Accelerating the TOGAF Preliminary Phase: An Empirical Evaluation of LLM-Generated Artifacts in Banking

Izaaz Azaam Syahalam[1]
Department of Information Technology
Pradita University
Tangerang, Indonesia
izaaz.azaam@student.pradita.ac.id

Alfa Ryano Yohannis[2]
Department of Information Technology
Pradita University
Tangerang, Indonesia
alfa.ryano@pradita.ac.id

Teddy Mantoro[3]
School of Computer Science
Nusa Putra University
Sukabumi, Indonesia
teddy.mantoro@nusaputra.ac.id

***Abstract*—Enterprise Architecture (EA) facilitates alignment between business strategy and IT governance, yet the TOGAF Preliminary Phase remains resource-intensive and time-consuming. This study empirically evaluates GPT-4's capability to generate TOGAF Preliminary Phase artifacts within a real banking environment. Using a mixed-methods approach, three certified EA experts assessed five artifacts across clarity, completeness, relevance, accuracy, and usability dimensions using a 25 item ISO/IEC 25010 based instrument. Results indicate high perceived quality (mean = 4.26/5.0; α = 0.86) and efficiency gains of up to 97 % compared to manual development. Despite a Gunning Fog Index (GFI) of 25.03, experts rated clarity highly (4.33/5.0), revealing a Complexity-Appropriate Paradox linguistic complexity does not reduce comprehension among expert audiences. Qualitative findings emphasize efficiency and structure as key strengths and contextual understanding as the main limitation. The results highlight LLMs as accelerators rather than replacements in EA processes, advancing a Human AI collaboration framework for architecture governance.**

Keywords-Enterprise Architecture, TOGAF, Large Language Models (LLM), Preliminary Phase, Banking Industry.

I. INTRODUCTION

Enterprise Architecture (EA) provides an integrative framework aligning business processes, information systems, and IT infrastructure [1]. Among available methodologies, TOGAF is the most widely adopted due to its structured lifecycle and governance rigor [2]. However, the Preliminary Phase defining architectural principles, governance structures, and stakeholder readiness demands extensive expert time and contextual knowledge, often spanning weeks of effort[3], [4].

Meanwhile, Large Language Models (LLMs) such as GPT-4 exhibit remarkable text-generation capability [4], [5], [6]. Their application in enterprise contexts is emerging, yet empirical validation remains limited, particularly for knowledge-intensive, context-specific domains such as EA. Previous studies show promise in using LLMs for structured documentation and technical drafting [7] and software engineering [8], whereas EA involves deeper contextual and organizational complexity.

This study bridges that gap through a real-world evaluation of GPT-4 generated TOGAF Preliminary Phase artifacts within a banking institution. The study seeks to answer:

1. RQ1: How do experts assess the quality and effectiveness of such artifacts?
2. RQ2: What strengths, limitations, and implications arise from applying LLMs in strategic EA processes?

This paper contributes: (1) empirical evidence of LLM performance in EA artifact generation, (2) identification of the Complexity-Appropriate Paradox, and (3) a Human–AI collaboration framework for architecture development acceleration.

II. RELATED WORK

This section reviews literature across three interconnected domains: Enterprise Architecture frameworks and their implementation challenges, LLM applications in knowledge work, and evaluation methodologies for AI-generated content.

A. *Enterprise Architecture and TOGAF Implementation*

Enterprise Architecture frameworks provide structured approaches for aligning business and IT strategy. Among them, TOGAF, developed by The Open Group, is the most widely applied, comprising sequential phases from Preliminary to Architecture Change Management [1].

Kornyshova and Deneckère [1] emphasize that effective EA requires situational adaptation, as rigid TOGAF application often fails to reflect organizational specificity a challenge central to this study. Oberle et al. [2] illustrate TOGAF's flexibility by integrating it with IEC 62264 for manufacturing IT architectures, demonstrating cross sector adaptability.

The Preliminary Phase forms the foundation for defining architecture principles, governance, and stakeholder models [3]. Anggraini et al. [3] identify it as the most time intensive stage, particularly in cloud adoption, due to extensive consultation and documentation. Yet, limited research explores how this phase can be accelerated without compromising quality a gap this paper addresses

B. *Large Language Models in Knowledge Work*

Recent advances in LLMs (e.g., GPT-3, GPT-4) have transformed knowledge intensive work, enabling coherent, context aware text generation across reasoning and synthesis tasks [4], [5]. Gupta et al. [4] propose a framework for GenAI adoption, highlighting governance and acceptance factors. Kumar et al. [5] identify task suitability, organizational readiness, and human AI collaboration as key success drivers.

Dwivedi et al. [8] discuss ChatGPT's cross-domain implications, while Fui-Hoon Nah et al. [9] emphasize that optimal human AI synergy arises when AI automates structured tasks and humans provide contextual judgmental integration model tested in our EA context.

C. *Evaluation of LLM-Generated Content*

Evaluating LLM outputs remains complex, especially for domain specific tasks requiring accuracy and contextual

validity. Rasool et al. [6] show that LLMs struggle with precise retrieval and numerical reasoning, demonstrating the need for domain adapted evaluation. Vaira et al. [7] compare GPT-4, Bard, and human-written medical documents using hybrid expert quantitative assessment, finding acceptable quality but stressing the need for expert validation.

Most prior evaluations focus on domains with explicit correctness (e.g., medicine, law, translation). In contrast, EA artifact generation demands a blend of technical accuracy, strategic alignment, and contextual appropriateness dimensions not easily quantifiable.

D. *Research Gap*

Although prior work establishes LLM potential and evaluation frameworks, no empirical studies assess their capability in strategic, context-dependent EA artifact generation. This gap is critical, as EA outputs require organizational context, stakeholder acceptance, and expert validation. Our study fills this void through an empirical evaluation of GPT-4 generated TOGAF Preliminary Phase artifacts, combining quantitative quality measures with qualitative expert assessment in a real organizational setting.

III. METHODOLOGY

This study adopts a mixed-methods single case design, integrating quantitative quality assessment and qualitative expert feedback to evaluate LLM generated EA artifacts. This approach aligns with prior recommendations for evaluating AI in professional settings [7], [9], balancing objective measurement with contextual insight.

A. *Research Design and Context*

A single embedded case study [10] was conducted within a mid-sized commercial bank undergoing digital transformation. The organization classified as EA Maturity Level 2 (Developing) operates 397 branches with over 11,000 employees under OJK and BI regulations. This setting provided both contextual depth and access to domain experts, making it ideal for examining LLM performance in a regulated industry.

B. *Participants*

Three experts were purposively selected based on EA and banking expertise, direct involvement in transformation projects, and familiarity with TOGAF. Though limited (n = 3), shown in TABLE I such sampling ensures methodological rigor in expert evaluation [7].

TABLE I Participant Profiles

ID	Role	Experience	EA Knowledge	Banking Experience
P1	IT Analyst	4 Years	EA familiar	3 Years
P2	Enterprise Architect	10 Years	TOGAF certified	4 Years
P3	Business Analyst	3 Years	EA Knowledge	3 Years

C. *LLM Configuration and Artifact Generation*

Artifacts were produced using GPT-4 (gpt-4-0613) via the OpenAI API. The model received a structured 10-page brief covering organizational strategy, IT landscape, regulatory context, and stakeholder profiles

Prompts following best practices [11]. Example Prompt: "You are a TOGAF certified Enterprise Architect. Based on the provided context, generate an Architecture Principles document for the Preliminary Phase, defining 8–10 principles with name, statement, rationale, and implications, using professional EA terminology suitable for executive presentation" Artifacts Generated:

1. Architecture Principles Document
2. Stakeholder Management Approach
3. Architecture Governance Framework
4. Tailored Architecture Framework
5. Request for Architecture Work

Each was produced in a single session without post-editing to simulate realistic enterprise use.

D. *Evaluation Instruments*

A 25-item questionnaire (Likert scale 1–5) was adapted from ISO/IEC 25010 and EA literature to assess five quality dimensions: clarity, completeness, relevance, accuracy, and usability. The tool was pilot-tested with two EA practitioners and reviewed by an academic expert for construct validity.

Following the survey, participants joined semi-structured interviews (45–60 min) discussing strengths, weaknesses, and adoption potential. Interviews were recorded, transcribed, and analyzed thematically.

E. *Objective and Mixed-Methods Analysis*

Readability was measured using the Gunning Fog Index (GFI) to assess linguistic complexity:

$$\mathrm{GFI} = 0.4 \times \left(\frac{\text{words}}{\text{sentences}} + 100 \times \frac{\text{complex words}}{\text{words}} \right) \qquad (1)$$

Quantitative data were analyzed through descriptive statistics, radar visualizations, and inter-rater reliability using Krippendorff's alpha. Qualitative data were examined via Braun and Clarke's six-phase thematic analysis, with independent coding and consensus validation. Both datasets were merged through a convergent mixed-methods approach to ensure triangulation and interpretive validity.

F. *Ethical and Validity Measures*

All participants provided informed consent; data were anonymized, encrypted, and stored securely. The study received ethics board approval. Construct, internal, and external validity were addressed through triangulation, validated instruments, and transparent case reporting. Reliability was ensured via systematic data procedures and inter-rater agreement checks.

G. *Limitations*

The study's scope was confined to GPT-4 and a single banking organization. While depth compensates for breadth, broader replication across industries, models, and TOGAF phases is recommended to enhance generalizability.

IV. RESULT AND DISCUSSION

This section presents findings from our mixed-methods evaluation, organized into three subsections: quantitative quality assessments, objective readability analysis, and qualitative stakeholder feedback.

A. *Quantitative Quality Assessment*

Overall Quality Ratings: TABLE II summarizes descriptive statistics across five quality dimensions and 15 evaluator ratings.

TABLE II. Quality Dimension Ratings

Dimension	Mean	Median	SD	Min	Max	95% CLI
Relevance	4,33	4	0,58	4	5	(3.68, 4.99)
Completeness	4,33	4	0,58	4	5	(3.68, 4.99)
Clarity	4,33	5	1,15	3	5	(3.03, 5.64)
Accuracy	4,33	4	0,58	4	5	(3.68, 4.99)
Usability	4	4	0	4	4	(4, 4)
Overall	4.26	4	0.71	3	5	

All dimensions scored above 4.0 ("Agree"), indicating strong expert acceptance. Clarity showed the greatest variation (SD = 1.15), suggesting differing perspectives on readability across roles.

Artifact Level Analysis: TABLE III presents mean ratings by artifact type.

TABLE III. Quality Ratings by Artifact Type

Artifact	PI	P2	P3	Overall
Architecture Principles	5	4	4	4.33
Stakeholder Management	4	4	5	4.33
Governance Framework	3	5	5	4.33
Tailored Framework	4	5	4	4.33
Request for Arch Work	4	4	4	4.00

Evaluator Level Analysis: Fig. 1 shows illustrates the sources of variation in clarity ratings. This pattern holds across most dimensions, indicating differences in perspective based on role.

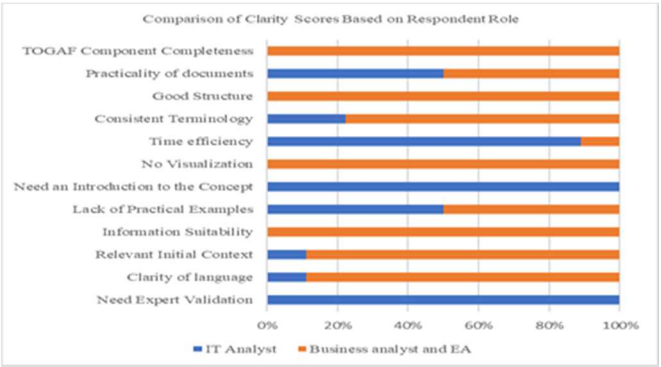


Fig. 1 Comparison of Clarity Scores Based on Respondent Role

Request for Architecture Work rated slightly lower due to context sensitivity. Evaluator differences were minor, Enterprise Architect (P2) and Business Analyst (P3) provided slightly higher clarity ratings than the IT Analyst (P1). Inter-rater reliability yielded **α = 0.86**, indicating high agreement and confirming the instrument's consistency.

B. *Objective Readability Analysis*

The Gunning Fog Index (GFI = 25.03) indicates postgraduate level linguistic complexity. Nevertheless, experts rated clarity at 4.33/5.0, revealing the Complexity Appropriate Paradox despite high textual complexity, experts perceived the content as clear and precise. This paradox suggests that in specialized domains, technical language may enhance clarity rather than hinder comprehension. It reflects alignment between content complexity and expert cognitive models, where domain-specific terminology increases precision and perceived credibility.

C. *Qualitative Findings*

Thematic analysis identified three main strengths and three main limitations, which were consistent across all interviews as in TABLE IV.

TABLE IV. Summary of Thematic Insights

Strengths	Limitations
Comprehensive coverage and TOGAF-aligned structure	Limited contextual adaptation to specific organization.
Significant time efficiency (90–97% reduction)	Absence of organizational history and precedent.
Professional tone and correct EA terminology.	Need for expert review and validation.

Participants described outputs as "complete, structured, and professional," noting that manual preparation would require weeks of effort. However, they emphasized that contextual nuances and institutional knowledge were missing. All agreed that LLMs serve as accelerators, not replacements effective for producing structured drafts that require expert refinement.

D. *Sentiment Analysis*

Overall sentiment distribution Fig. 2 was **47% positive**, **23% neutral**, and **30% negative**, mirroring quantitative findings. Positive sentiment focused on efficiency and structure, while negative feedback targeted context gaps

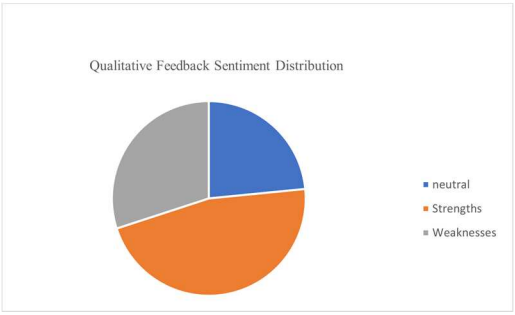


Fig. 2 Qualitative Feedback Sentiment Distribution

E. *Integrated Findings Summary*

Triangulating quantitative and qualitative data reveals a consistent pattern:

1. High overall quality (M = 4.26/5.0)
2. Efficiency transformation time savings up to 97%.
3. Complexity Appropriate clarity for expert audiences.
4. Contextual limitation insufficient adaptation to local nuances.
5. Human AI synergy experts validate and contextualize LLM outputs.

These findings confirm that GPT-4 can generate professional-quality EA artifacts, provided expert oversight ensures contextual fidelity.

F. *Discussion*

Findings confirm that LLMs can generate professional-grade Enterprise Architecture (EA) artifacts (M = 4.33), extending prior AI evaluation research [6], [7], [8] into strategic, context-driven domains. GPT-4 accurately applied TOGAF principles, suggesting internalized domain reasoning, though variation in clarity highlights the continuing role of expert interpretation.

The Complexity–Appropriate Paradox emerges as a key theoretical insight, challenging assumptions that linguistic simplicity ensures clarity [7]. Experts perceived complex outputs as clearer because precise terminology and structured reasoning align with expert cognition and discourse norms. Thus, readability metrics for AI-generated documents must consider audience expertise rather than rely on general standards.

Efficiency gains of 90–97% significantly transform EA practice, compressing Preliminary Phase work [3] from weeks to hours. Such acceleration democratizes architectural development and enables faster iteration but still requires robust validation to ensure contextual and strategic integrity.

Theoretically, this study extends LLM evaluation to strategic artifact generation, introduces the Complexity–Appropriate Paradox, and refines collaboration theory by delineating complementary strengths. Practically, it advocates LLM adoption as expert supervised accelerators combining efficiency with professional rigor. Limitations include reliance on a single model (GPT-4), one organizational case, and a small expert sample (n = 3). Broader, multi-domain evaluations are needed to generalize findings.

V. Conclusion

This study bridges a major gap in Enterprise Architecture (EA) and Artificial Intelligence research by providing the first empirical evaluation of LLM-generated TOGAF Preliminary Phase artifacts. Through a mixed-methods case study in the banking sector, findings show that Large Language Models can produce high-quality EA artifacts (M = 4.33/5.0) while achieving efficiency gains of 90–97%, indicating substantial potential to accelerate EA practices.

LLMs demonstrated consistent capability in generating professional-grade outputs across clarity, completeness, relevance, accuracy, and usability, with all dimensions scoring above 4.0. High accuracy and clarity (M = 4.33) reflect effective assimilation of TOGAF principles and domain logic, validating LLMs' utility in structured documentation.

The study introduces the Complexity–Appropriate Paradox, showing that linguistic complexity (GFI = 25.03) can coexist with perceived clarity among experts. Contrary to traditional readability assumptions, technical precision enhances comprehension in specialized domains. This highlights the need for audience-specific evaluation frameworks for AI-generated professional content.

However, LLMs still exhibit limited contextual sensitivity. Despite structural coherence, outputs lack awareness of organizational nuances, history, and informal stakeholder dynamics. Thus, LLMs serve best as accelerators, reinforcing the Human–AI collaboration model where human expertise ensures contextual and strategic alignment.

Theoretically, this research extends LLM evaluation to strategic, context-dependent domains, introduces the Complexity–Appropriate Paradox, and refines Human–AI collaboration theory by clarifying complementary strengths. Practically, it validates LLM assisted EA development, proposes a five-dimensional assessment framework, and outlines workflows balancing efficiency with contextual accuracy.

For practitioners, LLMs should function as accelerators under expert supervision, supported by prompt engineering, structured review processes, and professional oversight. For organizations, integration should include governance frameworks for validation, robust security and privacy controls, transparent communication on collaboration roles, and platform selection aligned with domain needs.

References

[1] Kornyshova, E., & Deneckère, R. (2022). A proposal of a situational approach for enterprise architecture frameworks: Application to TOGAF. Procedia Computer Science, 207, 3493–3500. https://doi.org/10.1016/j.procs.2022.09.408.

[2] Oberle, M., Yesilyurt, O., Schlereth, A., Risling, M., & Schel, D. (2022). Enterprise IT architecture greenfield design combining IEC 62264 and TOGAF by example of battery manufacturing. Procedia Computer Science, 200, 136–146. https://doi.org/10.1016/j.procs.2022.12.209.

[3] Anggraini, N., Binariswanto, & Legowo, N. (2019). Cloud computing adoption strategic planning using ROCCA and TOGAF 9.2: A study in government agency. Procedia Computer Science, 161, 1316–1324. https://doi.org/10.1016/j.procs.2019.11.247.

[4] Gupta, R., Nair, K., Mishra, M., Ibrahim, B., & Bhardwaj, S. (2024). Adoption and impacts of generative artificial intelligence: Theoretical underpinnings and research agenda. International Journal of Information Management Data Insights, 4(1), 100232. https://doi.org/10.1016/j.jjimei.2024.100232.

[5] Kumar, A., Shankar, A., Hollebeek, L. D., Behl, A., & Lim, W. M. (2025). Generative artificial intelligence (GenAI) revolution: A deep dive into GenAI adoption. Journal of Business Research, 189, 115160. https://doi.org/10.1016/j.jbusres.2024.115160.

[6] Rasool, Z., Maskeliūnas, R., Damaševičius, R., Misra, S., & Singh, S. K. (2024). Evaluating LLMs on document-based QA: Exact answer selection and numerical extraction using CogTale dataset. Natural Language Processing Journal, 8, 100083. https://doi.org/10.1016/j.nlp.2024.100083.

[7] Vaira, L. A., Lechien, J. R., Abbate, V., Allevi, F., Audino, G., Beltramini, G. A., Bergonzani, M., Bolzoni, A., Crimi, S., De Riu, G., Dell'Aversana Orabona, G., Favia, G., Fusetti, S., Galizia, I., Giordano, L., Gobbi, R., Graziano, F., Iaconetta, G., Ionna, F., ... Salzano, G. (2025). Evaluating AI-generated informed consent documents in oral surgery: A comparative study of ChatGPT-4, Bard gemini advanced, and human-written consents. Journal of Cranio-Maxillofacial Surgery, 53(1), 18–23. https://doi.org/10.1016/j.jcms.2024.10.002.

[8] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., ... Wright, R. (2023). 'So what if ChatGPT wrote it?' Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. International Journal of Information Management, 71, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642.

[9] Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Journal of Information Technology Case and Application Research, 25(3), 277–304. https://doi.org/10.1080/15228053.2023.2233814.

[10] Yin, R. K. (2018). Case study research and applications: Design and methods (6th ed.). SAGE Publications.

[11] Doshi, A. R., Bell, J. J., Mirzayev, E., & Vanneste, B. S. (2025). Generative artificial intelligence and evaluating strategic decisions. Strategic Management Journal, 46(3), 583–610. https://doi.org/10.1002/smj.3677.