-called specific key behaviors that can also be observed, developed, and assessed.

Key behavior is a series of behavioral indicators required as evidence of mastery of related competencies. Key behaviors, when displayed effectively, make the soft competency effective. Key behaviors can be used to evaluate an individual's readiness to carry out their role, both in the role of the position and the function entrusted to them, and to direct their development plans for behaviors that are not yet effective. The order in which key behaviors are written does not indicate the order of achievement.

*C. Data-to-Text*

Data-to-text aims to produce text that can accurately and fluently describe non-linguistic structured data. There are two main challenges in data-to-text tasks: "What to say" and "How to say". "What to say" refers to the process of analyzing and selecting structured data, where part or all the data is taken for abstraction and association. Meanwhile, "How to say" refers to the ability to present the selected data clearly and fluently using natural language [4].

The approaches in data-to-text generation can be broadly categorized into two types: (a) the traditional approach and (b) the end-to-end neural model. The traditional approach follows pipeline architecture for natural language generation (NLG), which typically consists of three main stages: document planning, micro planning, and surface realization. Within this traditional modular approach, there are two primary categories. The first is the template-based modular approach, which uses predefined rules and templates created by humans to generate text in a fixed structure. The second is the statistical modular approach, which employs probabilistic models—such as probabilistic language generation or probabilistic context-free grammar—to produce text based on statistical patterns learned from data.

The classic NLG pipeline architecture, which includes text (or document) planning, sentence (or micro) planning, and linguistic (or surface) realization, was first introduced in [5]. The template-based approach provides accurate and controlled outputs but lacks flexibility and requires extensive manual configuration to design and maintain templates [4]. In addition, there is a statistical modular approach method that uses a probabilistic model to learn patterns from training data and generate text based on probability. This method is more flexible but is prone to syntax errors and is highly dependent on training data. Both of these methods are included in the classical method.

Apart from the classical method, there is a method that is more commonly used, namely the end-to-end method using deep learning or neural networks [4]. Models can be created and trained from scratch with neural network configurations according to user needs, or pre-trained models can be used and then retrained with data in more specific domains for specific tasks.

Reference [6] developed a template-based text generator model that was adapted from the architecture in [5]. The architecture added the use of template generation and also used facts

representation [7]. In the evaluation conducted by humans, it was found that there were too many unnecessary repetitions in the articles created, such as repetition of names or locations. In addition, it was also found that the generated articles were too long, and readers were bored because of the minimal variation in the articles.

Paraphrasing, one alternative to dealing with variation problems, is a restatement using different words while maintaining the original meaning [8]. Research related to paraphrasing was conducted by [9] using a Pre-Trained Language Model (PLM) based on a neural network. The study was conducted by fine-tuning the PLM with training data in the form of sentence pairs and paragraph pairs that had the same meaning but used different words. The data originated from human-made paraphrases. The results of the study showed that the model could produce paraphrases not only at the sentence level but also at the paragraph level.

NLG uses neural networks, one of which was researched in [10] and [11]. Both of them utilized PLM to perform data-to-text generation tasks with tabular data first created in flat string format (linearization) input and paired with paragraphs as the output in the PLM fine-tuning dataset. Study in [11] performed text generation in the form of sentences, while in [10] the data is made in paragraph form. The model developed in [11] could produce good output, while in [10] model, which produced paragraphs, had difficulties, such as producing long texts.

## 3. Research Methodology

This section discusses the solution design for developing an NLG model with a classical template-based approach and a deep learning approach to data-to-text and paraphrasing, with the ability to generate reports from tabular data that resemble reports made by humans.

### *A. Dataset*

The production of competency dynamics reports is currently still done manually by assessors. Reports are created based on key behaviors fulfilled by assessment participants. Key behaviors have codes and behaviors required in the context of work. Each key behavior is then interpreted into sentences. Each sentence can represent one or more key behaviors. As an illustration, examples of key behavior data instances and transformations into competency cluster dynamics reports are presented in Tables 1 and 2.

Table 1. Key Behavior Score Results of Assessment Center Participants

| Code | Key Behaviour | Value |
|---|---|---|
| ADA-KB1 | Mencoba memahami perubahan-perubahan<br>(Trying to understand changes) | 1 |
| ADA-KB2 | Menyikapi perubahan dengan pikiran positif<br>(Responding to changes with positive thoughts) | 1 |
| ADA-KB3 | Menyesuaikan perilaku<br>(Adjusting behavior) | 1 |
| ADA-KB4 | Menyesuaikan interaksi antar pribadi<br>(Adjusting interpersonal interactions) | 0 |
| ADA-KB5 | Bertindak mandiri<br>(Acting independently) | 0 |
| ADA-KB6 | Melakukan lebih dari yang diminta<br>(Doing more than what is asked) | 0 |

The data in this study used data from the assessment center of a company in Indonesia. From January 2023 to June 2024, there were 467 competency cluster dynamics report data made by 27 different assessors. The data was stored in the company's information system and was used as a dataset in this study. In this study, augmentation was done by generating tabular data using template-based. The same data instances were generated using the template-based model in the desired amount as needed.

Table 2. Competency Cluster Dynamics Report

| *Competence Cluster Dynamics Report* |
| --- |
| Sdr. mampu mengenali terjadinya perubahan dalam organisasi atau lingkungan kerja dan memahami hal tersebut sebagai sesuatu yang wajar. Kondisi tersebut ia jadikan sebagai suatu kesempatan untuk melakukan perbaikan-perbaikan pada sejumlah aspek. Berbagai pendekatan ia coba lakukan guna menyelesaikan permasalahan yang ada. Namun demikian, ia perlu lebih menunjukkan inisiatifnya untuk bertindak sesuai kewenangannya tanpa menunggu perintah/arahan.<br><br>(Mr. X is able to recognize changes in the organization or work environment and understand them as something normal. He uses this condition as an opportunity to make improvements in several aspects. He tries various approaches to solve existing problems. However, he needs to show more initiative to act according to his authority without waiting for orders/directions.) |

*B. Template-Based*

In its implementation, the template-based approach can be directly used to generate the competency cluster dynamics report without the need for initial processing. This approach can flexibly produce paragraph output from tabular data input.

However, despite its simplicity and efficiency, the template-based approach has limitations in terms of flexibility to data variations and adaptability to changes in data patterns or structures. This approach relies heavily on pre-defined rules or templates. As such, if the incoming data does not match the expected structure, the output results could be less accurate.

Template-based NLG in [6] requires handling in terms of variations in the text output produced. A small modification by adding a "mini corpus" in the template-based model can be an alternative solution to increase output variations. The "mini corpus" contains variations of phrases so that the output of the template-based model varies according to the amount of data in the "mini corpus".

The design of the mini corpus and template-based usage flow is built based on the distribution of key behavior data in the dataset. Each key behavior can be in a one-sentence form in the report, or more than one key behavior is narrated in one sentence.

The creation of a template-based model was done by dividing it into four modules and seven submodules. The modules and submodules are as follows:

1. Mini Corpus
   Mini corpus is a collection of templates that function to supply templates needed by the model. The mini corpus consists of individual key behavior groups. There are also conjunction groups and combo key behavior groups, which are templates containing a combination of two or three key behaviors in one sentence. The number of templates in the mini corpus was around 450 templates.
2. Document Planner
   The document planner consists of two sub-modules, namely:
   a. Content determination, refers to selecting relevant data according to the main goal to be achieved.
   b. Document structuring, refers to determining how content is sorted or grouped to organize information for easier understanding. In this model, sorting is done based on the key behavior score, where the score of one is placed first.

Table 3. Pseudocode Document Planner

```
Define a function sorting(data):
    Initialize an empty list `positif`
    Initialize an empty list `negatif`

    For each column in `data`:
        If the value of the column is 1:
            Append the column name to the list `positif`
        Otherwise:
            Append the column name to the list `negatif`

    Return the lists `positif` and `negatif`
```

3. Microplanner
   The microplanner consists of three sub-modules, namely:
   a. Lexicalization, refers to the process of converting tabular data into a language that can be easily understood by humans. This involves translating numbers into clear descriptive sentences so that readers can easily understand the meaning of the data.
   b. Aggregation, refers to the combination of several sentences or phrases into one sentence without losing its main meaning. This process involves condensing information and/or using appropriate conjunctions to maintain the relationship between ideas. There are several functions for performing aggregation in this model, including the use of the compound phrase group with more than one key behavior so that a combination of sentences is obtained, representing more than one key behavior.

Table 4. Pseudocode Aggregation

```
Define function `combo_ada56(data)`:
  If `data['ada_5']` and `data['ada_6']` are equal to 1:
    If True:
      Generate a random integer between 0 and 1
      If random number is 0:
        Select random template from mini corpus ‘ada56’
        Store it in `report_temp`
      Else (random number is 1):
        Combine mini corpus ['ada_5'] and ['ada_6']
```

   c. Referring expression generation, refers to reducing the repetition of stating the same entity by using pronouns or synonyms to replace repeated names or terms. This aims to keep the text clear and easy to read without mentioning the same entity too often.

Table 5. Pseudocode Referring Expression Generation

```
Define function `combine_report(pos, neg)`:
    Get random `conj` from mini corpus
    Text = f"{pos}. {conj} Ia {neg}."

    Return Text
```

4. Surface Realizer
   The surface realizer consists of two sub-modules, namely:
   a. Linguistic realization, which refers to how phrases are formed through affixation processes such as adding prefixes, infixes, or suffixes, reduplication, and word composition.

Table 6. Pseudocode Linguistic Realization

```
Define function `gen_conj_pos(data1, data2)`:
   Initialize an empty string `text`
   Get random `conj`
   Get random `after_ia`
   Text = "{data1}. {conj} Ia {after_ia} {data2}"

   Return Text
```

b. Structure realization, which refers to changing every word or phrase into a structured sentence. In this module, a function is created to call other functions and combine them into one complete paragraph.

Table 7. Structure Realization

```
Define function `generate_report(data)`:
   Apply `lexicalization(data)`
   Initialize an empty string `report`

   Generate a random `pre_report`

   Call `combo_ada123(data)`
   Call `combo_ada456(data)`
   Call `combo_cle124(data)`
   Call `combo_cle356(data)`
   Call `sorting(data)` to obtain `positif` and `negatif`

   Pos = `data` for columns in `positif`
   Neg = `data` for columns in `negatif`

   Call `combine_pos(pos)
   Call `combine_neg(neg)

   Body_report = Call `combine_report(pos_text, neg_text)`

   Report = Concatenate `pre_report` and `Body_report`

   Return Report
```

*C. Data-to-Text*

Converting tabular data to paragraphs can be done in the data-to-text approach to produce a report in paragraph form, although it requires a little preprocessing of the tabular data into a flat string (linearization) input. This linearization aims to change tabular data that is structural in nature into a linear text representation that is easier to process by a data-to-text based model.

The linearization process usually involves transforming data into a key-value pair format or a simple structured sentence that represents the contents of the table concisely. The challenge of this approach is the need to train the model with a large enough dataset to ensure good generalization. With limited data, the model cannot produce text according to desire and can even produce output that has no meaning at all. An example of flat string (linearization) input data is shown in Table 3.

Table 8. Example of Data Flat String (Linearization) Input

| Flat String (Linearization) Input |
|---|
| <Code>ADA-KB1</Code><Key Behaviour>Trying to understand changes</Key Behaviour><Value>1</Value><Code>ADA-KB2</Code><Key Behaviour>Responding to changes with positive thoughts</Key Behaviour><Value>1</Value><Code>ADA-KB3</Code><Key Behaviour>Adjusting behavior</Key Behaviour><Value>1</Value><Code>ADA-KB4</Code><Key Behaviour>Adjusting interpersonal interactions</Key Behaviour><Value>0</Value><Code>ADA-KB5</Code><Key Behaviour>Acting independently</Key Behaviour><Value>0</Value><Code>ADA-KB6</Code><Key Behaviour>Doing more than requested</Key Behaviour><Value>0</Value> |

The model in the research conducted in [11] and [10] was trained using a large amount of data. Although a large amount of data was used, the generated text in the form of paragraphs did not produce good results.

Tabular data was converted into flat string (linearization) text and paired with the original expert-made report as a dataset to perform fine-tuning of the pre-trained language model. The data of 467 instances was then divided into training data, validation data, and test data with a composition of 90:5:5. Furthermore, fine-tuning was carried out to obtain a data-to-text model. The implementation flow of this approach is presented in Figure 1.

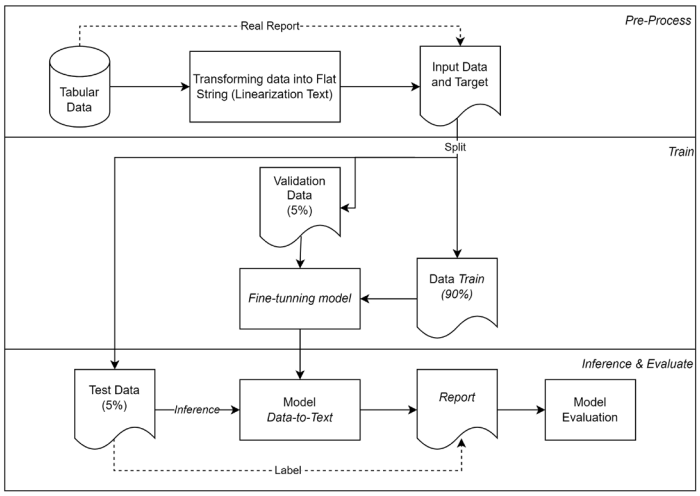


Figure 1. Data-to-Text Implementation Diagram

*D. Text-to-Text*

Unlike the other two approaches, the text-to-text approach requires input data in the form of text. Thus, this approach cannot be used directly to generate the competency cluster dynamics report which has a data source in the form of tabular data. Initial processing needs to be done so that the tabular data is first converted into text data so that it can be used as input for this model.

This approach can still be utilized by first transforming tabular data into text format. The transformation involves the process of encoding tabular data into text representations that represent important information from the data. Once the data is converted into text, the text-to-text model can be trained to generate reports with a specific narrative pattern based on the processed input. As a result, the use of the text-to-text approach requires a longer pipeline.

Initially, the tabular data was converted into text data before conducting PLM training with a paraphrasing task to generate the competency cluster dynamics report. The development of this

model utilized the output of a template-based model that converted tabular data into text. Furthermore, PLM training was carried out with a paraphrasing task.

This paraphrasing process aims to make the model more flexible in producing different variations of text but still maintain the same meaning. Hence, it can be applied in various scenarios and provide more natural and less rigid reports. Although the process is more complex, this text-to-text approach offers greater flexibility and adaptability in generating data-based reports or texts. The approach scheme is illustrated in Figure 2.

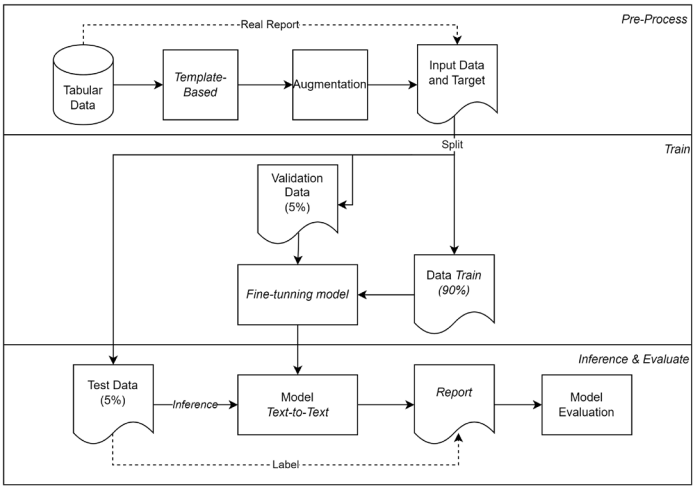


Figure 2. Text-to-Text Implementation Diagram

## 4. Evaluation And Results

This section describes the research conducted along with the evaluation of the experimental results. It presents the experimental results, their evaluation, and a discussion of the results to support the feasibility of the proposed approach.

### *A. Evaluation Method*

The evaluation metrics used were BLEU [12] and ROUGE [13]. Both metrics have slightly different approaches but they are conceptually similar. The metrics use a reference as a standard to compare the quality of the system's output, followed by calculating the similarity between the text output from the system and the reference. These metrics have limitations in identifying deeper semantic relationships. While both metrics explicitly focus on word correspondence, a machine learning-based metric called BERTScore [14] was used to understand the context and meaning of words in the text.

Specifically, BERTScore is used to measure semantic similarity between texts. BERTScore uses contextual representations from the BERT (Bidirectional Encoder Representations from Transformers) model. Word embeddings generated by BERT, which capture the meaning of words in their context, are used as contextual representations. Unlike metrics such as BLEU or ROUGE, BERTScore not only compares n-grams but also considers semantic relationships between words.

In addition to using an evaluation matrix, expert evaluation was also used in this study. Reference [6] employed qualitative evaluation. The evaluation was conducted on a scale of 1 to 4 for 10 assessment aspects. This research adopted the aspects used in the aforementioned research by combining them according to the needs compiled by experienced assessors in its creation.

The created aspects were then categorized into three groups [4]. The first group is fluency, which assesses whether the resulting text can be read fluently without striking grammatical errors. Faithfulness, the second group, ensures whether the resulting text remains in accordance

with the information or instructions given in the input. The last group is coherence, which is an aspect that assesses whether the text output is logically structured and whether the order of sentences and ideas is in accordance with the patterns commonly used in human writing. The qualitative evaluation employs 10 aspects with 3 categories as shown in Table 4.

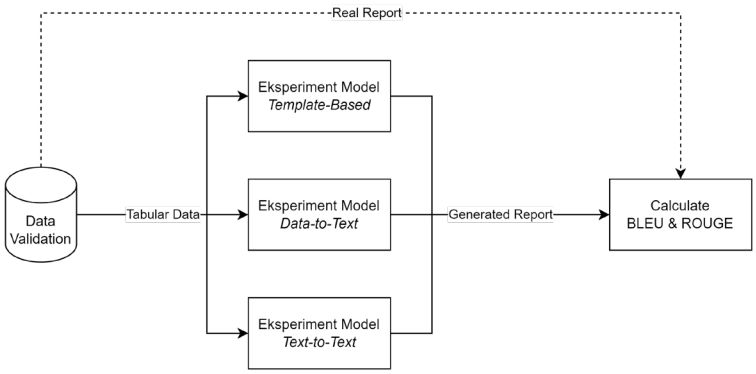


Figure 3. Evaluation Diagram

Table 9. Qualitative Evaluation

| Category | Aspect |
|---|---|
| Fluency | 1. The language used is clear, straightforward, and unambiguous. This aspect is directly related to fluency because clear and unambiguous language ensures that the text is easy to read and understand without any doubt.<br>2. The report uses common, simple words that are understandable to the general public. Using common and simple words affects fluency, ensuring that the text is easy to understand for a wide audience.<br>3. Using enhanced spelling. Using correct and standard spelling is part of fluency as spelling errors can disrupt the fluency of text reading. |
| Faithfulness | 1. Key behavior conformity. This aspect focuses on conformity as it ensures that the generated text is inline with the information provided in the input or existing data.<br>2. The sentences used describe the dynamics of all competencies in the cluster. Ensure that the text includes all the necessary information according to the input and instructions, thus reflecting suitability.<br>3. Still given development areas even though the target has been met. This section is related to the suitability of providing accurate information by ensuring that additional information remains connected to the main topic of the report. |
| Coherence | 1. Information is presented in a structured manner. Arranging information in a structured format ensures coherence, where the reader can follow the flow of information easily.<br>2. The report is presented in an interesting way and is able to attract the reader's attention. Maintaining the reader's attention through an interesting presentation contributes to coherence.<br>3. Use positive and constructive language. Positive language increases coherence because it creates a narrative that is enjoyable and easy to follow. |

| | |
|---|---|
| | 4. Describe the participant's areas of strength first, and, then their weaknesses. Structure the information to create a narrative that is enjoyable and easy to follow. |

Experts who are experienced in compiling competency cluster dynamics reports assessed the reports generated by the model. The evaluation scheme performed by experts was selected for three instances. Tabular data from the three instances was then generated by ChatGPT [15] using the one-shot prompting method. Then, the reports that was previously made by experts from these three instances were also assessed by experts, so that for one instance, five report variations were assessed by experts. They were asked to assess using a scale score of 1 to 4. A score of 1 means very inappropriate, 2 is inappropriate, 3 is appropriate, and 4 means very appropriate.

*B. Experiments Setup*

Experiments on the data-to-text model were conducted in three models. The experiments conducted were limited since there was not sufficient data. In addition, the initial validation results also did not produce good scores as shown in Table 5.

Table 10. Data-to-Text Validation Data Evaluation Matrix

| No. | Model | Amount of Data | Epoch | BLEU | Max BLEU | ROUGE-L |
|---|---|---|---|---|---|---|
| 1 | IndoT5-Base | 467 | 16 | 0.011 | 0.030 | 0.157 |
| 2 | IndoT5-Large | 467 | 16 | 0.012 | 0.050 | 0.184 |
| 3 | mT5 | 467 | 16 | 0.003 | 0.004 | 0.067 |

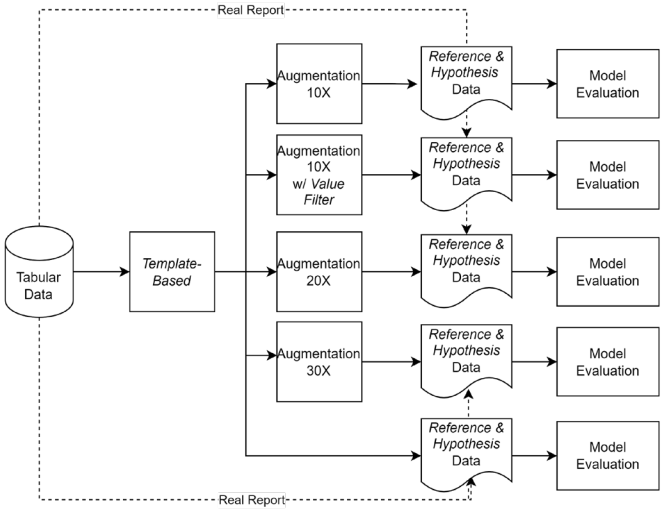


Figure 4. Template-Based Scheme

The highest BLEU score was 0.012 with the maximum BLEU score of 0.050, while the ROUGE L score was 0.184. All the highest scores were generated from the IndoT5-Large model.

The template-based model approach does not specifically carry out a test scenario. The results of the template-based model were used as training data for the text-to-text model. The results were different because there is a configuration in the augmentation. Furthermore, the output results were also selected as samples for qualitative testing by the assessors.

The configuration carried out was related to the use of filter values when selecting the template used. If the filter value is active, the template that has been used cannot be reused for the next report generation. This makes the use of filter values produces different variations in

each use. Furthermore, the output of the template-based model become a dataset for the text-to-text model after being combined with the original report made by the assessors. The template-based scheme is presented in Figure 3.

More experimental models were run in the text-to-text approach compared to the other approaches. The experiments were divided into three stages aimed at finding model configurations, the amount of augmentation data, prompts, and epochs. The first stage of the experiment tested the model that could produce the best performance out of the models used, such as IndoT5-Large, IndoT5-Base-Parafrase, and Llama-Instruct 3.2 1B.

In the second stage of the experiment, the model that produced the best performance from the first stage was selected. The experiments were then carried out with configurations on the amount of data and prompts. Lastly, in the third stage, the configuration was employed on the prompt and the number of epochs. The details of the experiment's prompts are presented in Table 6.

Table 11. Prompts of Experiments

| No. | Prompt |
|---|---|
| Prompt 1 | "Anda adalah seorang asesor assessment center yang berpengalaman selama 10 tahun membuat laporan assessment center. Tugas Anda adalah membuat parafrase laporan dengan tetap mempertahankan makna dan inti dari informasi. Identifikasi ide utama setiap paragraf dalam laporan input. Buatlah ulang setiap ide utama tersebut menggunakan sinonim, struktur kalimat berbeda, atau gaya bahasa yang lebih naratif. Perhatikan untuk menjaga akurasi informasi terkait kompetensi dan hasil asesmen. Gunakan narasi yang membangun dan hindari kata maupun kalimat negatif, serta pastikan paragraf baru tetap terhubung secara logis antar kalimat. Gunakan bahasa profesional dan narasi informasi faktual.\n\nParafrase laporan hasil asesmen center berikut.\n\n"<br><br>("You are an assessment center assessor with 10 years of experience in creating assessment center reports. Your task is to paraphrase the report while maintaining the meaning and essence of the information. Identify the main idea of each paragraph in the input report. Rewrite each main idea using synonyms, different sentence structures, or a more narrative style. Pay attention to maintaining the accuracy of the information related to competencies and assessment results. Use constructive narratives and avoid negative words or sentences, and ensure that new paragraphs remain logically connected between sentences. Use professional language and factual narrative information.\n\nParaphrase the following assessment center report.\n\n") |
| Prompt 2 | "Buat parafrase laporan dengan tetap mempertahankan makna dan inti dari informasi. Gunakan narasi yang membangun dan hindari kata maupun kalimat negatif, serta pastikan paragraf baru tetap terhubung secara logis antar kalimat. Parafrase laporan berikut:"<br><br>("Paraphrase the report while maintaining the meaning and essence of the information. Use constructive narrative and avoid negative words and sentences, and make sure the new paragraph remains logically connected between sentences. Paraphrase the following report:") |
| Prompt 3 | "Paraphrase :" |

The augmentation option could only be done up to 10 times for the value filter usage option. This is due to the number of mini corpus to produce different augmentations for each same instance is not sufficient if the iteration is done more than 10 times. As such, the amount of data used in this first stage of the experiment had no variation, which was 4203:233:234 for training data, validation data, and test data, respectively. The data was fully augmented before being split into three data. Prompt variations were also carried out to determine the effect of the prompt on the results of the model. The results of the first stage of the experiment are presented in Table 7.

Table 12. Experiment Results of First Stage

| No. | Model | Value Filter | Masked Name | Prompt | BLEU | Max BLEU | ROUGE-L |
|---|---|---|---|---|---|---|---|
| 1 | IndoT5-Large | x | x | - | 0.068 | 0.787 | 0.226 |
| 2 | IndoT5-Large | x | x | 1 | 0.074 | 0.640 | 0.231 |
| 3 | IndoT5-Large | v | x | 1 | 0.073 | 0.517 | 0.224 |
| 4 | Llama 3.2 1B Instruct | v | x | 1 | 0.011 | 0.062 | 0.137 |
| 5 | IndoT5-Large | v | v | 1 | 0.031 | 0.252 | 0.201 |
| 6 | IndoT5-Base-Parafrase | v | x | 3 | 0.045 | 0.305 | 0.209 |
| 7 | IndoT5-Base-Parafrase | v | x | 2 | 0.040 | 0.286 | 0.204 |

The best score was obtained from the second experiment with the IndoT5-Large model configuration, augmentation without value filters and masked names, and the first version of prompts. However, the highest BLEU score was obtained from a model that did not use prompts. The second stage of the experiment could be done by adding the number of augmentations because augmentation with filter values was not used. Thus, there was no limit to the number of data augmentations performed. Next, the result of second stage experiment are presented in Table 8.

Table 13. Experiment Results of Second Stage

| No. | Prompt | Amount of Data | BLEU | Max BLEU | ROUGE-L |
|---|---|---|---|---|---|
| 1 | 1 | 9340 (20X) | 0.221 | 1.000 | 0.351 |
| 2 | 1 | 14010 (30X) | 0.776 | 1.000 | 0.815 |
| 3 | 2 | 9340 (20X) | 0.263 | 1.000 | 0.385 |
| 4 | 2 | 14010 (30X) | 0.781 | 1.000 | 0.819 |
| 5 | - | 9340 (20X) | 0.223 | 1.000 | 0.352 |
| 6 | - | 14010 (30X) | 0.675 | 1.000 | 0.731 |

The best combination was obtained from the second stage experiment, i.e., 30x data augmentation and with the second prompt. The BLEU score for the experiment with 30X augmentation was very high compared to the first stage experiment. In addition, the maximum BLEU scores of the second experiment were all one, indicating a possible overfitting.

The results of the sampling performed on many data instances with the same name in the validation data obtained identical or at least close scores, especially in instances with high scores. This can be due to split data without exclusively separating the data instances so that they do not appear more than once of the data sets between training, validation, or test data.

Next, the third stage of the experiment explored the prompt and epoch. In addition, the amount of data was also added to the data obtained from July to November 2024. In this experiment, changes were made during data splitting. Instead of performing data augmentation, followed by splitting, in this third stage of the experiment, the data was split first and then augmented to the training data only. This is expected to reduce the model overfitting.

The training data, which previously only had 420 instances, was added with 397 instances, making up a total training data of 867 instances. The new data was not split but it was directly combined into a training data because the data was not time-bound. The results of the third stage of the experiment are shown in Table 9.

Table 14. Experiment Results of Third Stage

| No. | Prompt | Epoch | Early Stop | Amount of Data | BLEU | Max BLEU | ROUGE-L |
|---|---|---|---|---|---|---|---|
| 1 | 2 | 16 | x | 867 | 0.025 | 0.069 | 0.208 |
| 2 | 2 | 16 | x | 8670 | 0.036 | 0.185 | 0.200 |
| 3 | 2 | 32 | x | 867 | 0.000 | 0.002 | 0.038 |
| 4 | 2 | 32 | x | 8670 | 0.036 | 0.255 | 0.200 |
| 5 | - | 16 | x | 8670 | 0.025 | 0.074 | 0.204 |
| 6 | 1 | 16 | x | 8670 | 0.036 | 0.128 | 0.200 |
| 7 | 2 | 10 | x | 8670 | 0.029 | 0.126 | 0.213 |
| 8 | 2 | 32 | v | 8670 | 0.036 | 0.121 | 0.200 |

The BLEU score that previously increased in the second stage experiment was decreased again in the third stage experiment. This may occur since the task in this model involved a one-paraphrase paragraph where each instance can be different from the others, making it difficult to predict. In addition, the input style that comes from the template can be very different from the target style, which is human made. The reports were also made by more than 30 different assessors, resulting in report style variability, which, in turn, caused challenges for the model to imitate the target data specifically.

Following the third experiment, an evaluation to identify the best model from each approach was performed. Testing was done with test data. The evaluation was done quantitatively using BLEU, ROUGE, and BERTScore matrices, and then continued qualitatively by experts.

*C. Testing*

Testing was done with an evaluation matrix to determine the evaluation score of test data. After that, three reports with the highest BLEU and BERTScore scores on the text-to-text model were evaluated qualitatively to determine whether the models obtained good results.

Quantitative testing measures the BLEU, ROUGE-L, and BERTScore scores from reference data, namely the competency cluster dynamics report made by the assessors, compared to the hypothesis generated by the best model from each approach. The test utilized test data that was previously split into 24 instances. The models tested for each approach are presented in Table 10.

Table 11 shows the test results using the BLEU, ROUGE, and BERTScore evaluation matrices of the test data, suggesting that the text-to-text approach had the best evaluation results. Furthermore, three data were selected with the highest BLEU and BERTScore evaluation scores. The BLEU score from the text-to-text model was taken as the top three highest, then paired with reports from template-based and data-to-text with the same identity as the text-to-text report. In addition, the original reports made by the assessors were also assessed along with the ChatGPT report. The five reports for each instance were randomized. Each subsequent instance had a different report generation result sequence from each model to avoid assessment bias.

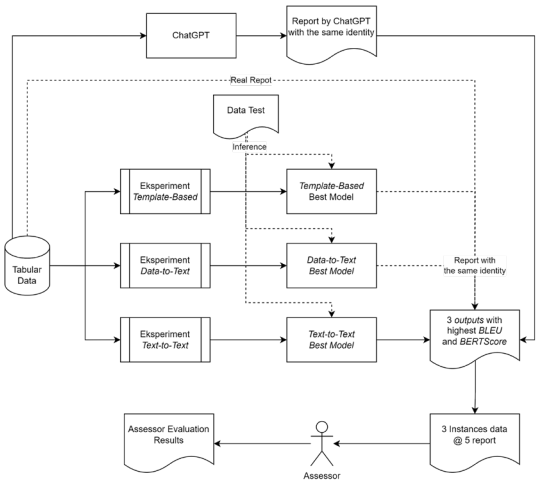


Figure 5. Test Scheme

Table 15. Best Model of Each Validation Data Evaluation Approach

| No. | Approach | Model | Prompt | Epoch | BLEU | ROUGE-L |
|---|---|---|---|---|---|---|
| 1 | Template Based | Template Based | - | - | 0.019 | 0.225 |
| 2 | Data-to-Text | IndoT5-Large | - | 16 | 0.012 | 0.184 |
| 3 | Text-to-Text | IndoT5-Large | 2 | 32 | 0.036 | 0.200 |

Table 16. Best Model Testing Results of Each Approach

| No. | Approach | Model | BLEU | Max BLEU | ROUGE-L | BERTScore |
|---|---|---|---|---|---|---|
| 1 | Template Based | Template Based | 0.024 | 0.057 | 0.184 | 0.740 |
| 2 | Data-to-Text | IndoT5-Large | 0.016 | 0.052 | 0.191 | 0.652 |
| 3 | Text-to-Text | IndoT5-Large | 0.038 | 0.165 | 0.209 | 0.745 |

The assessment form, in the form of a Google spreadsheet, was given to the assessors. The Google spreadsheets was used since this tool was considered as the easiest tool to be used by the assessors. Each assessor assessed three data instances consisting of five reports for each instance. In the end, each assessor assessed 15 reports with 10 assessment aspects. The reports were assessed by 9 assessors. The assessment results are presented in Figure 5, Figure 6, and Figure 7.

This method is validated through a blinded assessment process involving anonymous evaluators who rated each report without prior knowledge of whether it was generated by an AI model or a human author. By employing this double-blind evaluation method, potential bias toward either human- or machine-generated content was minimized, ensuring the reliability and objectivity of the scoring results.

The assessment results revealed that the data-to-text model had the smallest score close to 1 for each category. Meanwhile, the highest value in each evaluation category was obtained by ChatGPT, indicating that ChatGPT outperforms other models in every aspect.

The text-to-text model achieved the second highest score in the fluency category, below the ChatGPT score, and it matched the original report. This shows that the model in this study is as good as the original report in terms of fluency. The template-based model was in fourth place based on the fluency.

In the faithfulness category, the template-based and the text-to-text models ranked second and third, respectively, outperforming the original report. These results indicate that both models can produce a competency cluster dynamic report that is in accordance with the completeness of its key behavior. In fact, the original report by the assessor is considered inappropriate for fulfilling the key behavior in the report.

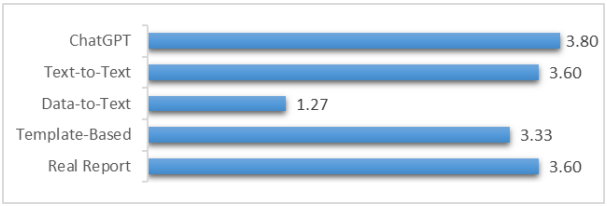


Figure 6. Fluency Evaluation Score

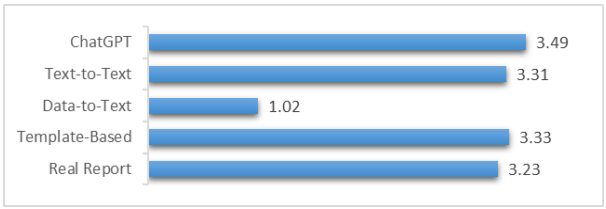


Figure 7. Faithfulness Evaluation Score

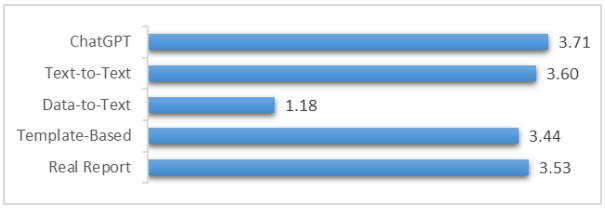


Figure 8. Coherence Evaluation Score

In terms of coherence, the text-to-text model outperformed the original report score by the assessor and other models. On the other hand, the template-based model was unable to gain a better score than the original report. This shows that the text-to-text model can imitate the logical structure and sequence of sentences and patterns commonly used by humans, although the score is still below ChatGPT.

The results of the qualitative evaluation show that the text-to-text model had good performance, where the results exceeded the original report in two of the three assessment categories. That being said, ChatGPT which "only" uses the One-Shot prompt can produce a report that exceeds other models and also the report made by the assessor in all categories.

*D. Error Analysis*

The test results were analyzed on three data samples that had undergone qualitative assessment. The best value in each approach was subjected to error analysis. In this case, the best value was the one with the best qualitative value.

In the template-based model, the report received the lowest category score assessment, namely fluency. Nonetheless, the report was good in the faithfulness category, which illustrates that the report is sufficient in representing conformity with the input data. This is evidenced by the completeness of the delivery of strengths and weaknesses and the use of good positive language. In addition, the development area was conveyed even though it had met the criteria. However, the report is still considered less interesting and monotonous, especially in the repetition of the pronoun "Ia" (He/She) or the use of the word "dalam" (in/at), which was too

much in one sentence. Another limitation found was the bridging from one sentence to another, which was considered not flexible enough. This is reflected in the fluency score, which is relatively far from the text-to-text model, the original report, and ChatGPT.

The variation problem seems to have been minimized by the presence of a mini corpus in the model. The next problem is in the transition from one topic to another or from one sentence to another, which still needs to be improved in terms of flexibility. In addition, the impression of monotony still appears because of the repetition of words without having too much or too close meaning. For example, the word ”dengan” (with) in the sentence ”lebih baik lagi jika Ia membangun sinergi yang kuat dengan tim dengan menyesuaikan pendekatan komunikatif” (It would be even better if he builds strong synergy with the team by adjusting the communicative approach.).

Furthermore, the inference of data-to-text model was unable to complete the text generation until the end. In addition, the model tended to produce words that were repeated continuously and the connection of each word in the paragraph did not have an actual meaning. The use of this model still requires further optimization, especially in terms of providing sufficient data.

Data limitations are one of the issues in this approach. The approach does require large data as reported in previous research. The need for data cannot be solved by augmentation because the input data is tabular data, hindering to perform augmentation. In the future, research can be conducted using prompts in this model. In addition, the use of data that is not originally human-made can be an alternative although it is not ideal in research.

Finally, the text-to-text model has excellent results as observed in the qualitative assessments. This model was able to produce reports that have good conformity with input and coherence. However, it was found that its fluency is an aspect that needs further improvement. Several aspects that can be improved are the use of words ”mendorong Ia” (push him), which can be simplified to ”mendorongnya” (push him but using pronouns that are integrated with verbs in Indonesian). Additionally, the conjunction ”untuk itu” (therefore) is still considered inappropriate with the context of the two connected sentences. Then, contextually, the first sentence is considered to reflect the key behavior of the business cluster competency.

## 5. Conclusion And Future Work

Report generation using NLG models from tabular data that resemble human reports can be done using three approaches, namely template-based, data-to-text approach that converts tabular data into flat strings (linearization text), which then becomes model input, and text-to-text approach by utilizing template-based output as model input. The text-to-text approach in this study can produce reports that resemble humans since the reports made by the model have an assessment score by the assessor that exceeds the original human-made reports in two of the three categories, although the assessment score is still below the report score produced by ChatGPT. Following BLEU, ROUGE, BERTScore, and assessor evaluations, the text-to-text approach has a better score than the template-based and the data-to-text approach.

Future research can be done related to the mini corpus used in template-based as it requires enrichment and re-validation in order to produce more varied and better reports. Additionally, research with the data-to-text approach requires a lot of data that can be used as an alternative by using data that is not originally made by humans, although it is not ideal. Lastly, further research can also be carried out by conducting an in-depth exploration of the existing hyper-parameters both when training and inferring the model.

## 6. References

[1] D. E. Rupp et al., “Guidelines and Ethical Considerations for Assessment Center Operations,” J. Manage., vol. 41, no. 4, hal. 1244–1273, 2015, doi: 10.1177/0149206314567780.

[2] R. Dale, “Natural language generation: The commercial state of the art in 2020,” Nat. Lang. Eng., vol. 26, no. 4, hal. 481–487, 2020, doi: 10.1017/S135132492000025X.

[3] PASSTI, Pedoman dan Etika Pelaksanaan Assessment Center Indonesia. 2022. [Daring]. Tersedia pada: https://assessmentcenter-indonesia.org/web/images/PEDOMAN_ETIKA_PELAKSANAAN_AC_FF_22.pdf
[4] Y. Lin, T. Ruan, J. Liu, dan H. Wang, "A Survey on Neural Data-to-Text Generation," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 4, hal. 1431–1449, 2024, doi: 10.1109/TKDE.2023.3304385.
[5] E. Reiter dan R. Dale, "Building applied natural language generation systems," Nat. Lang. Eng., vol. 3, no. 1, hal. 57–87, 1997, doi: 10.1017/S1351324997001502.
[6] S. Indrayani dan M. L. Khodra, "Data-Driven News Generation for Indonesian Municipal Election," ICAICTA 2018 - 5th *Int. Conf. Adv. Informatics Concepts Theory Appl.*, hal. 153–158, 2018, doi: 10.1109/ICAICTA.2018.8541337.
[7] L. Leppänen, M. Munezero, M. Granroth-Wilding, dan H. Toivonen, "Data-driven news generation for automated journalism," INLG 2017 - 10th Int. Nat. Lang. Gener. Conf. Proc. Conf., hal. 188–197, 2017, doi: 10.18653/v1/w17-3528.
[8] J. Zhou dan S. Bhat, "Paraphrase Generation: A Survey of the State of the Art," EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc., hal. 5075–5086, 2021, doi: 10.18653/v1/2021.emnlp-main.414.
[9] S. Witteveen dan M. Andrews, "Paraphrasing with large language models," EMNLP-IJCNLP 2019 - Proc. 3rd Work. Neural Gener. Transl., hal. 215–220, 2019, doi: 10.18653/v1/d19-5623.
[10] R. Yermakov, N. Drago, dan A. Ziletti, "Biomedical Data-to-Text Generation via Fine-Tuning Transformers," INLG 2021 - 14th Int. Conf. Nat. Lang. Gener. Proc., no. January, hal. 364–370, 2021, doi: 10.18653/v1/2021.inlg-1.40.
[11] M. Kale dan A. Rastogi, "Text-to-Text Pre-Training for Data-to-Text Tasks," INLG 2020 - 13th Int. Conf. Nat. Lang. Gener. Proc., no. 2019, hal. 97–102, 2020, doi: 10.18653/v1/2020.inlg-1.14.
[12] K. Papineni, S. Roukos, T. Ward, dan W. J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2002, hal. 311–318.
[13] C. Y. Lin, "Rouge: A package for automatic evaluation of summaries," Proc. Work. text Summ. branches out (WAS 2004), no. 1, hal. 25–26, 2004, Diakses: 2 Juni 2024. [Daring]. Tersedia pada: papers2://publication/uuid/5DDA0BB8-E59F-44C1-88E6-2AD316DAEF85
[14] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, dan Y. Artzi, "Bertscore: Evaluating Text Generation With Bert," 8th Int. Conf. Learn. Represent. ICLR 2020, hal. 1–43, 2020.
[15] OpenAI et al., "GPT-4 Technical Report," vol. 4, hal. 1–100, 2023, [Daring]. Tersedia pada: http://arxiv.org/abs/2303.08774

**Alfan Aris Setiawan** is a professional with a background in Information Systems. As part of the data analyst division, He has a strong interest in data analysis and information technology. Currently, he is pursuing a Master's degree in Informatics Engineering with a specialization in Data Science and Artificial Intelligence to deepen his understanding and application of technology in assessment and data analysis.

**Masayu Leylia Khodra** received her Doctoral degree from the Institute of Technology Bandung in 2012. She is currently a Faculty Member of the School of Electrical Engineering and Informatics, Institute of Technology Bandung. Her research interests include summarization, information extraction, classification, clustering, data mining, opinion mining, and knowledge-based systems