

Augmenting Enterprise Architecture With Large Language Models: A Single-Case Empirical Evaluation Of The TOGAF Preliminary Phase In PropTech

Rizki Darmawan¹ and Alfa Ryano Yohannis²

¹Department of Information Technology, Pradita University, Tangerang,

²Department of Information System, Pradita University, Tangerang, Indonesia

*Corresponding author: rizki.darmawan@student.pradita.ac.id

ABSTRACT

Designing Enterprise Architecture (EA) artifacts, particularly during the initiation phase such as the TOGAF Preliminary Phase, is a crucial foundation for business and technology alignment. However, this formulation process often creates an operational bottleneck as it demands high precision, involves heavy administrative documentation, and consumes a significant amount of time. This single-case study aims to evaluate the effectiveness of utilizing a Large Language Model (LLM)—specifically Gemini 3.1 Pro with a structured Prompt Engineering technique—as a solution to accelerate the design of these documents. Through a case study on the development of Smart Estate Management capabilities within the PropTech ecosystem of PT XYZ, the AI-generated EA artifacts were evaluated by five senior IT experts. The evaluation methodology employed an Expert Judgement approach (Likert scale) cross-validated with in-depth qualitative interviews. The findings indicate a strong agreement among the participating experts that the LLM output possesses a high degree of clarity and structural consistency (median score 5.00/5.00). Furthermore, stakeholders perceived that the validated use of AI could provide an estimated time savings of approximately 80% for initial drafting. Nevertheless, the practicality of document implementation was rated neutral (median score 3.00/5.00) because the AI outputs were found to be inherently generic and unable to accommodate tacit knowledge, legacy infrastructure, and company-specific compliance constraints. Ultimately, this research affirms the necessity of a human-in-the-loop paradigm in the EA discipline within this specific context. It suggests that LLMs function effectively as augmentation accelerators rather than human replacements, thereby shifting the enterprise architect's focus from manual drafting to analytical validation and strategic contextualization.

Keywords: TOGAF Preliminary Phase, Large Language Model (LLM), Prompt Engineering, Expert Judgement.

1. INTRODUCTION

Background

Digital transformation compels the property industry in Indonesia, including leading developers like PT XYZ, to transition from manual facility operations toward a data-driven Property Technology (PropTech) ecosystem. One of the strategic initiatives is the development of Smart Estate Management capabilities to automate services and seamlessly enhance the tenant experience. However, orchestrating innovations such as the Internet of Things (IoT) and artificial intelligence (AI) into this centralized service architecture requires a robust conceptual foundation through an Enterprise Architecture (EA) approach to strictly align IT capabilities with business strategies ([Pereira & Sousa, 2005](#); [Bhattacharya, 2018](#)).

The TOGAF framework, particularly the Preliminary Phase, provides a systematic method for establishing an architectural foundation (Ansyori et al., 2018); however, its execution often triggers operational bottlenecks as it demands high precision, involves heavy administrative

documentation, and consumes significant time. Conversely, Large Language Models (LLMs) have proven efficient in automating technical software engineering tasks. Nevertheless, the application of LLMs for formulating executive-level strategic artifacts like EA continues to provoke academic debate. The primary research gap in the current literature is the absence of empirical evaluation regarding the capability of LLMs to accelerate the design of architecture initiation documents—which rely heavily on tacit knowledge, legacy infrastructure, and organization-specific culture—without degrading their contextual accuracy.

To address this gap, this research offers novelty in the form of a single-case empirical evaluation of the effectiveness and practitioner perceptions regarding TOGAF Preliminary Phase artifacts generated by an LLM (Gemini 3.1 Pro) using structured Prompt Engineering techniques. Through a Smart Estate Management scenario at PT XYZ, this study does not merely test text automation, but rather validates the boundaries of human-in-the-loop practice feasibility within the EA governance discipline; it provides a straightforward literary contribution regarding the separation of roles between the generative model as a conceptual drafter and the human architect as a strategic validator in a highly regulated industry.

Research Problem

Based on the background outlined above, this study is formulated to answer two primary research questions:

1. RQ1: How do stakeholders perceive the effectiveness and quality of the TOGAF Preliminary Phase output generated by an LLM within their organizational context?
2. RQ2: What are the perceived strengths and limitations of using an LLM to generate the Preliminary Phase output, including views on AI-authored versus human-authored content and estimated time savings?

Research Objectives

Generally, this research aims to evaluate the quality and effectiveness of utilizing a Large Language Model (LLM) to generate Preliminary Phase output within the TOGAF framework in a real-world organizational context. Specifically, the objectives of this research are:

1. To analyze stakeholders' perceptions of the quality and effectiveness of LLM-generated TOGAF Preliminary Phase documents based on evaluation dimensions: relevance, completeness, clarity, consistency, and practicality.
2. To identify the main strengths and limitations of using LLMs in the context of enterprise architecture artifact preparation, including perceptions of authenticity (AI vs. human) and potential time efficiency.
3. To provide practical recommendations for PT XYZ regarding the potential adoption of LLMs to support future architecture design processes.

Research Contribution and Novelty

This research is expected to provide contributions and novelty in three main areas:

1. Empirical Contribution: Serving as one of the first empirical studies to evaluate the use of LLMs in generating strategic TOGAF Preliminary Phase artifacts using data and validation from actual stakeholders in the Indonesian property industry.
2. Methodological Contribution: Producing a mixed-method evaluation framework (utilizing Likert scales and thematic analysis) that can be used to assess the quality of AI-generated strategic artifacts within business and enterprise architecture contexts.
3. Academic and Practical Novelty: Bridging the literature gap regarding the application of LLMs to strategic enterprise architecture activities, moving beyond generic tasks like summarization or drafting. Furthermore, the findings provide practical insights for IT management and enterprise architects regarding the potential and risks of AI adoption in the architecture design process.

2. LITERATURE REVIEW

TOGAF Preliminary Phase

Based on its official specification (The Open Group, 2018), TOGAF defines the Architecture Development Method (ADM) as an iterative and systematic guide for designing architecture. However, a synthesis of the literature reveals a tension between the framework's theory and operational pragmatism. Despite being comprehensive, the rigid (out-of-the-box) implementation of TOGAF is often considered too complex and administratively burdensome (Kotusev, 2018), an observation empirically supported by numerous local EA implementation case studies across various sectors in Indonesia (Armi et al., 2024; Fikri et al., 2020; Jatmiko & Saepudin, 2023; Sika et al., 2024; Suyoto et al., 2023). Particularly in the Preliminary Phase—which is crucial for establishing Architecture Principles (Azhary, 2023)—practitioners face the challenge of integrating generic architecture principles with tacit knowledge and tangible company compliance constraints. The need for agile customization, often requiring situational framework adaptations (Kornysheva & Deneckère, 2022) particularly during the requirement management stages (Kornysheva & Barrios, 2021), alongside the heavy administrative burden create an urgency for acceleration tools, such as AI-based automation, to break the bottleneck in this initiation phase.

Large Language Models and Generative AI

Artificial Intelligence (AI) is generally defined as a machine's ability to perform cognitive tasks that resemble human intelligence. One rapidly developing branch is Generative AI, a type of AI focused on creating new content, such as text, images, or other data.

Large Language Models (LLMs), such as the Generative Pre-trained Transformer (GPT) models developed by OpenAI, form the foundation of many current Generative AI applications. LLMs are built on the Transformer architecture, utilizing attention mechanisms

to weigh the importance of different words in a sequence when generating output. These models are trained using vast text datasets from the internet, encompassing billions of words from books, articles, and websites.

LLM capabilities cover various natural language processing (NLP) tasks, such as answering questions, summarizing text, translating, and generating code. However, LLMs have several well-documented limitations, including:

- a. Hallucination: The tendency to convincingly generate incorrect or non-factual information.
- b. Bias: The model may replicate and amplify existing biases in its training data, including gender and racial biases.
- c. Inconsistency: The model may provide different answers to the same question at different times.

Application of LLMs in Business and EA Domains

The application of LLMs in business contexts and organizational decision-making demonstrates significant potential (Chuma & de Oliveira, 2023), albeit still in its early stages. LLMs like ChatGPT can be used as assistive tools for experts to gain a general overview of a business topic, but the generated answers are often generic and superficial, thus unable to replace the role of an expert in strategic decision-making. In specialized domains such as financial analysis, deploying general LLMs faces unique challenges related to specific terminology, linguistic structures, and complex multi-step reasoning requirements (Callanan et al., 2023).

Although LLMs have demonstrated high effectiveness in automating technical software engineering tasks (Hou et al., 2023), the literature indicates that their application at the strategic level of Enterprise Architecture continues to provoke academic debate. Conceptually, LLMs can act as a proxy domain expert (Görgen et al., 2024) capable of generating macro-level modeling elements (Fill & Fettke, 2023). Crucially, however, these models remain vulnerable to hallucinations and the loss of specific business semantics due to the absence of internal company context in their training data. A synthesis of recent literature (White et al., 2023; Ying & Sacks, 2024) suggests that these weaknesses can be mitigated—though not entirely eliminated—through structured Prompt Engineering techniques (role-based and context-injection). Therefore, rather than positioning LLMs as fully autonomous agents, current literature justifies the necessity of an augmentation paradigm (human-in-the-loop), wherein the generative speed of AI must be calibrated by the contextual validation of human architects.

Furthermore, in the property industry context, the integration of these digital technologies drives the evolution of Property Technology (PropTech). Ullah et al. (2018) asserted that modern PropTech data architectures demand convergence between legacy systems and smart

innovations like the Internet of Things (IoT) and AI to realize a Smart Estate ecosystem. Therefore, applying LLMs to design PropTech architecture, as explored in this research, is not merely text automation but an experiment in conceptual modeling to bridge AI agility with enterprise infrastructure and compliance boundaries.

Evaluation Methods and Document Quality

To assess the quality and effectiveness of LLM-generated documents, a multi-dimensional evaluation approach is required. Based on literature studies, several evaluation methods have been identified and applied:

- a. Linguistic and Readability Metrics: This quantitative approach measures text properties. Linguistic metrics include word count, sentence length, and pronoun usage. Meanwhile, readability metrics employ formulas such as the Flesch Reading Ease, Flesch-Kincaid Grade Level, and Gunning Fog Index to assess text difficulty and determine the educational level required to comprehend it.
- b. Sentiment Analysis: This method is used to measure emotional nuances in text, classifying them as positive, negative, or neutral, and identifying specific emotions such as trust or anticipation.
- c. Expert Panel Evaluation: Content quality, particularly technical accuracy, is evaluated by a panel of domain experts. In the medical context, this is termed medical fidelity ([Kharko et al., 2024](#)), which can be adapted into technical fidelity in the EA context. Experts provide qualitative ratings and feedback that can subsequently be analyzed using thematic analysis.
- d. Human Comparison Metrics: LLM output quality can be directly compared against human output using four primary metrics: Accuracy (informational correctness), Completeness, Comprehensibility, and the time required to generate the output.

AI Adoption in Organizations

Although AI is viewed as a critical and disruptive technology, its adoption in large organizations remains in a relatively early stage ([Benbya et al., 2020](#)), with many projects still experimental or in the proof-of-concept phase. There are several primary challenges to AI adoption at the organizational level:

- a. Deployment Problem: One of the most significant challenges is transitioning AI projects from the pilot phase to full-scale production. This process is often complex as it requires integration with legacy technology architecture, greenfield infrastructure designs ([Oberle et al., 2023](#)), business process modifications, and substantial data engineering.
- b. Talent Availability: The lack of skilled AI talent, such as data scientists and AI engineers, is a barrier for many organizations, particularly those outside the technology sector.

- c. Social and Ethical Dysfunction: Managers must recognize and plan ways to circumvent potential AI dysfunctions, such as algorithmic bias, unexplainable outcomes, blurred accountability boundaries, and data privacy issues.

3. METHODS

Research Design

This study adopts a qualitative case study approach, wherein the researcher acts as a practitioner applying this methodology in the PT XYZ work environment. The primary aim of this design is to deeply explore and evaluate the use of the Gemini 3.1 Pro Large Language Model (LLM) in generating Preliminary Phase artifacts of the TOGAF framework. The methodological steps conducted are as follows:

1. Input Document Collection: Identifying and collecting relevant internal documents at PT XYZ as inputs for the LLM. These documents include, but are not limited to, business strategy, IT strategic plans, business principles and objectives, and existing organizational models or structures. For visual documents like organizational charts, the information will be converted into descriptive text format.
2. Hypothetical New Capability Identification: Proposing a new capability relevant to the PT XYZ context. Based on a needs analysis in the property industry, the proposed capability is the "Smart Estate Management System". This aspirational capability aims to improve operational efficiency and the tenant experience, and will serve as the focus for the generated output.
3. Prompt Formulation, Execution, and Ethical Procedures: Formulating a series of structured prompts for the LLM based on internal documents and the proposed new capability. In compliance with ethical procedures and corporate data confidentiality, all input documents were anonymized (data sanitization) prior to being processed by the public model. This experiment utilized the Gemini 3.1 Pro model via the web interface (gemini.google.com), accessed between May and June 2026. The model was executed using default parameters, where the hallucination rate and determinism were strictly controlled through rigorous Output Constraints within the prompt. To ensure output reliability, the generation process was executed for two iterations (number of runs). The final output selected for submission to the evaluators was based on the best Markdown structure completeness from the two runs.
4. Artifact Compilation: Compiling the LLM-generated output into standard TOGAF Preliminary Phase document formats, such as the organizational model for enterprise architecture, the tailored architecture framework, and the initial architecture repository.
5. Stakeholder Evaluation: Submitting the Gemini 3.1 Pro LLM-generated documents for evaluation by five key stakeholders identified within the PT XYZ IT organizational structure. The evaluators are:

- a. Manager of Information Technology for All Unit
 - b. Head of IT Development Division
 - c. Head of IT Network Division
 - d. Project Manager of IT Development Division
 - e. Senior IT System Analyst
6. Evaluation Data Collection: Conducting evaluation through in-depth document reviews followed by semi-structured interviews. These interviews aim to gather stakeholder perceptions regarding the quality, effectiveness, strengths, and limitations of the LLM-generated documents, as well as estimated time if the documents were prepared manually.

Prompt Engineering Instrument Design

To ensure validity, transparency, and reproducibility (the ability of the research to be replicated by other researchers), interaction with the Large Language Model (LLM) was not conducted randomly, but utilized structured Prompt Engineering techniques. The approach utilized in this research is Role-Based Prompting combined with Context-Injection. The prompt design is divided into four main components:

1. System Role: Defines the specific persona, expertise, and perspective the AI must adopt.
2. Context Injection: Provides boundaries for business scope, operational strategy, and specific domain knowledge related to the property industry to ensure outputs are not overly generic.
3. Task Definition: Instructs the LLM to execute precise steps in accordance with Enterprise Architecture framework standards.
4. Output Constraints: Determines the format, grammar, and technical parameters of the generated document.

The operational details of the LLM instructions were configured through Prompt Engineering techniques consisting of four main components: System Role, Context Injection, Task Definition, and Output Constraints. To maintain article readability while ensuring research reproducibility, the complete text of the prompt utilized, along with the model's parameter settings, are enclosed in Appendix A at the end of this document.

The use of specific instructions in the Context Injection section aims to direct the model to not merely generate out-of-the-box architecture principles from the TOGAF manual, but rather align them with the data-driven Smart Estate Management innovations that have become the company's strategic requirement.

Evaluation Dimensions for RQ1

To answer the first research question (RQ1) regarding stakeholder perceptions of output effectiveness and quality, this study employs a structured Expert Judgement approach. Initial data collection was conducted using a Likert scale instrument (1-5) to measure the level of consensus (agreement) among experts on each evaluation dimension. This approach was chosen not as a statistical significance test—given the limited size of the expert panel (N=5)—but rather as a qualitative calibration instrument. Subsequently, semi-structured interviews were utilized to explore the contextual and strategic reasoning behind the given consensus scores. The evaluation dimensions used are as follows:

1. **Relevance:** The extent to which the document content aligns with the needs and business context of PT XYZ.
2. **Completeness:** Whether all critical aspects required in the TOGAF Preliminary Phase have been adequately covered in the document.
3. **Clarity:** How clear and comprehensible the document content is, including its structure, language, and terminology used.
4. **Consistency:** Whether the document content is internally consistent and aligns with standard TOGAF terminology.
5. **Practicality/Actionability:** The extent to which the generated document can be directly used or adapted for practical needs within the organization.

Evaluation Questions for RQ1

The following is the list of evaluation questions posed to stakeholders to answer RQ1:

1. How relevant is this document to the needs and context of your organization? (Answer on a 1-5 Likert scale and explain the reasoning behind your rating).
2. Does this document cover all crucial aspects required in the TOGAF Preliminary Phase? (Answer on a 1-5 Likert scale and explain the reasoning behind your rating).
3. How clear and comprehensible is the content of this document? (Answer on a 1-5 Likert scale and explain the reasoning behind your rating).
4. Is the content of this document consistent with and in accordance with the TOGAF terminology you know? (Answer on a 1-5 Likert scale and explain the reasoning behind your rating).
5. How practical is this document for direct use or adaptation within your organization? (Answer on a 1-5 Likert scale and explain the reasoning behind your rating).

Methods to Answer RQ2

To answer the second research question (RQ2) regarding the strengths, limitations, and perceptions related to LLM usage, this study employed a qualitative approach through semi-structured interviews. This approach aims to uncover deep understanding regarding the benefits, risks, perceived authorship authenticity (AI vs. human), and the perceived efficiency of the generated documents. The transcripts of stakeholder responses and qualitative feedback were analyzed using a structured thematic analysis procedure (Braun & Clarke, 2006). This

procedure included the phases of data familiarization, initial deductive code generation based on EA evaluation dimensions, theme searching, and the reviewing and defining of final themes. The coding process was conducted manually using Microsoft Excel and Word software to crystallize narratives regarding the strengths, limitations, and perceived authenticity of the AI-generated documents.

Evaluation Questions for RQ2

The following is the list of evaluation questions used in the interviews to answer RQ2:

1. In your opinion, what are the primary strengths of this LLM-generated Preliminary Phase document?
2. What are the limitations or weaknesses you found in this document?
3. Based on the content and writing style of the document, do you think this document was:
 - a. Written entirely by a human,
 - b. Written entirely by AI, or
 - c. Written by AI with human editing? Please explain the reasons for choosing your answer.
4. In your estimation, if you were to create/write a document of similar quality yourself without AI assistance, how much time would it take to complete? (Provide estimates in hours or days).
5. Do you see the potential of using LLMs like this to support your work in the future? Why?

Limitations and Threats to Validity

This research has several limitations and potential threats to validity that must be noted:

1. Generalization of Findings: As a single case study at PT XYZ, the findings obtained are contextual and may not be directly generalizable to other organizations with different characteristics.
2. Input Quality Variation: The quality of the LLM output heavily depends on the quality and completeness of the input documents provided by the researcher. Differences in input formulation can affect the final outcome.
3. Respondent Evaluation Bias: Stakeholder assessments may be influenced by their personal preferences toward AI, their level of understanding of TOGAF, or internal organizational biases, which could potentially affect evaluation objectivity.
4. Construct Validity Threats: There is a possibility that the evaluation dimensions used have not fully covered all quality aspects, or there may be differences in dimensional interpretation among respondents.
5. Internal Validity Threats: The non-experimental research design means there is a possibility of external factors (e.g., changes in internal company policy) affecting respondent perceptions during the evaluation period.

6. External Validity Threats: Research outcomes may not apply to organizations with extremely high or extremely low architectural maturity levels, as this context could affect the relevance of LLM usage.

Despite these limitations, this research is expected to provide an important initial contribution to understanding the potential and challenges of using LLMs to support the TOGAF Preliminary Phase in a real-world organizational environment.

4. RESULT AND DISCUSSION

4.1. Analysis of RQ1: Stakeholder Perceptions on the Quality and Effectiveness of LLM Output

The analysis for the first research question (RQ1) integrates quantitative data from the Likert scale ratings with qualitative data from the respondents' explanations.

Table 1. Participant Profile

Role	Experience Level (years)
Manager of IT-All Unit	>15
Head of IT Development Division	11-15
Head of IT Network Division	11-15
IT Development Project Manager	5-10
Senior IT Systems Analyst	5-10

Based on the participant profile, three out of five respondents (60%) possess more than 10 years of experience (two respondents in the 11-15 years range, and one >15 years). This seniority level adds significant qualitative weight to the evaluation results.

Expert Consensus Assessment Results

The structured evaluation from the five key stakeholders regarding the LLM-generated Preliminary Phase document is summarized in Table 2. The assessment was based on a Likert scale (1-5), where consensus was measured by the uniformity of scores given by the panelists. Ratings were conducted using a Likert scale, where 1 means "Strongly Disagree/Poor" and 5 means "Strongly Agree/Excellent".

Table 2. Expert Evaluation on LLM Output

Evaluation Dimension	Respondent 1	Respondent 2	Respondent 3	Respondent 4	Respondent 5	Median Score
D1. Relevance	4	4	4	4	4	4.00
D2. Completeness	4	4	4	4	4	4.00

D3. Clarity	5	5	5	5	5	5.00
D4. Consistency	5	5	5	5	5	5.00
D5. Practicality	3	3	3	3	3	3.00

In accordance with methodological guidelines, Table 2 presents the median scores for each evaluation dimension. The evaluation results indicate a strong agreement among the panelists, where the Clarity (Median = 5.00) and Consistency (Median = 5.00) dimensions received the highest scores, indicating that the LLM output was rated exceptionally well in terms of ease of understanding and adherence to standard terminology. Relevance (4.00) and Completeness (4.00) were also rated well. On the other hand, Practicality (3.00) received the lowest score, signaling reservations among the respondents regarding the extent to which the document could be directly utilized without modification.

Qualitative Analysis and Discussion

1. Discussion on Relevance (D1) and Completeness (D2) Dimensions - Score 4.00 (Good)

All respondents collectively agreed (100%) that the LLM-generated document is relevant to the organization's needs and complete in covering the critical aspects of the TOGAF Preliminary Phase. Stakeholders were impressed by the LLM's ability to contextualize general architecture principles into the property industry and the specific "Smart Estate Management" initiative. However, the score did not reach perfection due to notes indicating the document lacked depth in capturing the nuances and undocumented "tacit knowledge" distinct to PT XYZ. The document was viewed as an excellent foundation but still requires enrichment with internal insights.

2. Discussion on Clarity (D3) and Consistency (D4) Dimensions - Score 5.00 (Excellent)

Both of these dimensions achieved absolute agreement with perfect scores. Respondents praised the LLM output for its highly structured, professional language that was easily understandable, even by managers without a deep technical background. The logical document structure and consistent application of TOGAF terminology were recognized as major strengths. One respondent stated that "the clarity and format exceeded the standards of internal documents we often manually create," which highlights the LLM's potential for documentation standardization.

3. Discussion on Practicality Dimension (D5) - Score 3.00 (Neutral)

This dimension was unanimously agreed upon as the weakest area. Although rated as an excellent initial draft, all respondents agreed the document could not be directly implemented (lacked practicality) without significant validation and adjustment by the internal team. The underlying reason was that the LLM output tended to be generic and did not yet account for practical organizational constraints, such as budgets, internal politics, and team readiness. This

document was viewed more as an "accelerator" or a starting point for discussion, rather than a final solution.

4.2. Analysis of RQ2: Strengths, Limitations, and Perceptions of LLM Usage

The analysis for RQ2 is based on interview responses from the questionnaire.

Theme 1: Primary Strengths (P6)

All respondents identified similar strengths. The most prominent primary strength was the speed and efficiency in generating a comprehensive draft document. Furthermore, the professional and logical structure, along with consistent formal language usage, were highly appreciated. The LLM-generated document was considered an "excellent initial idea" capable of drastically cutting down brainstorming and framework-drafting time.

Theme 2: Primary Limitations and Weaknesses (P7)

The most frequently cited limitation was the lack of company-specific context. The LLM was incapable of capturing culture, internal dynamics, and undocumented tacit knowledge. This caused several sections of the document to feel overly generic. Consequently, the greatest weakness is that LLM output necessitates stringent human validation to review accuracy and contextual relevance, and to anticipate potential "hallucinations" or inaccurate information.

Critical Analysis of Contextual Gaps (Visual Evidence)

Although the LLM output was rated as having very high clarity and consistency, panelists gave the lowest score on the Practicality dimension (D5), citing results that "tended to be generic". To understand this gap more objectively, Table 3 presents visual evidence comparing the raw output from the LLM with the human refinement performed by internal IT architects.

Table 3. Comparison of LLM Output and Expert Refinement on the Preliminary Phase Document

TOGAF Aspect	Original LLM Output (Raw Output)	Expert Refinement (Human Refinement)	Gap Analysis
Data Architecture Principle	"Tenant data must be securely stored and managed using a centralized cloud-based relational database to facilitate access."	"Sensitive tenant data is managed using an on-premise database integrated with the local Data Center in accordance with the company's internal compliance policies. Non-sensitive analytical data can be synchronized to the hybrid-cloud ecosystem."	The LLM lacks context regarding internal data security policies and local privacy regulations—specifically referring to the Republic of Indonesia Law Number 27 of 2022 concerning Personal Data Protection (UU PDP)—which mandates strict protection of sensitive data, thereby justifying the use of on-premise infrastructure within PT XYZ.
Application Integration	"The integrated facility management	"The smart estate inspection application is integrated via	The LLM provided industry-standard recommendations (Open API), but

TOGAF Aspect	Original LLM Output (Raw Output)	Expert Refinement (Human Refinement)	Gap Analysis
(Smart Estate)	application must be connected through an open Application Programming Interface (API)."	the company's existing Secure API Gateway mechanism, adopting the Microservices protocol to minimize legacy server loads."	failed to account for the existence of legacy servers and API Gateways already owned and established as operational standards by the organization.

Based on the comparison in Table 3, the architectural limitations of the LLM are evident. In the data architecture principle, the LLM proposed a centralized cloud solution which is conceptually sound and modern. However, in practice, human Enterprise Architects had to radically alter this principle toward a hybrid or on-premise approach due to internal compliance constraints, tenant data sensitivity (tacit knowledge), and applicable legal regulations.

Additionally, in the application integration aspect, the LLM recommendation remained at a macro conceptual level. Systems analysts had to translate those generic sentences by incorporating variables of legacy infrastructure that are genuinely operating within the company. This substantiates previous findings that the LLM acts exceptionally well as a "baseline framework creator," yet it lacks spatial reasoning capabilities concerning actual infrastructure and the company's operational budget constraints. This contextualization process consumes the remaining 20% of the working time and underscores that the role of a human architect (human-in-the-loop) is absolutely essential before architectural artifacts are ratified.

Theme 3: Authorship Perception (AI vs. Human) (P8)

A total of 4 out of 5 respondents answered that the document was most likely "Written by AI with human editing". Their primary reasoning was that the level of linguistic consistency and perfect structure is difficult for a human writer to achieve in a single draft, yet the depth of content still felt as though it required final finishing touches from an expert for adjustment.

Theme 4: Estimated Time Savings (P9)

Regarding efficiency, four out of five respondents (80%) perceived that the document was authored through human-AI collaboration (hence, the perception pie chart has been removed for visual efficiency). Furthermore, stakeholder estimation indicates a potential 80% reduction in initial drafting time. It is crucial to note that these values are strictly perceptual projections based on the panelists' experience, rather than observationally measured productivity data.

Theme 5: Potential LLM Adoption in the Future (P10)

All respondents were highly optimistic concerning the future potential adoption of LLMs. They view LLMs as an extremely useful support tool, particularly for three main activities: (1)

Drafting initial outlines for various document types (reports, proposals, policies), (2) Brainstorming and idea exploration, and (3) Standardizing documentation format and quality across departments.

4.3. Implications of Findings

Practical Implications and Approach Generalizability

The findings of this study provide strong evidence that LLMs function as a strategic tool to improve the efficiency of the enterprise architecture planning process. Furthermore, the proposed human-in-the-loop approach possesses a high degree of generalizability for adoption by other large-scale property developers currently transitioning toward smart city ecosystems. Although the AI output proved capable of automating 80% of draft preparation, the role of human architects remains an absolute prerequisite that can be replicated across organizations, particularly for contextualizing frameworks into data-driven PropTech business models and integrating them with complex service architectures and legacy infrastructures. This confirms that the shifting role of the architect—from draft compiler to analytical validator—represents a new operational standard applicable in industries with strict compliance and data confidentiality regulations. It is recommended that PT XYZ begin formulating internal guidelines regarding ethical and effective LLM utilization, including a mandatory human validation process for any output intended for decision-making.

1. Theoretical Implications and Methodological Limitations

Theoretically, this study contributes empirical evidence on the application of LLMs for complex cognitive tasks such as Enterprise Architecture. Regarding the evaluation limitation of involving only five IT professionals, this quantitative limitation is methodologically compensated by a high degree of qualitative depth. The evaluators possess managerial seniority (the majority with >10 years of experience) and specific domain expertise in corporate-level system orchestration. This small-scale yet deep expert judgement evaluation is sufficiently robust to validate the structural consistency of AI outputs at this initial stage, although longitudinal studies with broader cross-departmental respondent samples are required in the future to achieve statistical significance. Ultimately, these findings reinforce the foundation of the human-AI collaboration model, wherein the agility of generative models must always be calibrated by human critical thinking

5. CONCLUSIONS

Conclusion

This study aimed to evaluate the quality and effectiveness of utilizing a Large Language Model (LLM) in generating TOGAF Preliminary Phase artifacts within a real corporate environment, specifically PT XYZ. Based on quantitative and qualitative data analysis from

evaluations conducted by five key IT stakeholders, several primary conclusions can be drawn as follows:

1. **LLMs Excel in Document Structure, Clarity, and Consistency:** The LLM-generated document received exceptionally high ratings (an average score of 5.00 out of 5.00) in the clarity and consistency dimensions. This indicates that the LLM is capable of producing artifacts that are professionally structured, easily understandable, and aligned with standard TOGAF terminology. This capability serves as one of the LLM's greatest strengths in documentation standardization.
2. **Relevant and Complete Foundation, Yet General in Nature:** In the relevance and completeness dimensions (average score of 4.00), the LLM output was rated as capable of providing a solid foundation covering nearly all essential required aspects. However, its deficiency lies in the LLM's inability to capture internal context, culture, and "tacit knowledge" specific to PT XYZ, resulting in outputs that trend toward being generic.
3. **The Role of LLM as an Accelerator, Not a Replacement:** The lowest score in the practicality dimension (average score of 3.00) confirms that the LLM output cannot be used directly (as-is). Experts view it as an excellent initial draft or accelerator capable of saving significant time (an estimated 80% on average), but it still requires validation, refinement, and context enrichment by human experts.
4. **Human-AI Collaboration is the Most Realistic Working Model:** The majority of respondents (4 out of 5) perceived the document as the result of human and AI collaboration. This reinforces the conclusion that the most effective and realistic working model for LLM adoption in strategic tasks is augmentation (enhancing human capability), rather than full automation. The LLM handles the drafting work, while humans focus on critical thinking, strategy, and validation.

Overall, within the confines of this single-case study, the findings indicate that LLMs possess significant potential to transform the workflow of Enterprise Architecture document preparation. These conclusions are strictly limited to the perceptions of internal stakeholders at PT XYZ. The primary value of generative AI in this initiation phase lies in its capability as a drafting accelerator that perceptually improves time efficiency, under the absolute prerequisite that contextual validation and analytical compliance by human experts (human-in-the-loop) remain irreplaceable.

Future Development Plans

Based on the findings and limitations identified in this study, the following are several development plans and directions for future research that can be explored:

1. **Application in Other TOGAF ADM Phases:** This study was limited to the Preliminary Phase. Future research can test LLM effectiveness in more complex phases of the

TOGAF ADM cycle, such as Phase B (Business Architecture) or Phase C (Information Systems Architecture), which demand more profound solution analysis and design.

2. Exploration of Advanced Prompt Engineering Techniques: To address the limitation of generic outputs, future research could explore more sophisticated prompt engineering techniques. This could include the development of a "Prompt Library" tailored for EA tasks or utilizing few-shot learning methods, where the LLM is provided several high-quality internal document examples as references to generate more contextual outputs.
3. Comparative Analysis Between LLM Models: This study utilized one type of LLM. Comparative research could be conducted to compare the performance of various LLMs available in the market (e.g., models from OpenAI, Google, Anthropic, or open-source models) in handling the same EA tasks, to identify which model is most suitable.
4. LLM Integration with EA Tools and Workflow Automation: A practical development plan for PT XYZ is to explore the possibility of integrating LLM capabilities directly into existing architecture management or modeling tools (e.g., ArchiMate-based software). This could create a more seamless workflow where the LLM not only generates text but also assists in directly drafting visual models or analyzing architecture data.
5. Longitudinal Studies on Adoption Impact: Conducting longitudinal studies to track the long-term impact of LLM adoption within architecture teams. Such studies could quantitatively measure its influence on project completion speed, decision quality, and business stakeholder satisfaction levels over time.

6. REFERENCES

- Ansyori, R., Qodarsih, N., & Soewito, B. (2018). A systematic literature review: Critical Success Factors to Implement Enterprise Architecture. *Procedia Computer Science*, 135, 43–51.
- Armi, N. M. D., Susila, A. A. N. H., & Pasirullo, M. A. (2024). Perancangan Arsitektur Enterprise Menggunakan Framework TOGAF (Studi Kasus: Badan Pusat Statistik Provinsi Bali). *Journal of Comprehensive Science*, 3(8), 3499-3505.
- Azhary, R. I. (2023). Analisis dan Perancangan Enterprise Architecture Menggunakan Framework Togaf Adm Pada Fungsi Perencanaan dan Pengendalian Strategis di PT Industri Telekomunikasi Indonesia. *Syntax Literate; Jurnal Ilmiah Indonesia*, 8(3), 2249-2268.
- Benbya, H., Davenport, T. H., & Pachidi, S. (2020). Artificial Intelligence in Organizations: Current State and Future Opportunities. *MIS Quarterly Executive*, 19(4).

- Bhattacharya, P. (2018). Aligning Enterprise Systems Capabilities with Business Strategy: An extension of the Strategic Alignment Model (SAM) using Enterprise Architecture. *Procedia Computer Science*, 138, 655–662.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101.
- Callanan, E., Mbakwe, A., Papadimitriou, A., Pei, Y., Sibue, M., Zhu, X., Ma, Z., Liu, X., & Shah, S. (2023). Can GPT models be Financial Analysts? An Evaluation of ChatGPT and GPT-4 on mock CFA Exams. *arXiv preprint arXiv:2310.08678*.
- Chuma, E. L., & de Oliveira, G. G. (2023). Generative AI for Business Decision-Making: A Case of ChatGPT. *Management Science and Business Decisions*, 3(1), 5–11.
- Fikri, A. H., Purnomo, W., & Putra, W. H. N. (2020). Perancangan Enterprise Architecture Menggunakan TOGAF ADM pada PT. Hafintech Prima Mandiri. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 4(7), 2032–2042.
- Fill, H. G., & Fettke, P. (2023). Conceptual modeling and large language models: impressions from first experiments with ChatGPT. *Enterprise Modelling and Information Systems Architectures (EMISAJ)*, 18, 1-15.
- Görgen, L., Müller, E., Triller, M., Nast, B., & Sandkuhl, K. (2024). Large Language Models in Enterprise Modeling: Case Study and Experiences. In *Proceedings of the 12th International Conference on Model-Based Software and Systems Engineering (MODELSWARD 2024)*, 74-85.
- Hou, X., Zhao, Y., Liu, Y., Yang, Z., Wang, K., Li, L., ... & Grundy, J. (2023). Large language models for software engineering: A systematic literature review. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, 33(5), 1-41.
- Jatmiko, W., Saepudin, R. P. S. (2023). Perancangan Model Enterprise Architecture Framework Togaf Pada Penjualan Alat - Alat Olahraga di Toko Sentral Sport. *JURSISTEKNI (Jurnal Sistem Informasi dan Teknologi Informasi)*, 5(2), 214-224.
- Kharko, A., McMillan, B., Hagström, J., Muli, I., Davidge, G., Häggglund, M., & Blease, C. (2024). Generative artificial intelligence writing open notes: A mixed methods assessment of the functionality of GPT 3.5 and GPT 4.0. *Digital Health*, 10, 1-14.
- Kornysheva, E., & Barrios, J. (2021). Process-oriented Knowledge Representation of the Requirement Management Phase of TOGAF-ADM: an Empirical Evaluation. *Procedia Computer Science*, 192, 2239–2248.
- Kornysheva, E., & Deneckère, R. (2022). A Proposal of a Situational Approach for Enterprise Architecture Frameworks: Application to TOGAF. *Procedia Computer Science*, 207, 3493–3500.

- Kotusev, S. (2018). TOGAF-based Enterprise Architecture Practice: An Exploratory Case Study. *Communications of the Association for Information Systems*, 43, 321-359.
- Oberle, M., Yesilyurt, O., Schlereth, A., Risling, M., & Schel, D. (2023). Enterprise IT Architecture Greenfield Design Combining IEC 62264 and TOGAF by Example of Battery Manufacturing. *Procedia Computer Science*, 217, 136–146.
- Pereira, C. M., & Sousa, P. (2005). Enterprise Architecture: Business and IT Alignment. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, 1344-1345.
- Sika, X., Astri, L. Y., Kisbianty, D., Setiawan, R., & Irawan. (2024). Implementasi Framework TOGAF ADM Bagian Kemahasiswaan Universitas Dinamika Bangsa Jambi. *Jurnal Ilmiah MEDIA SISFO*, 18(1), 1-7.
- Suyoto, I. D. M., Indra, Wedi, S., & Setiawan, K. (2023). Perancangan Arsitektur Sistem dan Teknologi Informasi Menggunakan TOGAF ADM (Studi Kasus Kantor Pertanahan ABC). *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIK)*, 10(4), 909-918.
- The Open Group. (2018). *The TOGAF® Standard, Version 9.2*. Van Haren Publishing.
- Ullah, F., Sepasgozar, S. M., & Wang, C. (2018). A systematic review of smart real estate technology: Drivers of, and barriers to, the use of digital disruptive technologies and online platforms. *Sustainability*, 10(9), 3142.
- White, J., Hays, S., Fu, Q., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382*.
- Ying, H., & Sacks, R. (2024). From Automatic to Autonomous: A Large Language Model-driven Approach for Generic Building Compliance Checking. *Proceedings of the CIB W78 Conference 2024*.