Assessing Large Language Models for Zero-Shot Dynamic Question Generation and Automated Leadership Competency Assessment

I Gusti Bagus Yogiswara Ghheartha[1], Adiwijaya[2,*], Ade Romadhony[3], Yusfi Ardiansyah[4]

[1,2,3]*Telkom University, School of Computing, West Java, Bandung, 40257, Indonesia,*

[4]*Telkom Indonesia, Corporate Strategic Planning, Jakarta, 12710, Indonesia*

(Received: May 10, 2025; Revised: July 5, 2025; Accepted: October 25, 2025; Available online: December 1, 2025)

Abstract

Automated interview systems powered by artificial intelligence often rely on fine-tuned models and annotated datasets, limiting their adaptability to new leadership competency frameworks. Large language models have shown potential for generating questions and assessing answers, yet their zero-shot performance, operating without task-specific retraining remains underexplored in leadership assessment. This study examines the zero-shot capability of two models, Qwen 32B and GPT-4o-mini, within a multi-turn self-interview framework. Both models dynamically generated questions, interpreted responses, and assigned scores across ten leadership competencies. Professionals representing the role of Digital Marketing and Account Manager participated, each completing two AI-led interview sessions. Model outputs were evaluated by certified experts using a structured rubric across three dimensions: quality of behavioral insights, relevance of follow-up questions, and fit of assigned scores. Results indicate that Qwen 32B generated richer insights than GPT-4o-mini (mean = 2.86 vs. 2.62; p less than 0.01) and provided more differentiated assessments across competencies. GPT-4o-mini produced more consistent follow-up questions but lacked depth in interpretation, often yielding generic outputs. Both models struggled with accurate scoring of candidate responses, reflected in low answer score ratings (Qwen mean = 2.35; GPT mean = 2.21). These findings suggest a trade-off between insight richness and scoring stability, with both models demonstrating limited ability to fully capture nuanced leadership behaviors. This study offers one of the first empirical benchmarks of zero-shot model performance in leadership interviews. It underscores both the promise and current limitations of deploying such systems for scalable assessment. Future research should explore competency-specific prompt strategies, fairness evaluation across demographic groups, and domain-adapted fine-tuning to improve accuracy, reliability, and ethical alignment in high-stakes recruitment contexts.

Keywords: Large Language Models, Dynamic Question Generation, Automated Assessment, Zero-Shot Learning, Self-Interview, Competency-Based Scoring, Leadership Competency Assessment

1. Introduction

In the pursuit of achieving Indonesia's Golden Vision 2045, a national vision to become a sovereign, advanced, and sustainable nation, Indonesia must invest heavily in strengthening the quality and competitiveness of its human resources. One of the most strategic efforts toward this vision is the development of an effective talent mapping and selection process, especially for identifying leadership potential among candidates across sectors. However, the massive volume of applicants in both public and private sector recruitment processes has posed significant challenges to conventional approaches, which are often time-consuming, costly, and prone to subjectivity. At the same time, global trends in Human Resource (HR) technology indicate an accelerated adoption of digital tools, such as asynchronous video interviews and Artificial Intelligence (AI)-based assessments, to streamline recruitment and talent evaluation [1], [2]. Particularly, AI-driven systems that utilize Natural Language Processing (NLP) offer promising capabilities in automating question generation and answer evaluation in interview scenarios [3], [4].

Recent studies have explored the integration of AI and language models into recruitment and behavioral assessment, providing a foundation for zero-shot evaluation research. Black and van Esch [1] highlighted how AI-enabled recruiting has reshaped managerial practices in candidate evaluation, while Gonzalez et al. [5] examined applicant reactions toward AI/ML-based hiring processes, signaling the importance of trust and acceptance in automated assessments. In

***Corresponding author: Adiwijaya (adiwijaya@telkomuniversity.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i1.970

parallel, Garcia et al. [6] demonstrated that virtual interviews can significantly influence hiring decisions compared to traditional profile-based evaluation, underscoring the role of conversational data in candidate screening. At the methodological level, Otter et al. [3] and Serrano et al. [4] provided broader perspectives on the capabilities and limitations of deep learning and Large Language Models (LLMs), establishing the technical underpinnings of zero-shot evaluation. More directly related to competency and personality assessment, Zhang et al. [7] conducted one of the first comprehensive evaluations of LLMs in asynchronous video interviews, reporting on validity, reliability, and fairness, while Bounab et al. [8] applied NLP on interview transcriptions to estimate personality traits, demonstrating that behavioral evidence can be extracted from textual responses. Collectively, these studies illustrate that while LLM-based assessment has been investigated in recruitment and interview settings, a systematic benchmark for zero-shot leadership competency evaluation remains absent. Current AI-based interview systems still largely rely on annotated training data and fine-tuned models that are limited in their ability to capture the full complexity of human behavior and leadership competencies [5], [6]. Therefore, there is a lack of empirical studies evaluating zero-shot LLMs for dynamic question generation and answer scoring in leadership interviews, especially with structured expert validation in the context of Indonesian State-Owned Enterprises (SOEs) and similar high-volume, high-stakes environments. Most prior research focuses on personality trait inference or general employability prediction, rather than competency-based leadership assessment using zero-shot LLMs with human-in-the-loop validation [7], [8].

To address this gap, this study proposes a zero-shot dynamic question generation and answer assessment framework using LLMs. Unlike traditional supervised approaches, the proposed method allows the model to generate interview questions and evaluate answers without additional training data, relying solely on the inherent capabilities of pre-trained transformer-based models such as GPT and Qwen [9]. The evaluation framework is structured around 10 key leadership competencies, including Digital Leadership, Strategic Orientation, Customer Focus, and others, as aligned with organizational leadership models in Indonesian SOE environments. This study engages 12 candidates across Digital Marketing and Account Manager roles, segmented by experience level (Junior less than 3 years, Middle 3–6 years, Senior more than 7 years of experience). Each candidate participated in two AI-led interview sessions, each powered by a different LLMs (GPT-4o-Mini and Qwen 32B). To evaluate model performance in this setting, our study incorporates structured expert validation, whereby certified assessors reviewed model-generated outputs across evaluation dimensions insight quality, relevance of follow-up questions, and score fit. Insight Quality is defined as the accuracy and depth with which the model interprets candidate responses, measured through a five-point scale where 1 indicates completely irrelevant or inaccurate interpretation, 3 represents adequate understanding with some missing nuances, and 5 signifies comprehensive, contextually rich analysis that captures behavioral indicators relevant to leadership competencies. Follow-up Question Relevance evaluates the appropriateness and exploration potential of model-generated questions, with scores ranging from 1 (completely unrelated to candidate response or competency) to 5 (highly targeted, probing questions that effectively explore deeper behavioral evidence). Score Fit assesses the alignment between assigned competency scores and actual response content, where 1 indicates completely misaligned scoring, 3 represents reasonable but imperfect alignment, and 5 signifies precise score assignment that accurately reflects demonstrated competency level. This human-in-the-loop validation provides a grounded basis for evaluating the performance of LLMs in leadership-oriented interview tasks [7].

Qwen 32B consistently outperforms GPT-4o-mini in generating contextually rich insights (Insight Mean = 2.86 vs. 2.62; $t = 2.96$, $p = 0.004$, Cohen's $d = 0.39$), while also producing more differentiated assessments across competencies. However, Qwen's follow-up questions were more variable in relevance (SD = 0.83), and occasionally too general. In contrast, GPT-4o-mini achieved more consistent scores in the Next Question metric (SD = 0.76), but often failed to provide depth in behavioral interpretation and produced generic insights. Both models scored low in Answer Score evaluation (Qwen Mean = 2.35; GPT Mean = 2.21), indicating limited capability to accurately judge candidate responses against leadership criteria. These findings highlight the potential and limitations of zero-shot LLMs for scalable, automated leadership assessment. While LLMs offer adaptive and efficient solutions for talent evaluation, further refinement is needed to improve scoring accuracy and align model outputs with expert judgment. The study provides a foundation for developing more robust, human-in-the-loop AI assessment systems for leadership competencies.

2. Literature Review

2.1. AI-Based Self-Interview Systems

AI-based self-interview systems refer to automated platforms that allow candidates to respond to interview questions asynchronously, typically via video or text, with the support of artificial intelligence in processing, analyzing, and evaluating their responses. These systems aim to improve efficiency, objectivity, and scalability in talent acquisition, especially in high-volume recruitment scenarios [1], [2]. Self-interview platforms are often enhanced by NLP capabilities to generate dynamic questions, extract behavioral indicators, and assess candidate fit [4]. Compared to traditional face-to-face interviews, AI-based self-interviews provide greater flexibility and scalability, allowing candidates to participate from anywhere at any time. These systems can process thousands of candidates quickly while maintaining consistent and objective assessments [10], [11]. Furthermore, by leveraging optimized models, they can reduce human bias and promote fairness in evaluation [12], [13].

NLP plays a central role in extracting structured insights from unstructured candidate responses. Common NLP tasks in recruitment systems include sentiment analysis, keyword extraction, topic modeling, and behavioral scoring [3]. Transformer-based models have pushed the boundaries of NLP by enabling contextual understanding and dynamic response generation, which are critical for assessing competencies in interviews [14].

The adoption of AI-based interviews spans across various sectors, including large corporations managing high-volume recruitment, educational institutions for student assessments, and government agencies seeking operational efficiency. These systems are applied in use cases such as candidate selection, competency evaluation, employee rotation, promotions, and skill development through simulated interviews [10], [11], [15].

2.2. Zero-Shot Learning with Large Language Models (LLMs)

Zero-shot learning refers to the ability of a model to perform a task without any task-specific fine-tuning, using only general prompts and its pre-trained knowledge. In the context of interviews, zero-shot LLMs can be used to generate questions dynamically and evaluate answers based on the prompt instructions alone [9]. This approach reduces the need for costly labeled data and enhances adaptability to new domains or candidate types. LLMs are built upon the transformer architecture, which enables them to process and generate human-like text by capturing complex dependencies between words in a sequence. Introduced by Vaswani et al. [16], the transformer framework serves as the foundational backbone for state-of-the-art LLMs such as GPT, BERT, and T5. The Transformer architecture, introduced by Vaswani et al. [16], revolutionized NLP by relying entirely on attention mechanisms without using recurrence or convolution. This architecture is particularly relevant for AI-based self-interview systems, as it enables the parallel processing of candidate responses and the generation of follow-up questions through its multi-head attention mechanism. The core components of Transformer's encoder and decoder allow for efficient sequence modeling, with the encoder analyzing input sequences and the decoder generating contextually appropriate outputs, such as interview questions or feedback.

Transformer-based models can be classified into three types: encoder-only, decoder-only, and encoder-decoder. Encoder-only models like BERT [17] and IndoBERT [18] are optimized for understanding input text, making them ideal for assessing candidate responses during self-interviews. Decoder-only models such as GPT-3 [19] and GPT-4 [20] specialize in text generation, useful for crafting dynamic and personalized interview questions. These models can simulate interviewer behavior in asynchronous interactions, improving engagement and realism. GPT and Qwen are well-suited for AI-based self-interview systems due to their strong capabilities in natural language understanding and generation. GPT, particularly models like GPT-3 [19] and GPT-4 [20], excels in generating contextually relevant questions and evaluating candidate responses through semantic analysis. Its decoder-only architecture allows for dynamic follow-up questioning, making the interview process feel interactive and personalized. Studies have demonstrated that GPT-based models can reliably assess personality traits and performance in asynchronous interviews [7]. Qwen, developed by Alibaba DAMO Academy, offers a scalable and efficient alternative with competitive performance. Qwen are optimized for multilingual contexts, including Bahasa Indonesia, making them ideal for global or localized self-interview platforms. Qwen's open-source nature and support for fine-tuning enhance its adaptability to domain-specific needs such as behavioral assessment and competency evaluation [21].

Prompt design plays a central role in zero-shot learning, as both the question generation and answer assessment tasks rely on zero-shot prompting, without any task-specific fine-tuning. Constructed prompts are used to guide the model to generate behaviorally aligned questions, produce interpretive insights, and assign competency scores. Furniturewala et al. [22] explore how structured prompting can mitigate bias and improve fairness in LLM-generated evaluations, particularly in high-stakes decision-making scenarios such as hiring, university admissions, and criminal justice risk assessments. The authors propose a reflective two-step prompt structure where the model is first instructed to explain its reasoning before producing a judgment or score [22].

2.3. Question Generation

Question generation is a subfield of NLP concerned with automatically generating questions from text or contextual input. In the context of AI-based self-interviews, question generation is a key component for simulating interactive dialogues that can adaptively explore a candidate's competencies. Effective question generation models must generate questions that are not only grammatically fluent but also semantically aligned with behavioral goals such as leadership evaluation or problem-solving exploration. Transformer-based LLMs have enabled the transition from static, rule-based question generation to dynamic, context-aware generation in both zero-shot and few-shot settings. This advancement allows models like GPT or Qwen to formulate interview questions based on the candidate's previous answers, without requiring additional fine-tuning. Such zero-shot capabilities are particularly valuable in real-world applications where labeled datasets are limited and domain variation is high [19].

In educational and dialogic contexts, studies have shown that LLMs can generate diverse and relevant questions when guided by well-structured prompts [23]. It analyzed how instruction tuning improves the quality of generated questions by aligning model behavior with target educational outcomes [23]. Similarly, in the interview domain, EZInterviewer demonstrates how question generation can be tailored to specific roles or behavioral scenarios, enhancing the realism and challenge level of mock interview simulations [24].

Beyond basic generation, some systems adopt specialized strategies such as decoder fusion or reinforcement learning to improve question quality, diversity, and goal orientation. Handover QG [25], question generation by Decoder Fusion and Reinforcement Learning) proposes a mechanism for combining multiple decoding styles and optimizing question relevance through reward-based training. These advancements demonstrate how modern question generation frameworks move beyond surface-level syntax toward deeper semantic modeling, which is essential in high-stakes contexts like behavioral interviews.

This study adopts a zero-shot question generation approach using general-purpose LLMs to maintain high adaptability, minimal implementation cost, and rapid scalability. By leveraging prompt-based generation alone, the system can dynamically produce competency-aligned questions across a wide range of roles and industries without the need for annotated interview corpora or retraining. This strategy is especially relevant in contexts such as early-stage prototyping, limited-resource deployments, or exploratory studies where flexibility and generalization are prioritized over optimality. Furthermore, the zero-shot setup mirrors realistic recruitment scenarios in which interview topics evolve based on candidate input, reinforcing the need for generalizable, context-aware question generation.

2.4. Answer Assessment

Answer assessment in AI-based interview systems refers to the process by which candidate responses are automatically evaluated and scored, typically in relation to certain predefined behavioral or personality frameworks. The goal is to simulate or support human judgment in identifying relevant traits, competencies, or communication quality based on natural language input. Recent developments in LLMs have significantly expanded the potential of answer assessment, particularly by enabling models to generate context-sensitive interpretations and scoring without requiring task-specific training. In LLM-powered systems, answer assessment can be conducted either through direct scoring prompts, justification-based scoring, or via classification architectures trained on annotated interview data [7]. In the study by Zhang et al. [7], GPT-4 was used to infer personality traits (Big Five) from video transcriptions using carefully crafted prompts. The results were compared against human rater judgments to evaluate validity (R^2 correlation), reliability (test-retest and split-half), and fairness (across gender and ethnicity). This approach confirmed that LLMs are capable of producing assessments with comparable consistency to trained human evaluators, although some variations still exist due to prompt sensitivity and demographic biases.

Similarly, Gong et al. [8] demonstrated that transformer-based models could extract personality traits and screening labels from interview transcriptions. Their model employed multi-task learning to predict both trait scores and candidate suitability. This study underscores the growing feasibility of using LLMs to interpret open-ended candidate responses without structured input. In the domain of communication skill assessment, Thakkar et al. [26] conducted a comparative study showing that LLM-based classifiers could outperform traditional handcrafted feature-based models, especially when combined with domain adaptation techniques. This suggests that LLMs are more effective at capturing subtleties in candidate language and behavior that are often lost in rigid feature engineering pipelines. The rationale for using LLMs in this way is not to replace human evaluators entirely, but to reduce effort, standardize assessments, and make preliminary screening more scalable. As emphasized by Yadav et al. [27], LLMs can also be used to reflect and critique interviewer strategies, indicating that these models are capable of not only evaluating answers but also meta-assessing interactional dynamics in interviews.

2.5. Leadership Competency Assessment in Indonesia

The assessment of leadership competencies in Indonesia has undergone rapid transformation, aligning with both national developmental goals and the demands of digital-era human capital management. State-Owned Enterprises (SOEs) and private organizations have institutionalized competency frameworks that reflect a synthesis of global best practices and national core values, such as those prescribed by the Ministry of SOEs (AKHLAK: Trustworthy, Competent, Harmonious, Loyal, Adaptive, and Collaborative). These frameworks typically distinguish between strategic business leadership and organizational capabilities, including key dimensions such as Digital Leadership, Strategic Orientation, Customer Focus, Building Strategic Partnerships, Managing Diversity, and Driving Innovation.

Leadership competency assessment in Indonesia reflects distinct cultural and organizational characteristics that differ substantially from Western models. Indonesian managerial competency frameworks emphasize three core components: conceptual skills, human resources skills, and technical skills, with particular attention to hierarchical relationships and collective decision-making processes [28]. Effective competency models must account for local governance structures and cultural expectations, with successful implementations showing better alignment between organizational vision achievement and competency selection processes [28]. The Indonesian National Competency Standards (SKKNI) provide a regulatory framework that mandates participatory development processes through Competency Standards Committees, ensuring cultural relevance while maintaining technical rigor [29]. Cultural dimensions significantly influence leadership assessment validity in Indonesian contexts. Indonesia's high collectivism scores indicate that group harmony and interpersonal relationships take precedence over individual achievements, requiring assessment frameworks that capture collaborative leadership behaviors rather than individualistic competencies. The cultural preference for consensus-building and consultative decision-making styles presents unique challenges for automated assessment systems that may not recognize subtle cultural communication patterns [30]. Research indicates that leadership effectiveness in Indonesia correlates strongly with cultural alignment, suggesting that assessment tools must incorporate indigenous behavioral indicators to maintain validity.

Linguistic considerations present substantial challenges for LLM-based leadership assessment in Indonesian contexts. Indonesian (Bahasa Indonesia) exhibits significant syntactic and semantic differences from English, particularly in sentence structure, tense systems, and idiomatic expressions. Studies on Indonesian NLP development reveal that current language models achieve limited performance on complex linguistic tasks, with accuracy rates below 85% for sentiment analysis, named entity recognition, and dependency parsing. The prevalence of code-switching between Indonesian and regional languages in professional settings adds complexity to automated language processing, as leaders often communicate using mixed linguistic registers that reflect cultural nuances [31]. These regional and linguistic factors have direct implications for zero-shot LLM performance in leadership assessment. The limited availability of high-quality Indonesian language datasets for training [32], combined with cultural communication patterns that differ from Western norms, suggests that LLM models may exhibit reduced accuracy when evaluating Indonesian leadership responses.

2.6. Human in the Loop Evaluation

While LLMs demonstrate increasing capability in generating and assessing interview responses, their reliability and fairness remain subject to scrutiny, especially in high-stakes applications such as recruitment and leadership selection.

To address this, Human-In-The-Loop (HITL) evaluation serves as a crucial validation mechanism, enabling experts to verify, calibrate, and interpret LLM outputs before they are used for decision-making. In the work of Zhang et al. [7], human raters provided reference scores on Big Five personality traits from candidate videos, which were then used to validate the LLM-generated personality assessments. Their approach used metrics such as R^2 (coefficient of determination) to evaluate validity, split-half and test-retest reliability, and demographic fairness across gender and ethnicity.

Building upon this, the current study employs human-in-the-loop evaluation not by asking experts to assign new scores, but to judge the appropriateness of LLM-generated competency scores for each candidate's answer. This reduces annotation fatigue while preserving high judgment integrity. Experts rate the alignment between LLM outputs and actual response content using a Likert scale (1–5), making it possible to quantify agreement levels without requiring full rescoring. This annotation strategy is particularly suitable for zero-shot LLM scenarios, where the model is not fine-tuned but expected to generalize across prompts. Since such models can hallucinate or overgeneralize, expert judgment acts as a calibration layer, ensuring that behavioral interpretations and scoring remain contextually grounded and ethically appropriate [33].

Other studies, such as Bounab et al. [8], highlight that incorporating human oversight into model evaluation boosts stakeholder trust and facilitates the integration of AI systems into organizational decision processes. HITL evaluation bridges the gap between model capability and practical deployment, balancing automation with accountability. In this study, it ensures that the zero-shot, prompt-based LLM architecture remains usable and ethically aligned, particularly in assessing leadership potential across diverse candidates.

2.7. Research Gap

Most studies on LLM-based assessment focus on personality trait inference [7] or general employability prediction [8]. However, there is little to no exploration of how LLMs can be applied to behavioral leadership frameworks, such as the 10 Key Leadership Behaviors used in organizational leadership development. This leaves a gap in evaluating LLM effectiveness in high-level roles where strategic, collaborative, and adaptive leadership competencies are essential. While advanced models such as Handover QG [25] show strong results in generating questions for education or dialogue systems, they require task-specific training and complex architectures. Very few studies have tested zero-shot question generation and answer assessment in interview contexts, where domain-specific datasets are often unavailable. This study introduces a zero-shot approach using general-purpose LLMs to bridge this gap with higher flexibility and deploy ability.

Most research treats question generation and answer assessment as isolated tasks, focusing either on generating coherent questions or on scoring answer quality. For instance, EZInterviewer [24] emphasizes question generation but does not evaluate how generated questions affect downstream assessment [24], [25]. Although some works evaluate model output against human ratings, these are mostly limited to personality scoring or general fairness testing [7], [8], [12], [26]. There is a lack of structured expert validation protocols for scoring outputs of LLMs per leadership competency per question, particularly in high-context domains like digital leadership or strategic orientation. Table 1 summarize the gap of current research.

Table 1. State of the art from previous research

Paper	Focus	Method	QG	AA	Assessment Type	Human Eval
Zhang et al. [7]	Personality trait assessment from video interviews (data video)	Prompted GPT-4 scoring + human validation	No	Yes	Big Five traits	Yes (R^2, agreement)
Bounab et al. [8]	Personality traits and job screening from video transcripts	Transformer + multi-task learning	No	Yes	Suitability + traits	Yes (traits vs ground truth)
Thakkar et al. [26]	Communication skill scoring using deep learning	Deep learning classifier + domain adaptation	No	Yes	Communication score	No
Li et al. [24]	Mock interview generation with role-specific QG (EZInterviewer)	Pre-trained model + scenario-specific prompt	Yes	No	No scoring, focus on QG	No

Chung et al. [25]	Reinforced QG using decoder fusion (education, Handover QG)	Fusion decoder + RL training (non-zero-shot)	Yes	No	No scoring, focus on QG	No
Sun et al. [34]	Multi-role and multi-behavior (LLM as Interviewer and candidate) simulation using LLMs for recruitment	Role-specific multi-agent prompting + scripted scenario	Yes	Yes	Interviewer and Candidate score	Yes (human judgment)
Fang et al. [12]	NLP for classifying the content of interviews with cancer patients	Multi-label text classification based on BERT model	No	Yes	Classification (QoL, symptom, other)	Yes (human label)
Yadav et al. [27]	Automated feedback to the interviewer based on interview record	Classification using BERT (question) and T5 (event)	No	No	Interview metrics (speaking ratio, etc)	Yes (human judgement)
This Study	Leadership competency scoring with zero-shot QG and AA	Zero-shot LLM prompting + human-in-the-loop eval	Yes	Yes	10 leadership competencies	Yes (human judgement)

Note: QG = question generation, AA = Answer Assessment

3. Methodology

This study follows an experimental design that integrates LLMs into a self-interview framework for assessing leadership competencies. The overall research flow is illustrated in figure 1 and comprises seven main stages, ranging from problem formulation to expert evaluation and final reporting. This design allows for a structured evaluation of the LLM's performance in generating interview questions and scoring candidate responses using a zero-shot prompting approach.

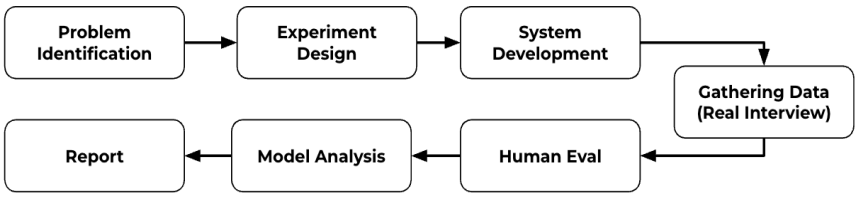

Figure 1. Research flow

3.1. Research Design

This study adopts an experimental research design to evaluate the performance of two LLMs such as GPT-4o-mini and Qwen 32B in handling AI-based self-interviews for leadership competency assessment. The research aims to address the increasing need for scalable and adaptive interview systems by analyzing how well LLMs can generate relevant interview questions and assess candidate responses in zero-shot settings.

The core tasks investigated are question generation and answer assessment, both executed without task-specific fine-tuning. The models are expected to dynamically generate behaviorally relevant follow-up questions based on candidate responses, provide short interpretive insights about those responses, and assign a score (1–5) to each of 10 predefined leadership competencies. To guide this research, three primary questions are formulated. RQ1: which LLMs is more suitable for handling AI-based interview use cases? RQ2: How does the LLM perform in generating questions and assessing candidate responses based on leadership competencies? RQ3: What are the appropriate evaluation metrics for validating model-generated scores, insights, and questions?

The objectives of this study are focused around evaluating the effectiveness of LLMs in the context of AI-based self-interviews. Specifically, the research aims to analyze and compare the quality of candidate evaluations generated by GPT-4o-mini and Qwen 32B, with a focus on how well each model supports dynamic leadership competency assessment. This includes examining the model's ability to generate relevant and explorative outputs, such as behavioral insights, follow-up questions, and competency scores, based on candidate responses. Furthermore, the study seeks to validate these outputs through expert-based score fit rating, providing a structured mechanism for measuring

how accurately the model-generated elements reflect the actual content and meaning of the candidate's answers. Together, these objectives support the broader goal of understanding how LLMs can be reliably integrated into scalable, adaptive talent evaluation systems.

The evaluation of this study relies exclusively on expert judgment, with all assessments conducted via subjective rating scales of output quality for each competency. The study does not employ manual labeling of ground truth competency detection, nor does it compare model predictions to predefined correct answers. As a result, the analysis is limited to descriptive and inferential statistics derived from expert ratings, rather than detection-based performance metrics. All reported findings should be interpreted as reflecting perceived quality and competency relevance according to expert review, rather than as measures of classification accuracy or detection correctness.

3.2. System Development

This subsection describes how the self-interview system was developed to support the automatic evaluation of candidates using LLMs. The system is designed to generate interview questions, assess candidate responses against predefined leadership competencies, and manage the interview flow dynamically. The architecture follows a modular design, utilizing webhook-based communication and API calls to two LLMs namely GPT-4o-mini and Qwen 32B as explained in figure 2.

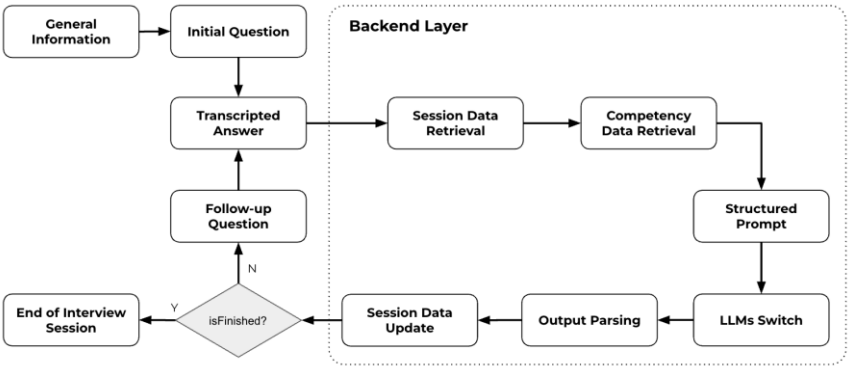

Figure 2. System flow

The system flow is divided into two main layers, there are interview interaction layer and backend layer. The process begins when a candidate submits their basic profile information, after which the system delivers an initial question based on the targeted competency. Candidate responses are then transcribed and forwarded to the backend. On the backend, the system triggers a structured prompt call to the selected LLM through a switch mechanism that routes to Model 1 QWEN 32B) or Model 2 (GPT-4o-mini). The model returns three outputs, there are behavioral insight derived from the answer, a follow-up question, and a set of competency scores. These outputs are parsed and stored in the database through an HTTP request. The system evaluates a conditional flag to determine whether the interview session should continue. If the 10 competencies have not been detected from candidate's answer, the system presents the follow-up question and loops back to collect the next answer. When the 10 competencies have been detected and scored, it ends the interaction and saves the full interview data for expert-based evaluation in the validation phase. The system terminates the interview when all 10 competencies achieve a minimum detection threshold of score ≥ 1 on the 5-point scale. Competency detection is determined by the LLM's explicit scoring output for each dimension, where scores of 1 indicate insufficient evidence and scores of 2-5 represent progressive levels of competency demonstration.

3.3. Model Setup

This study utilizes two state-of-the-art LLMs, GPT-4o-mini and Qwen 32B, as the backbone of the AI-based self-interview system. Both models are accessed in a zero-shot setting, meaning that no fine-tuning or task-specific retraining is applied. Instead, prompt engineering is used to elicit task-specific outputs such as follow-up questions, behavioral insights, and competency scores. The two models were chosen to provide contrast across different deployment and performance categories. GPT-4o-mini represents a commercial, optimized, inference-efficient model widely used in industry. GPT-4o-mini is a compact version of OpenAI's GPT-4 model, optimized for faster inference and lower latency. While its exact parameter count is undisclosed, it is generally considered a lightweight variant

 ISSN 2723-6471

suitable for real-time dialogue applications. While Qwen 32B represents a powerful open-source alternative with greater model capacity and transparency. By comparing these two models under the same task and prompt structure, the study aims to explore the trade-off between generation quality, evaluation accuracy, and practical deployability in the context of AI-based self-interviews. The detailed differences between these 2 models are explained in table 2.

GPT-4o-mini is OpenAI's optimized variant of the GPT-4 architecture, specifically designed for efficient inference and reduced computational overhead. While OpenAI does not disclose exact parameter counts, GPT-4o-mini is estimated to contain approximately 8-20 billion parameters based on its performance characteristics and computational requirements. The model is trained on diverse internet text, books, and curated datasets with a focus on instruction-following and conversational applications. Its training emphasizes safety alignment, multilingual capabilities, and reduced hallucination rates compared to earlier GPT variants. GPT-4o-mini represents the commercial, closed-source approach to LLM deployment, offering enterprise-grade reliability and consistent API access. Qwen 32B is Alibaba DAMO Academy's open-source large language model containing 32 billion parameters, built on a decoder-only transformer architecture. The model is trained on a comprehensive dataset including web documents, books, academic papers, and multilingual corpora with particular strength in Chinese and English languages. Qwen 32B incorporates advanced training techniques including instruction tuning, Reinforcement Learning From Human Feedback (RLHF), and careful data curation to enhance reasoning capabilities. Its open-source nature provides full transparency regarding architecture, training methodology, and model weights, enabling fine-tuning and customization for domain-specific applications.

Table 2. GPT-4o-mini and Qwen 32B comparison

Specification	GPT-4o-mini	Qwen 32B
Architecture	Decoder-only Transformer	Decoder-only LLM Transformer
Provider	OpenAI	Alibaba Cloud
Deployment Method	API (cloud-based)	Open-source, self-hosted (Hugging Face or local)
Language Support	Multilingual (with emphasis on English)	Multilingual (strong support for Chinese and English)
Data Source	Web data, books, Wikipedia, code, dialog datasets	Books, web documents, Wikipedia, academic

3.4. Prompt Design

Prompt design plays a central role in this study, as both the question generation and answer assessment tasks rely on zero-shot prompting, without any task-specific fine-tuning. Constructed prompts are used to guide the model to generate behaviorally aligned questions, produce interpretive insights, and assign competency scores. In alignment with structured prompt design principles recommended in recent LLM research [22], the prompts in this study are designed to elicit not only output but also reasoned, interpretable evaluations. The prompt used in the answer assessment task begins by instructing the model to assume the role of an AI interview evaluator, followed by a clear job-specific context and a list of the 10 leadership competencies to be considered.

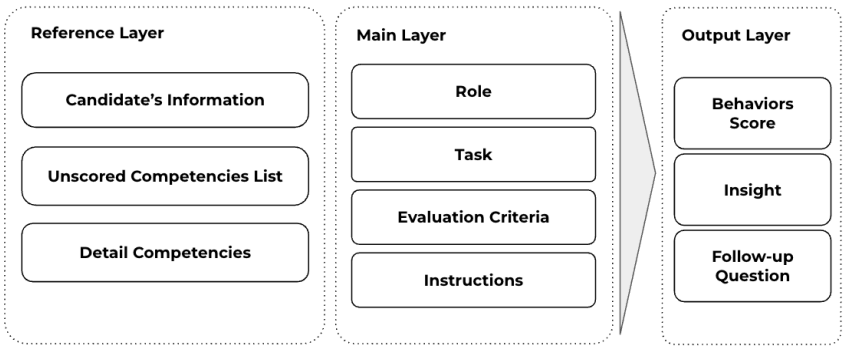

Figure 3. Prompt Structure

Figure 3 shows the prompt structure used in this study. It organized into three logical layers, namely Reference Layer, Main Layer, and Output Layer to guide the LLMs evaluation process in a clear and controlled manner. The Reference Layer provides contextual inputs such as the candidate's profile and a list of leadership competencies to frame the assessment. The Main Layer defines the model's role as an AI interview evaluator, outlines the task such as tailoring assessment to the candidate's position, and specifies evaluation criteria such as score scale, language, and tone. The final Output Layer governs the expected response format, requiring the model to produce structured outputs that include behavioral insights, competency scores, and follow-up question. This layered structure ensures interpretability, consistency, and alignment with competency-based evaluation standards. An example of a Structured Prompt is available in APPENDIX A.

The output generated by the LLMs in this study is structured to reflect the dual-task objective of the system, there are interpreting candidate responses, scoring leadership competencies, and generate follow-up question. Each LLM whether GPT-4o-mini or Qwen 32B receives the same structured prompt and is expected to produce three main components per candidate answer as shown in figure 3.

3.5. Tools and Platform

This study employs a modular system architecture supported by a combination of open-source tools and modern cloud platforms to enable the development, deployment, and evaluation of the AI-based self-interview system. The frontend is developed using Svelte, along with HTML, CSS, and JavaScript, and is hosted through Vercel for efficient deployment and continuous integration. Development activities are facilitated using StackBlitz as a browser-based IDE, with version control maintained via GitHub. On the backend, the system utilizes n8n, a low-code workflow automation platform that handles API integration, decision logic, and LLMs invocation through structured webhooks. For data storage we use Supabase, an open-source backend built on PostgreSQL to manage candidate records, interview transcripts, and model-generated outputs. The entire infrastructure is secured through Cloudflare, ensuring encrypted traffic, performance optimization, and basic threat protection. Core services such as n8n and containerized deployments are hosted using Docker on a self-managed Ubuntu VPS, while orchestration and environment consistency are handled via Coolify, an open-source DevOps platform. This technology stack provides the flexibility, scalability, and reliability required to support iterative experimentation with LLMs in a production-ready interview environment. Figure 4 details the tech stack used for this research.

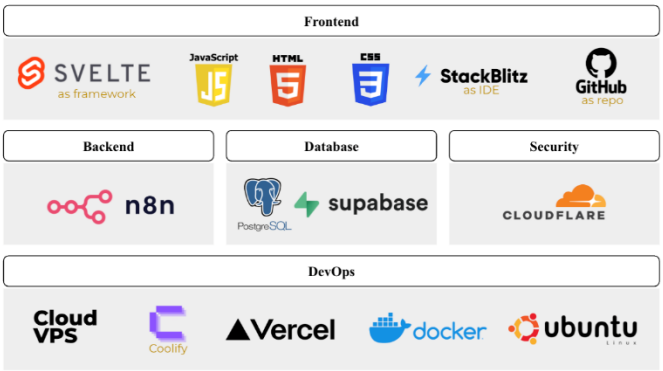

Figure 4. Tech stack

3.6. Dataset and Participants

To maintain realism and domain relevance, interviews are conducted with 12 candidates, 6 from the Account Manager role and 6 from Digital Marketing across Junior (less than 3 years of experience), Middle (3–6 years), and Senior (more than 7 years) experience levels. Each candidate responds to structured prompts within the self-interview system powered by both LLMs, generating a dataset of model outputs for expert evaluation. Each participant completed two structured self-interview sessions in sequential order: all participants first completed the interview with Qwen 32B (Model 1), followed by GPT-4o-mini (Model 2). This fixed-order, within-subjects design resulted in 24 total model sessions (12 per model). The sequential design was chosen to maintain consistency in the experimental protocol, though it introduces potential order effects that are acknowledged as a limitation. Each session operated as a standalone,

independent interview with no cross-session context sharing. The models had no access to previous session data, candidate responses, or competency scores from the first interview. Each LLM started with identical baseline information (candidate profile and role specifications) and generated questions, insights, and assessments independently based solely on real-time candidate responses within that specific session. This design ensures that model performance comparisons reflect genuine differences in zero-shot capabilities rather than contextual advantages from previous interactions. Participants were explicitly instructed to treat each session as a separate, independent interview experience without referencing or building upon their previous session responses. During each session, participants responded to a series of interview questions aligned with 10 predefined leadership competencies, such as Digital Leadership, Global Business Savvy, Customer Focus, Building Strategic Partnership, Strategic Orientation, Driving Execution, Driving Innovation, Developing Organizational Capabilities, Leading Change, and Managing Diversity. The models then generated an insight, a follow-up question, and a competency score for each response.

The dataset consists of all model-generated outputs along with human expert evaluations, covering candidate identifiers, interview timestamps, LLM used (Model 1 or Model 2), and system outputs (insight, next question, and 10 competency scores). This structured and annotated dataset forms the basis for both comparative model evaluation and statistical analysis of model–human alignment.

3.7. Expert Evaluation

To ensure the quality, fairness, and reliability of model-generated outputs, this study employed a human-in-the-loop evaluation process involving expert raters from Assessment Center Indonesia (ACI), a nationally recognized authority in behavioral assessment and talent development. The selected experts possess extensive experience in competency-based evaluation and are certified in leadership assessment practices. A panel of expert raters assesses the quality and appropriateness of model-generated outputs across three dimensions insight quality, follow-up question, and score fit rating. All expert evaluations are conducted using a 5-point Likert scale, which allows for the quantitative comparison of both LLMs. The research design therefore not only compares model performance (GPT-4o-mini vs Qwen 32B) but also establishes a metric-based protocol for validating AI-assisted interviews, filling a current gap in AI-for-HR literature.

The expert evaluation output in this study is recorded in a structured format, using a 5-point Likert scale for each of the three assessment dimensions mentioned before. For every candidate response processed by each LLM (GPT-4o-mini and Qwen 32B), the experts rated how accurately the model-generated insight reflected the content of the answer, how contextually appropriate and behaviorally relevant the follow-up question was, and how well the model's assigned competency scores matched the actual substance of the candidate's response. Each evaluated entry includes metadata such as candidate ID, role, experience level, the original interview question, the candidate's response, the model's output (insight, question, and scores), and the expert's three ratings. This evaluation output serves as the ground truth for analyzing model performance in terms of interpretability, contextual alignment, and behavioral scoring accuracy.

4. Results and Discussion

This chapter presents the analysis and discussion of experimental results obtained from the expert evaluation of model-generated outputs. The focus lies in assessing the performance of two LLMs (GPT-4o-mini and Qwen 32B) in producing structured interview insights, follow-up questions, and leadership competency scores based on candidate responses. Using expert-validated fit ratings as reference, the results are analyzed across multiple dimensions, including model comparison, competency-level accuracy, and alignment between model behavior and human expectations. This discussion aims to interpret both quantitative trends and qualitative observations to derive implications for the development of scalable and fair AI-based interview systems.

4.1. Overview of Evaluation Results

The evaluation of the Expert, system involved comprehensive analysis of two LLMs Qwen 32B (as Model 1) and GPT-4o-mini (as Model 2). The total number of expert evaluation rows was 288, comprising 202 evaluations for Qwen 32B and 86 for GPT-4o-mini. This distribution imbalance is a consequence of the model's differing approaches to competency detection during each interview session. Specifically, GPT-4o-mini often identifies and scores multiple competencies in response to a single candidate answer, thus requiring fewer question–answer iterations to cover all

target competencies within a session. In contrast, Qwen 32B typically detects only one or occasionally two competencies per candidate response, resulting in a greater number of iterations and correspondingly more evaluation rows per session to achieve complete coverage of all ten competencies. As a result, sessions led by Qwen 32B naturally produced more discrete evaluation units for expert review compared to those involving GPT-4o-mini. The higher number of evaluation rows for Qwen 32B reflects a more granular, step-wise approach rather than a greater exposure to candidate answers, and may influence the variance or robustness of aggregate scoring indicators. The overall evaluation revealed significant performance differences between the models across all measured competencies, with competency-based assessment requiring rigorous evaluation instruments that define sub-competencies and indicators with clarity.

Table 3. Results overview

Metric	N	Mean	Stdev	Min	Max
Insight	288	2.78	0.63	1	5
Next Question	275	2.97	0.82	1	4
Answer Score	275	2.31	0.59	1	4

As explained in table 3, the scoring patterns revealed distinct distributions across the three evaluation metrics. For Insight Fit, scores ranged from 1 to 5 points, with the most frequent score being 3 (64.9% of cases), followed by 2 (24.3%) and 4 (7.6%). Only one instance achieved the maximum score of 5 (0.4%). Next Question scores showed a similar pattern, with score 3 being most common (54.9%), followed by score 4 (24.7%) and score 2 (13.1%). Answer Score demonstrated the most restrictive distribution, with the majority of evaluations receiving score 2 (63.6%) and only 29.8% achieving score 3, indicating generally low scoring in this dimension. The mean scores across all evaluations were Insight Fit, Next Question, and Answer Score, suggesting moderate performance levels with Answer Score showing the lowest average and least variability. The insight metric, with 288 observations, has a mean score of 2.78 (on a scale of 1 to 5), indicating a moderate level of insightful responses overall, with a relatively low standard deviation (0.63) suggesting consistency in the evaluation. The next question and answer score metrics, both with 275 observations, have mean scores of 2.97 and 2.31 respectively, showing that the ability to generate relevant follow-up questions is slightly higher than the quality of the answer score, though both metrics also exhibit moderate consistency (standard deviations of 0.82 and 0.59). The minimum and maximum values for all metrics indicate that there is a range in the quality and depth of responses, with some responses scoring the lowest possible value and others reaching the maximum. This spread suggests variability in performance, but the relatively low standard deviations imply that most responses cluster around the mean. Overall, the results highlight that while there is room for improvement, particularly in answer quality, the system demonstrates a fair level of insight generation and follow-up question formulation.

4.2. Performance Comparison Between Model

Based on table 4 and figure 5, Model 1 (Qwen 32B) demonstrated superior performance across most evaluation metrics compared to Model 2 (GPT-4o-mini). For the Insight metric, Qwen 32B achieves a higher mean score (2.86) compared to GPT-4o-mini (2.62), with similar standard deviations (0.62 and 0.64, respectively). The t-test yields a value of 2.96 with a p-value of 0.004, indicating a statistically significant difference in favor of Qwen 32B. The Cohen's d of 0.39 suggests a small to moderate effect size, meaning that while the difference is statistically significant, the practical impact is moderate. This indicating a large effect size and robust superiority of Model 1 in generating contextually appropriate insights. Qwen 32B not only exhibits higher mean performance but also maintains consistency, as indicated by its comparable standard deviation with GPT-4o-mini. This consistency implies that Qwen 32B's ability to generate insightful content is robust and reliable across various contexts. The statistically significant difference (p-value = 0.004) further emphasizes this advantage, reinforcing the model's suitability for tasks requiring deeper cognitive engagement and nuanced understanding.

In the Next Question metric, Qwen 32B again outperforms GPT-4o-mini, with mean scores of 3.04 and 2.81, respectively. The standard deviations are 0.83 for Qwen 32B and 0.76 for GPT-4o-mini. The t-value of 2.14 and p-value of 0.034 indicate a statistically significant difference, though the effect size (Cohen's d = 0.27) is small. This suggests that Qwen 32B is somewhat better at generating relevant follow-up questions, but the margin is not large. Although Qwen 32B continues to surpass GPT-4o-mini, the reduced Cohen's d suggests that the practical implications

 ISSN 2723-6471

of this performance gap are narrower. This implies that experts generally perceived Qwen 32B as more effective in formulating relevant follow-up questions compared to GPT-4o-mini. However, the modest difference in scores indicates that both models were perceived similarly competent by the experts, suggesting only slight qualitative differences in their capability to generate follow-up questions.

Table 4. Results of Statistical Tests Between Models

Metric	Model 1		Model 2		t	p-value	Cohen's d
	Mean	Stdev	Mean	Stdev			
Insight	2.86	0.62	2.62	0.64	2.96	0.004	0.39
Next Question	3.04	0.83	2.81	0.76	2.14	0.034	0.27
Answer Score	2.35	0.63	2.21	0.46	2.14	0.034	0.25

For the Answer Score metric, Qwen 32B achieves a mean of 2.35, slightly higher than GPT-4o-mini's 2.21, with standard deviations of 0.63 and 0.46, respectively. The t-value and p-value (2.14 and 0.034) again show statistical significance, but with the smallest effect size among the three metrics (Cohen's d = 0.25). This modest difference underscores that while Qwen 32B delivers slightly superior answer quality, the practical distinction between the two models might be negligible in typical use cases. The notably smaller standard deviation for GPT-4o-mini (0.46) compared to Qwen 32B (0.63) indicates greater agreement among experts regarding GPT-4o-mini's answer quality. This suggests that expert assessments of GPT-4o-mini's answers were relatively uniform, even though the overall perceived quality was slightly lower compared to Qwen 32B.

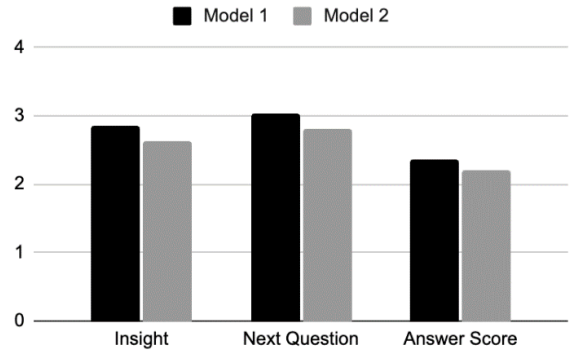

Figure 5. Mean scores by model

Collectively, these expert evaluations provide meaningful insights into perceived qualitative differences between Qwen 32B and GPT-4o-mini. Experts consistently rated Qwen 32B slightly higher across all three evaluation metrics including Insight, Next Question, and Answer Score. This indicating its perceived strength in delivering contextually richer and more relevant outputs. However, GPT-4o-mini's lower variability in the Answer Score metric suggests that it maintains a stable performance level, even if slightly lower overall. Therefore, the differences highlighted by expert assessments reflect nuanced perceptions of qualitative output, guiding practitioners in selecting a model based on the desired depth and consistency of conversational interactions rather than strictly technical performance.

Table 5. Comparison of Model Scores by Role Type and Experience Level

Model	Role	Experience	Avg. insight	Avg. Next Question	Avg. Answer Score
Model 1	Account Manager	Junior	2.97	3.16	2.40
		Middle	2.78	2.94	2.29
		Senior	2.75	3.29	2.36
	Digital Marketing	Junior	2.86	2.89	2.43
		Middle	3.03	2.79	2.21
		Senior	2.75	3.11	2.47
Model 2	Account Manager	Junior	2.58	3.00	2.33
		Middle	2.67	2.64	2.33

	Senior	2.58	3.00	2.17
Digital Marketing	Junior	2.92	2.45	2.42
	Middle	2.50	3.00	1.95
	Senior	2.53	2.71	2.20

Based on table 5, the average expert assessments show variations in model performance across each combination of roles and experience levels. In general, Model 1 consistently produced slightly higher Insight and Next Question scores than Model 2 for almost all role and level combinations, with the best scores appearing at the Middle Digital Marketing level (average insight 3.03) and Senior Account Manager (average next question 3.29). On the other hand, Model 2 performed more consistently in several aspects, such as the Answer Score for Junior Digital Marketing (2.42), although its overall average remained lower than Model 1, particularly in the Middle Digital Marketing category, which only received a 1.95 Answer Score.

For all roles, both Account Manager and Digital Marketing, there was no pattern that experience consistently correlated positively with improvements in all metrics. For example, Insight and Next Question scores for Senior candidates sometimes decreased compared to Middle or Junior candidates. This can be explained in two ways. First, LLM in a zero-shot setting relies heavily on the prompt and candidate response patterns. When Senior candidates provide more generic or normative responses, which often occur due to high experience, the model tends to fail to uncover unique insights or rate the answers more highly. Second, the model's tendency to provide relevant follow-up questions (Next Questions) is more stable across roles/levels, likely influenced by data peculiarities or prompt framing that are more easily explored in middle roles and experience. Overall, these results underscore the need for more adaptive LLM assessment strategies to adapt to variations in candidate behavior at different levels and roles, as well as the importance of expert assessment rubrics to control model bias toward responses that appear "safe" or normative.

4.3. Competency-Level Evaluation

The competency-level analysis reveals significant patterns in how both models detect and assess leadership competencies, providing insights into their strengths and limitations across the ten evaluated dimensions. This analysis examines competency detection frequency, scoring accuracy, and the relationship between competency assessment and expert evaluation scores. The evaluation draws on expert annotations for both Qwen 32B (Model 1) and GPT-4o-mini (Model 2) across the ten targeted leadership competencies Digital Leadership (DL), Global Business Savvy (GBS), Customer Focus (CF), Building Strategic Partnership (BSP), Strategic Orientation (SO), Driving Execution (DE), Driving Innovation (DI), Developing Organizational Capabilities (DOC), Leading Change (LC), and Managing Diversity (MD).

Both models demonstrate varying capabilities in competency recognition, with Qwen 32B showing more conservative detection patterns compared to GPT-4o-mini's broader competency identification approach. This difference in detection sensitivity has important implications for comprehensive leadership assessment, as conservative models may miss subtle demonstrations of certain competencies, while more inclusive models may attribute competencies in cases where evidence is less clear. Qwen 32B displays particular strengths in specific competency domains and limitations in others. The conservative approach may result in fewer identified competencies overall but with greater confidence in each detection. This balance in detection sensitivity is particularly significant in contexts where thorough competency coverage is essential for valid leadership evaluation. GPT-4o-mini exhibits different characteristics, with broader patterns of competency identification that may capture a wider range of expressions but also include cases where competency relevance is uncertain. The models also differ in their approaches to scoring, as reflected by variations in how numerical values are assigned to detected competencies, which in turn influences overall assessment reliability.

Table 6 presents a detailed breakdown of the average scores and standard deviations for both Qwen 32B (Model 1) and GPT-4o-mini (Model 2) across ten leadership competencies, evaluated on three dimensions (Insight, Next Question, and Answer Score). When competencies were detected, both models generally assigned high scores, with most detected competencies receiving scores of 4.0 on the 5-point scale. For Model 1 (Qwen 32B), descriptive averages suggest that competencies such as BSP and Strategic Orientation (SO) tended to score higher across dimensions, particularly in Insight and Next Question categories (mean range 3.05–3.26). In contrast, competencies such as DOC and Global

 ISSN 2723-6471

Business Savvy (GBS) appeared lower (mean range 2.63–2.95), reflecting areas where Model 1 outputs were less aligned with expert expectations. For Model 2 (GPT-4o-mini), similar descriptive patterns were observed, although with slightly lower overall averages. BSP again showed relatively higher scores in generating follow-up questions (mean = 3.21), whereas competencies such as Managing Diversity (MD) and DOC consistently appeared at the lower end (mean range 2.07–2.47). However, one-way ANOVA tests indicated that these competency-level differences were not statistically significant for either model (Model 1: $F(9,2550)=0.139$, $p=0.999$; Model 2: $F(9,1070)=0.057$, $p=0.999$). This suggests that although descriptive variations exist, the models evaluated competencies in a statistically uniform manner. Therefore, competency-level differences should be interpreted as observational tendencies rather than systematic effects.

Comparing descriptive accuracy across key competencies reveals that both models are reasonably accurate in evaluating competencies such as Customer Focus (CF) and Execution (represented as DE), with closely matched average scores and standard deviations, indicating general expert consensus on model performance in these areas. However, competencies that involve subtler or more context-dependent behaviors, such as DL and Managing Diversity (MD), present greater variability, suggesting potential challenges for both models in consistently interpreting complex competency demonstrations. Furthermore, examining correlations between competency scores and the overall expert Score Fit highlights interesting relationships. Competencies with consistently higher scores (e.g., Building Strategic Partnership and Strategic Orientation) generally correlated positively with expert perceptions of overall accuracy and relevance (Score Fit). Conversely, competencies with lower scores and higher variability (such as DOC and MD) showed weaker correlations with expert Score Fit ratings, indicating that models struggle more significantly to provide accurate and relevant evaluations in these areas.

Table 6. Evaluation based on competencies

Competency	Model 1						Model 2					
	Insight		Next Question		Answer Score		Insight		Next Question		Answer Score	
	Avg	Stdev	Avg	Stdev	Avg	Stdev	Avg	Stdev	Avg	Stdev	Avg	Stdev
DL	2.84	0.50	3.21	0.71	2.50	0.62	2.58	0.69	3.05	0.71	2.21	0.42
GBS	2.79	0.54	3.13	0.96	2.39	0.61	2.63	0.60	2.72	0.75	2.16	0.37
CF	2.79	0.42	3.37	0.68	2.63	0.60	2.63	0.60	3.05	0.52	2.16	0.37
BSP	3.05	0.52	3.26	0.56	2.63	0.68	2.63	0.68	3.21	0.63	2.32	0.58
SO	3.05	0.52	2.95	0.78	2.26	0.45	2.74	0.65	2.78	0.65	2.26	0.45
DE	2.79	0.42	3.26	0.65	2.53	0.51	2.58	0.61	2.95	0.52	2.16	0.37
DI	3.00	0.67	3.00	0.87	2.33	0.59	2.68	0.75	3.05	0.62	2.21	0.42
DOC	2.63	0.60	2.95	0.71	2.32	0.58	2.47	0.61	2.50	0.82	2.21	0.42
LC	2.89	0.32	3.11	0.66	2.47	0.51	2.63	0.60	2.84	0.60	2.21	0.42
MD	2.95	0.52	2.71	0.85	2.39	0.50	2.47	0.51	2.07	0.92	2.11	0.46

Other relevant observations include the variability of expert scores as represented by standard deviations. High variability in competencies like Driving Innovation (DI) and Global Business Savvy (GBS) for Next Question evaluations suggest inconsistencies in model-generated follow-up questions, pointing to potential opportunities for targeted model improvements or further fine-tuning to ensure consistent performance across diverse competency contexts. In conclusion, this competency-level evaluation reveals specific strengths and weaknesses of both models in assessing leadership competencies. It underscores the importance of competency-specific considerations in model selection, training, and refinement to enhance reliability and accuracy in leadership assessment contexts.

Table 7. Detailed table of distribution per competency

Competency	*Model 1 (Score Count)*														
	Insight					Next Question					Answer Score				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
DL	0	4	14	1	0	0	3	9	7	0	0	10	7	1	0
GBS	0	5	13	1	0	2	0	8	6	0	0	12	5	1	0
CF	0	4	15	0	0	0	2	8	9	0	0	8	10	1	0
BSP	0	2	14	3	0	0	1	12	6	0	0	9	8	2	0

 ISSN 2723-6471

SO	0	2	14	3	0	0	6	8	5	0	0	14	5	0	0
DE	0	4	15	0	0	0	2	10	7	0	0	9	10	0	0
DI	0	4	11	4	0	2	0	11	4	0	0	13	4	1	0
DOC	1	5	13	0	0	1	2	13	3	0	1	11	7	0	0
LC	0	2	17	0	0	0	3	11	5	0	0	10	9	0	0
MD	0	3	14	2	0	2	3	10	2	0	0	11	7	0	0

Model 2 (Score Count)

Competency	Insight					Next Question					Answer Score				
	1	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**5**
DL	0	10	7	2	0	0	4	10	5	0	0	15	4	0	0
GBS	0	8	10	1	0	1	5	10	2	0	0	16	3	0	0
CF	0	8	10	1	0	0	2	14	3	0	0	16	3	0	0
BSP	0	9	8	2	0	0	2	11	6	0	0	14	4	1	0
SO	0	7	10	2	0	1	3	13	1	0	0	14	5	0	0
DE	0	9	9	1	0	0	3	14	2	0	0	16	3	0	0
DI	0	9	7	3	0	0	3	12	4	0	0	15	4	0	0
DOC	0	11	7	1	0	3	2	11	0	0	0	15	4	0	0
LC	0	8	10	1	0	0	5	12	2	0	0	15	4	0	0
MD	0	10	9	0	0	5	3	6	0	0	1	15	3	0	0

The distribution tables presented in table 7 provide a clear overview of how expert ratings are spread across each competency, for every assessed category (insight, next question, answer score), and for both Model 1 and Model 2. Each cell in the table shows the frequency of a specific score (ranging from 1 to 5) assigned to a given competency and evaluation category. Most ratings fall within scores 2 and 3 for both models, indicating that expert judgments generally position the model outputs around a "satisfactory" or "good" level, with very few instances of "excellent" (score 5) or "poor" (score 1). The distribution patterns for Model 2 mirror those of Model 1, though Model 2 tends to receive slightly more scores of "2" for both insight and answer score, reflecting marginally lower performance according to expert assessment.

Some competencies, such as LC and MD, show a concentration of ratings at score 3, suggesting that the model's responses to these competencies were regarded as consistent and average by the experts. In contrast, competencies like SO, DE, and BSP obtained more score "4" ratings in Model 1 than in Model 2, which may indicate a performance drop or less accurate interpretation of these competencies in Model 2.

4.4. Inter-Rater Reliability of Expert Assessment

Inter-Rater Reliability (IRR) is a key indicator for evaluating the degree of consistency among expert's assessments using a given instrument. In this study, IRR was measured using the Intraclass Correlation Coefficient (ICC), a coefficient widely recommended for multi-rater Likert scale and expert panel evaluations. The ICC captures the level of agreement among raters, both for individual and average ratings across all judges [35]. The ICC analysis was conducted using several models, but the most relevant for practical applications with multiple expert panels are ICC1k (average raters absolute agreement), ICC2k (average random raters), and ICC3k (average fixed raters). These indices reflect the reliability of using the average score from all raters, under both random and fixed rater assumptions.

Table 8. Average-rater (k) ICC values for the assessment instrument

ICC Type	ICC	95% Confidence Interval	Interpretation
ICC1k	0.62	0.41 - 0.78	Moderate - Good
ICC2k	0.65	0.44 - 0.80	Moderate - Good
ICC3k	0.71	0.54 - 0.83	Good

Interpretation of table 8 demonstrates that inter-rater reliability for the mean scores of all experts is in the moderate to good range (ICC1k = 0.62, ICC2k = 0.65, ICC3k = 0.71). According to the categorization, ICC values above 0.7 are generally considered to indicate good reliability. This level of consistency indicates that the evaluation process is sufficiently objective and trustworthy, so that the aggregated mean scores of the expert panel are appropriate for further evaluation or decision-making. These findings also suggest that the effect of individual rater subjectivity on aggregate scores is relatively minor, as variation is minimized by averaging across the panel. Accordingly, the level of inter-rater

reliability observed in this instrument meets the standard benchmarks recommended in the literature for multi-rater assessment.

4.5. Model Strength and Weakness

The qualitative analysis from expert evaluations highlights several critical strengths and weaknesses for both Model 1 (Qwen 32B) and Model 2 (GPT-4o-mini). Expert observations provide nuanced insights into specific cases where models performed effectively or where notable shortcomings occurred.

Insights identified as particularly strong by experts included cases where the model clearly articulated relevant and contextually appropriate interpretations. For instance, some expert comments indicated positive recognition when insights accurately reflected candidate responses with appropriate depth, such as effectively capturing participant behaviors and strategic orientations. Conversely, several weak insights were highlighted, notably characterized by vague or overgeneralized conclusions. A frequent critique was "lack of sharp insight" or overly generic interpretations, with some insights described as exhibiting "overconfidence" without sufficient evidence or context provided by candidates. Regarding follow-up questions, experts distinguished sharply between questions deemed highly relevant versus overly generic or poorly targeted. Model-generated questions rated positively by experts were those that effectively probed deeper into specific competencies or candidate responses, facilitating richer and more meaningful elaboration. In contrast, poorly rated follow-up questions were frequently criticized for being too broad or generic, not adequately focusing on the candidate's specific statements or competencies under evaluation, thus limiting their utility in generating valuable responses.

Scoring discrepancies were another notable area of concern. Experts pointed out numerous cases where model-generated scores did not accurately reflect the quality of the candidate responses. Comments frequently highlighted a lack of alignment between insight quality and assigned scores, such as scenarios where strong insights were coupled with disproportionately low answer scores or vice versa. For example, comments indicated situations of " insight and answer score are not aligned" pointing toward inconsistencies in the model's scoring logic.

Quantitative results reinforce these qualitative observations. For insights, 81.2% of Model 1 outputs were rated relevant compared to 53.5% for Model 2. In follow-up questions, 78.7% of Model 1 and 69.8% of Model 2 questions were deemed relevant. However, answer scoring emerged as the weakest dimension, with only 35.2% of Model 1 and 20.9% of Model 2 outputs rated as appropriately aligned with expert expectations.

Additional relevant observations included contextual misunderstandings, where expert justifications noted that certain competencies or contexts, such as Managing Diversity (MD) and Developing Organizational Capabilities (DOC), were not accurately captured or interpreted by the models. Experts specifically criticized instances where the model's scoring and insights did not align well with the actual behavioral indicators presented by participants, suggesting significant room for improvement in these areas. In summary, expert qualitative feedback underscores key strengths in the models' capability to occasionally deliver nuanced insights and precise follow-up questions but simultaneously highlights critical weaknesses. Specifically, the models struggle with maintaining consistency between insight generation, question formulation, and scoring accuracy, particularly in nuanced competencies. These findings indicate clear areas for further refinement and targeted improvements to enhance overall model reliability and validity. Table 9 provides a comparative overview of model strengths and weaknesses, highlighting trade-offs between insight quality, efficiency, and scoring consistency.

Table 9. Model strengths and weaknesses

Model	Strength	Weakness
Model 1 (QWEN 32B)	Provides highly relevant insights (81.19%) and follow-up questions (78.71%), effectively contextualizing candidate responses.	Scoring alignment remains occasionally inconsistent (only 35.15% consistent), leading to many answer scores that do not match evidence (64.85% misaligned).
Model 2 (GPT-4o-mini)	Covers the full set of competencies with 57.43% fewer iterations than Model 1, making it more efficient for rapid assessment.	Generates a higher percentage of irrelevant insights (46.51%) and answer scores (79.07%) misaligned with candidate responses, often lacking depth and specificity in assessment.

 ISSN 2723-6471

5. Conclusion

This study evaluated and compared the performance of two LLMs, Qwen 32B (Model 1) and GPT-4o-mini (Model 2), within the context of automated competency-based assessment using expert annotations. The evaluation centered on three metrics: Insight Fit, Next Question, and Answer Score, with detailed findings presented in Tables 4, 5, 6, and 7. The results indicate that Qwen 32B consistently outperformed GPT-4o-mini in generating contextually relevant and nuanced insights. The Insight scores showed a statistically significant difference in favor of Qwen (Mean = 2.86, SD = 0.62) compared to GPT-4o-mini (Mean = 2.62, SD = 0.64, t = 2.96, p = 0.004, Cohen's d = 0.39). Expert qualitative observations emphasized Qwen's strength in accurately interpreting and differentiating subtle variations in candidate competency demonstrations, particularly noted in competencies such as Strategic Orientation and Building Strategic Partnership.

However, Qwen exhibited variability in generating follow-up questions, evidenced by a slightly higher mean but greater variability (Mean = 3.04, SD = 0.83) compared to GPT-4o-mini (Mean = 2.81, SD = 0.76). This variability indicates that while Qwen often generates targeted and meaningful questions, there are occasions when questions are overly general or insufficiently targeted, limiting their utility in further probing candidate competencies. Conversely, GPT-4o-mini showed relative consistency in scoring, reflected by lower standard deviations across competencies, particularly in Answer Score evaluations (Mean = 2.21, SD = 0.46 compared to Qwen's Mean = 2.35, SD = 0.63). Experts noted GPT-4o-mini's consistent, conservative approach provided predictable baseline evaluations. However, this consistency came with notable weaknesses, as GPT-4o-mini frequently generated overly generic insights and follow-up questions lacking depth and specificity, especially in nuanced competencies such as Managing Diversity and Digital Leadership.

A critical finding is the generally modest Answer Score performance for both models, suggesting notable limitations in accurately assessing candidate responses. Although Qwen achieved slightly higher scores overall, the relatively low average ratings across both models (approximately mid-point on a 5-point scale) highlight a significant gap between model outputs and human expert judgments. Expert annotations consistently indicated discrepancies between model-generated scores and the qualitative quality of candidate responses, underscoring the current challenges of AI-driven assessment systems in comprehensively capturing the complexity of human competencies. This limitation is particularly concerning as accurate answer scoring is fundamental to competency-based assessment systems, where precise evaluation of candidate responses directly impacts the validity of competency conclusions [36]. The consistently low Answer Score classification suggests that current AI models may lack the nuanced understanding required to evaluate complex behavioral responses against established competency criteria, highlighting a significant gap between AI capabilities and human expert judgment in assessment contexts.

The Answer Score was rated low due to several interrelated root causes. First, the model frequently misinterprets nuanced language, resulting in implicit behavioral evidence not being incorporated into the assessment, particularly in bilingual contexts (Indonesian - English) and nuanced competencies like Managing Diversity and Digital Leadership. This pattern is consistent with literature showing disparities in cross-language comprehension and sensitivity to punctuation structure in large models, which can lead to interpretations that miss the candidate's intent and lead to scores that err on the side of response evidence. Second, a lack of contextual memory per session prevents the model from linking subsequent answers to previous conversational traces, which studies have shown can lead to error propagation. Third, prompt ambiguity can encourage generic and stable but shallow output, and overly technical process information without rational justification can diminish perceptions of fairness and clarity, leading to conservative and evidence-insensitive scoring decisions.

These findings emphasize the importance of selecting models aligned to specific assessment requirements, balancing depth of insight with scoring consistency. Future research should prioritize refining models to better mirror human expert judgment, potentially through targeted fine-tuning, improved training datasets, and hybrid assessment frameworks incorporating human oversight [37]. Fine-tuning with a domain-adapted training set based on annotated leadership interviews and expert-rated responses can help align model output more closely with the behavioral rubric. Additionally, RLEF is a viable approach to iteratively refine model scoring accuracy, ensuring consistent quality of insights and scores. Further investigations involving larger datasets, diverse competencies, and comprehensive

qualitative analyses are critical to enhancing the reliability and applicability of automated competency-based assessments. There is also a need for more explicit alignment of assessment rubrics with model output.

A limitation of the present study is the unequal number of expert evaluations for each model, which stems from systematic differences in each model's competency detection granularity per interview turn. This may impact the representativeness and direct comparability of model performance statistics. Future studies should aim for balanced evaluation sample sizes for each model to ensure greater statistical power and comparability in performance analysis.

Another limitation of this study is its narrow participant pool. All participants were drawn from Digital Marketing and Account Manager roles within Indonesian SOEs, which may limit generalizability to other job families (e.g., technical, creative, or operational roles) or to multinational environments operating with different leadership competency frameworks. Consequently, the observed model behaviors may be specific to this context and not directly transferable to other domains. Future research should examine whether the observed model performance extends to a wider range of job roles and organizational contexts. This includes testing with technical and operational positions that emphasize different behavioral indicators, as well as cross-cultural evaluations in multinational corporations where competency frameworks may diverge from those used in Indonesian SOEs. Such expansion would strengthen the external validity of AI-based leadership assessment systems.

6. Declarations

6.1. Author Contributions

Conceptualization: I.G.B.Y.G., A., A.R., and Y.A.; Methodology: I.G.B.Y.G., and A.R.; Software: I.G.B.Y.G.; Validation: I.G.B.Y.G., A., and Y.A.; Formal Analysis: I.G.B.Y.G., A., and A.R.; Investigation: I.G.B.Y.G. and A.; Resources: I.G.B.Y.G., and Y.A.; Data Curation: I.G.B.Y.G.; Writing Original Draft Preparation: I.G.B.Y.G., A., A.R., and Y.A.; Writing Review and Editing: I.G.B.Y.G., A., and A.R.; Visualization: I.G.B.Y.G.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] J. S. Black and P. van Esch, "AI-enabled recruiting: What is it and how should a manager use it?," *Business Horizons*, vol. 63, no. 2, pp. 215–226, Mar. 2020, doi: 10.1016/j.bushor.2019.12.001.

[2] A. K. Upadhyay and K. Khandelwal, "Applying artificial intelligence: Implications for recruitment," *Strategic HR Review*, vol. 17, no. 5, pp. 255–258, Oct. 2018, doi: 10.1108/shr-07-2018-0051.

[3] D. W. Otter, J. R. Medina, and J. K. Kalita, "A survey of the usages of deep learning in natural language processing," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 2020, no. Jul., pp. 604–624, 2020, doi: 10.1109/TNNLS.2020.2979670.

[4] S. Serrano, Z. Brumbaugh, and N. A. Smith, "Language models: A guide for the perplexed," *arXiv preprint arXiv:2311.17301*, vol. 2023, no. Nov., pp. 1–12, 2023, doi: 10.48550/arXiv.2311.17301.

[5] M. F. Gonzalez *et al.*, "Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes," *Computers in Human Behavior*, vol. 130, no. May, pp. 1–12, 2022, doi: 10.1016/j.chb.2022.107179.

[6] R. L. F. Garcia, Y. K. Huang, and L. Kwok, "Virtual interviews vs. LinkedIn profiles: Effects on human resource managers' initial hiring decisions," *Tourism Management*, vol. 94, no. Feb., pp. 1–12, 2023, doi: 10.1016/j.tourman.2022.104659.

[7] T. Zhang *et al.*, "Can large language models assess personality from asynchronous video interviews? A comprehensive evaluation of validity, reliability, fairness, and rating patterns," *IEEE Transactions on Affective Computing*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.1109/TAFFC.2024.3374875.

[8] Y. Bounab, M. Oussalah, N. Arhab, and S. Bekhouche, "Towards job screening and personality traits estimation from video transcriptions," *Expert Systems with Applications*, vol. 238, no. Mar., pp. 1–12, 2024, doi: 10.1016/j.eswa.2023.122016.

[9] C. Qin *et al.*, "Automatic skill-oriented question generation and recommendation for intelligent job interviews," *ACM Transactions on Information Systems*, vol. 42, no. 1, pp. 1–12, Aug. 2023, doi: 10.1145/3604552.

[10] S. Chowdhury, P. Budhwar, and G. Wood, "Generative artificial intelligence in business: Towards a strategic human resource management framework," *British Journal of Management*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.1111/1467-8551.12824.

[11] A. Ujlayan, S. Bhattacharya, and Sonakshi, "A machine learning-based AI framework to optimize the recruitment screening process," *International Journal of Global Business and Competitiveness*, vol. 18, no. S1, pp. 38–53, Dec. 2023, doi: 10.1007/s42943-023-00086-y.

[12] C. Fang, N. Markuzon, N. Patel, and J.-D. Rueda, "Patient-reported outcomes natural language processing for automated classification of qualitative data from interviews of patients with cancer," *Journal of Biomedical Informatics*, vol. 2022, no. 12, pp. 1995–2002, 2022, doi: 10.1016/j.jbi.2022.104565.

[13] U. Leicht-Deobald *et al.*, "The challenges of algorithm-based HR decision-making for personal integrity," in *Business and the Ethical Implications of Technology*, Springer, vol. 2022, no. 1, pp. 71–86, 2022, doi: 10.1007/s10551-019-04204-w.

[14] Y.-H. Chan and Y.-C. Fan, "A recurrent BERT-based model for question generation," in *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, vol. 2019, no. 1, pp. 154–162, 2019, doi: 10.18653/v1/D19-5821.

[15] S. Mahajan *et al.*, "Generative AI-based interview simulation and performance analysis," *International Research Journal of Modernization in Engineering, Technology and Science*, vol. 2024, no. 1, pp. 1–12, 2024, doi: 10.56726/IRJMETS56275.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, vol. 2017, no. 1, pp. 6000–6010, 2017, doi: 10.5555/3295222.3295349.

[17] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2019, no. 1, pp. 4171–4186, 2019, doi: 10.18653/v1/N19-1423.

[18] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *Proceedings of the 28th International Conference on Computational Linguistics*, vol. 2020, no. Nov., pp. 757–770, 2020, doi: 10.18653/v1/2020.coling-main.66.

[19] T. B. Brown *et al.*, "Language models are few-shot learners," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, vol. 2020, no. 1, pp. 1877–1901, 2020, doi: 10.5555/3495724.3495883.

[20] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, vol. 2023, no. 3, pp. 1–12, 2023, doi: 10.48550/arXiv.2303.08774.

[21] J. Bai *et al.*, "Qwen Technical Report," *arXiv preprint arXiv:2309.16609*, vol. 2023, no. Sep., pp. 1–12, 2023, doi: 10.48550/arXiv.2309.16609.

[22] S. Furniturewala *et al.*, "'Thinking' fair and slow: On the efficacy of structured prompts for debiasing language models," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, vol. 2024, no. 1, pp. 213–227, 2024, doi: 10.18653/v1/2024.emnlp-main.13.

[23] S. Al Faraby, A. Romadhony, and Adiwijaya, "Analysis of LLMs for educational question classification and generation," *Computers and Education: Artificial Intelligence*, vol. 7, no. Dec., pp. 1–12, 2024, doi: 10.1016/j.caeai.2024.100298.

[24] M. Li *et al.*, "EZInterviewer: To improve job interview performance with mock interview generator," in *Proceedings of the 16th ACM International Conference on Web Search and Data Mining (WSDM 2023)*, vol. 2023, no. Feb., pp. 1102–1110, 2023, doi: 10.1145/3539597.3570476.

[25] H. L. Chung, Y.-H. Chan, and Y.-C. Fan, "Handover QG: Question generation by decoder fusion and reinforcement learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, no. 1, pp. 3644–3655, 2024, doi: 10.1109/TASLP.2024.3426292.

[26] J. Thakkar, C. Thomas, and D. B. Jayagopi, "Automatic assessment of communication skill in real-world job interviews: A comparative study using deep learning and domain adaptation," in *Proceedings of the ACM International Conference*, vol. 2023, no. Dec., pp. 1–12, 2023, doi: 10.1145/3627631.3627636.

[27] K. Yadav *et al.*, "Interviewing the interviewer: AI-generated insights to help conduct candidate-centric interviews," in *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*, vol. 2023, no. Mar., pp. 723–736, 2023, doi: 10.1145/3581641.3584051.

[28] I. Irawati and M. D. E. Munajat, "The uniqueness of managerial competency model in Indonesian districts and cities," *Policy and Governance Review*, vol. 7, no. 3, pp. 237–260, 2023, doi: 10.30589/pgr.v7i3.774.

[29] World Bank, "Competency standards as a tool for human capital development: Assessment of their development and introduction into TVET and certification in Indonesia," *World Bank Report*, vol. 2020, no. 1, pp. 1–12, 2020, doi: 10.1596/33558.

[30] C. Triwibisono, "Leadership style in Indonesia: Does national culture affect it?," in *Proceedings of the Interuniversity Forum for Strengthening Academic Competency*, vol. 2019, no. 1, pp. 379–383, 2019.

[31] M. Marcellino, "Competency-based language instruction in speaking classes: Its theory and implementation in Indonesian contexts," *Indonesian Journal of English Language Teaching*, vol. 1, no. 1, pp. 1–12, 2005, doi: 10.25170/ijelt.v1i1.1405.

[32] N. Lin *et al.*, "IndoCL: Benchmarking Indonesian language development assessment," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, vol. 2024, no. 1, pp. 4873–4885, 2024, doi: 10.18653/v1/2024.findings-emnlp.280.

[33] C. Rigotti and E. Fosch-Villaronga, "Fairness, AI and recruitment," *Computer Law and Security Review*, vol. 53, no. Jul., pp. 1–12, 2024, doi: 10.1016/j.clsr.2024.105966.

[34] H. Sun *et al.*, "Facilitating multi-role and multi-behavior collaboration of large language models for online job seeking and recruiting," *arXiv preprint arXiv:2405.18113*, vol. 2024, no. 5, pp. 1–12, 2024, doi: 10.48550/arXiv.2405.18113.

[35] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/j.jcm.2017.10.001.

[36] S. Lloyd *et al.*, "Foundations for AI-assisted formative assessment feedback for short-answer tasks in large-enrollment classes," in *Proceedings of the Eleventh International Conference on Teaching Statistics*, vol. 2022, no. Dec., pp. 1–12, 2022, doi: 10.52041/iase.icots11.T3C3.

[37] N. A. Khayi, V. Rus, and L. Tamang, "Towards improving open student answer assessment using pretrained transformers," in *The International FLAIRS Conference Proceedings*, vol. 34, no. 1, pp. 1–12, 2021, doi: 10.32473/flairs.v34i1.128483.