



Optimization of LayoutLMv3 for DocVQA on Statistical Tables via Fine-Tuning and Semantic Mapping

Optimasi LayoutLMv3 untuk DocVQA pada Tabel Statistik Melalui *Fine-tuning* dan *Semantic Mapping*

D. Achmad Ansari A.^{1*}, Hazriani², Yuyun³

^{1,2}Sistem Komputer, Program Pascasarjana, Universitas Handayani Makassar, Indonesia

³Badan Riset dan Inovasi Nasional, Bandung

E-Mail: ¹dachmadansaria@gmail.com, ²hazriani@handayani.ac.id, ³yuyu010@brin.go.id

Received Mar 10th 2026; Revised Apr 26th 2026; Accepted May 08th 2026; Available Online Jun 20th 2026

Corresponding Author: D. Achmad Ansari A.

Copyright © 2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

This study optimizes the LayoutLMv3 model for the Document Visual Question Answering (DocVQA) task on statistical Tables through a combination of Continual Fine-tuning (CFT), which integrates 2D coordinate normalization and visual patch embedding, and a Semantic Similarity Mapping (SSM) logic layer to handle question variations. A custom dataset was constructed from 398 table images from the BPS publication 'Kabupaten Pinrang Dalam Angka 2025', comprising 19,900 QA pairs. Evaluation results show that the Training Loss decreased from 3.72 to 0.63, ROUGE-1 reached 74.06%, and human validation accuracy increased from 0% (zero-shot) to 41% (CFT) and 51% (CFT+SSM). The highest accuracy of 100% was achieved on tables with three-level headers. This multimodal approach effectively improves the understanding of statistical Table structures, but generalization to complex Tables remains limited by the dataset.

Keywords: *Continual Fine-Tuning, DocVQA, LayoutLMv3, Model Optimization, Statistical Table*

Abstrak

Penelitian ini mengoptimalkan model LayoutLMv3 untuk tugas *Document Visual Question Answering* (DocVQA) pada tabel statistik melalui kombinasi *Continual Fine-tuning* (CFT) yang mengintegrasikan normalisasi koordinat 2D dan *patch embedding* visual, serta lapisan logika *Semantic Similarity Mapping* (SSM) untuk menangani variasi pertanyaan. Dataset mandiri dibangun dari 398 gambar tabel publikasi BPS "Kabupaten Pinrang Dalam Angka 2025" dengan 19.900 pasangan QA. Hasil evaluasi menunjukkan *Training Loss* turun dari 3,72 menjadi 0,63, ROUGE-1 mencapai 74,06%, dan akurasi *human validation* meningkat dari 0% (*zero-shot*) menjadi 41% (CFT) dan 51% (CFT+SSM). Akurasi tertinggi 100% dicapai pada tabel dengan header tiga level. Pendekatan multimodal ini efektif meningkatkan pemahaman struktur tabel statistik, namun generalisasi pada tabel kompleks masih terbatas oleh dataset.

Kata Kunci: *Continual Fine-Tuning, DocVQA, LayoutLMv3, Optimasi Model, Tabel Statistik*

1. PENDAHULUAN

Dalam era digital saat ini, volume dokumen digital baik yang terstruktur maupun tidak terstruktur telah mengalami pertumbuhan eksponensial. Dokumen-dokumen ini, mulai dari laporan keuangan, publikasi statistik, hingga formulir administrasi, mengandung informasi krusial yang seringkali terkunci dalam format yang sulit diakses secara otomatis. Tantangan utama dalam bidang *Document Artificial Intelligence* (Document AI) adalah bagaimana mengekstrak, memahami, dan memanfaatkan informasi dari dokumen-dokumen kompleks ini secara efisien dan akurat [1], [2]. Salah satu tugas yang paling menantang dalam Document AI adalah *Document Visual Question Answering* (DocVQA), yang bertujuan untuk menjawab pertanyaan dalam bahasa alami berdasarkan konten visual dan tekstual dari sebuah gambar dokumen [3].

Tidak seperti VQA pada gambar umum, DocVQA memerlukan pemahaman yang mendalam tidak hanya terhadap teks yang diekstraksi melalui *Optical Character Recognition* (OCR), tetapi juga terhadap

struktur tata letak (*layout*) dokumen, seperti tabel, kolom, dan hierarki informasi [4]. Kompleksitas DocVQA terletak pada sifat multimodal dari data di mana sebuah model harus mampu memproses dan mengintegrasikan informasi dari tiga modalitas utama: teks (konten semantik) [5], tata letak (posisi spasial elemen pada halaman), dan visual (fitur gambar seperti jenis huruf, warna, atau garis tabel) [6], [7]. Publikasi statistik, seperti yang dikeluarkan oleh Badan Pusat Statistik (BPS), merupakan contoh utama dari dokumen yang padat informasi dan sangat bergantung pada struktur tabel. Ekstraksi informasi dari tabel-tabel ini secara manual adalah proses yang lambat, rentan terhadap kesalahan, dan tidak dapat diskalakan.

Perkembangan model *deep learning*, khususnya arsitektur Transformer, telah menunjukkan kemajuan yang signifikan dalam pemrosesan bahasa alami (NLP) dan visi komputer. Model-model seperti LayoutLMv3 telah menetapkan standar baru dalam pemahaman dokumen dengan mengintegrasikan informasi teks, tata letak, dan visual secara bersamaan dalam satu kerangka kerja *pre-training* terpadu [8]. Beberapa penelitian relevan dalam bidang ini antara lain Mathew et al. (2021) yang memperkenalkan *dataset* DocVQA sebagai standar benchmark [3], Huang et al. (2022) yang memperkenalkan LayoutLMv3 sebagai model *pre-training* terpadu [1], [9], serta Chanti et al. (2023) yang mengeksplorasi efektivitas LayoutLMv3 dalam memahami dokumen kompleks melalui tugas VQA [10], [11]. Pendekatan multimodal berbasis *feature fusion transformer* menunjukkan bahwa integrasi fitur visual dan tekstual secara terpadu dapat meningkatkan akurasi klasifikasi dan ekstraksi informasi pada data kompleks [12]. Strategi *fine-tuning* transfer yang efisien terbukti mampu mempercepat adaptasi model terhadap domain spesifik tanpa memulai pelatihan dari awal [13].

Namun demikian, terdapat beberapa keterbatasan kritis pada penelitian-penelitian terdahulu yang perlu dicermati. Pertama, *dataset* benchmark DocVQA yang diperkenalkan oleh Mathew et al. (2021) didominasi oleh dokumen berbahasa Inggris dengan struktur tata letak yang relatif seragam, sehingga kemampuan generalisasinya terhadap dokumen statistik berbahasa Indonesia yang memiliki hierarki tabel kompleks dan multibahasa belum teruji [3]. Kedua, meskipun Chanti et al. (2023) mendemonstrasikan potensi LayoutLMv3 pada tugas VQA, penelitian tersebut tidak mempertimbangkan adaptasi domain spesifik maupun penanganan variasi linguistik pertanyaan, yang merupakan tantangan nyata dalam konteks dokumen pemerintahan berbahasa Indonesia [10]. Ketiga, strategi *fine-tuning* yang ada umumnya diterapkan pada *dataset* generik tanpa mekanisme tambahan untuk menyelaraskan variasi semantik pertanyaan dengan kandidat jawaban dalam tabel, sehingga konsistensi respons model masih rendah ketika menghadapi pertanyaan yang secara struktural berbeda namun bermakna identik [13]. Kesenjangan-kesenjangan inilah yang menjadi motivasi utama penelitian ini.

Berdasarkan uraian tersebut, penelitian ini berfokus pada optimasi model LayoutLMv3 untuk tugas DocVQA pada dokumen tabel statistik berbahasa Indonesia. Kontribusi utama penelitian ini adalah mengusulkan kombinasi *Continual Fine-tuning* (CFT) berbasis *dataset* domain spesifik dengan lapisan logika *Semantic Similarity Mapping* (SSM) sebagai mekanisme kontrol semantik, yang secara eksplisit dirancang untuk mengatasi keterbatasan generalisasi model pada konteks tabel statistik pemerintahan berbahasa Indonesia. Secara spesifik, penelitian ini bertujuan untuk: (1) mengukur kinerja model LayoutLMv3 dalam menghasilkan anotasi tabel publikasi Pinrang Dalam Angka, dan (2) mengevaluasi kinerja model LayoutLMv3 dalam menghasilkan informasi visual pada tabel. Evaluasi dilakukan dalam tiga kondisi, yaitu *zero-shot* (pre-trained), setelah CFT, dan setelah penerapan CFT dengan SSM.

2. BAHAN DAN METODE

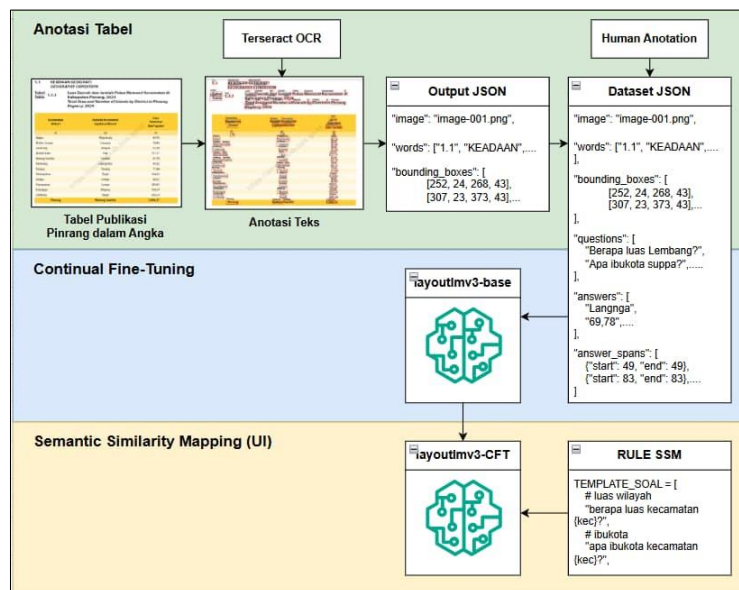
2.1. Rancangan Penelitian

Penelitian ini menggunakan pendekatan eksperimental kuantitatif dengan merancang dan mengevaluasi sistem DocVQA berbasis arsitektur LayoutLMv3 pada dokumen tabel statistik berbahasa Indonesia. Secara keseluruhan, sistem yang dibangun terdiri dari tiga komponen utama yang saling terintegrasi: (1) pembangunan *dataset* berbasis OCR dan human annotation, (2) *Continual Fine-tuning* (CFT) model LayoutLMv3, dan (3) penerapan lapisan logika *Semantic Similarity Mapping* (SSM). Gambaran arsitektur sistem secara keseluruhan ditunjukkan pada *Gambar 1*.

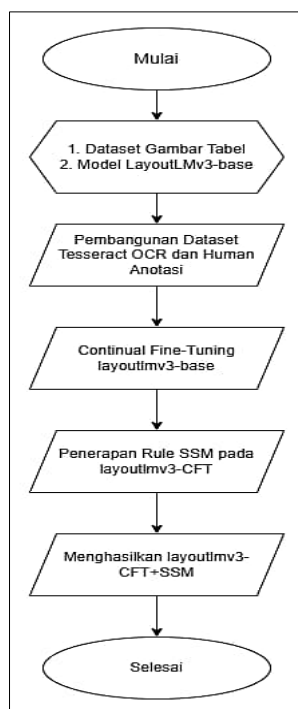
Seperti yang ditunjukkan pada *Gambar 1*, arsitektur sistem terdiri dari tiga tahap besar yang berurutan. Tahap pertama adalah anotasi tabel, di mana gambar tabel statistik dari publikasi BPS diekstraksi menggunakan Tesseract OCR untuk memperoleh informasi teks dan koordinat spasial, kemudian dilengkapi dengan anotasi pasangan pertanyaan-jawaban secara manual oleh peneliti menghasilkan *dataset* JSON. Tahap kedua adalah CFT, di mana *dataset* yang telah dibangun digunakan untuk melatih ulang model LayoutLMv3-base sehingga menghasilkan model LayoutLMv3-CFT yang telah teradaptasi terhadap domain tabel statistik berbahasa Indonesia. Tahap ketiga adalah SSM, di mana lapisan logika SSM diterapkan pada LayoutLMv3-CFT sebagai mekanisme pra-inferensi untuk menormalkan variasi linguistik pertanyaan, menghasilkan sistem akhir LayoutLMv3-CFT+SSM.

Prosedur penelitian secara keseluruhan dirancang secara sistematis untuk membandingkan kinerja model dalam tiga kondisi yaitu *pre-trained* (*zero-shot*), setelah CFT, dan setelah CFT dengan penerapan *rule* SSM. Evaluasi model bahasa modern tidak hanya bergantung pada metrik otomatis, tetapi juga memerlukan

validasi yang mempertimbangkan kualitas semantik dan konsistensi respons [14]. Metrik kuantitatif mencakup analisis *Training Loss*, *Validation Loss*, skor ROUGE [16], dan *Exact Match* [17]. Validasi manusia (*human validation*) oleh 5 mahasiswa Program Studi S2 Sistem Komputer Universitas Handayani Makassar digunakan sebagai parameter final [18]. Alur prosedur penelitian secara rinci ditunjukkan pada Gambar 2.



Gambar 1. Diagram Alur Rancangan Penelitian



Gambar 2. Diagram Alur Rancangan Penelitian

2.2. Dataset Penelitian

Sumber data utama yang digunakan adalah publikasi resmi dari Badan Pusat Statistik (BPS) yaitu "Kabupaten Pinrang Dalam Angka 2025" [19], [20]. *Dataset* penelitian dibuat dari 398 gambar halaman yang berisi tabel-tabel dari publikasi tersebut. Proses pembuatan *dataset* melibatkan dua langkah utama: (a) Ekstraksi OCR menggunakan *Tesseract OCR* untuk mengekstrak semua teks beserta informasi spasialnya berupa koordinat *bounding box* [14], [15], dan (b) Anotasi pertanyaan-jawaban (QA) secara manual sebanyak 50 pertanyaan per gambar, menghasilkan total 19.900 pasangan QA. *Dataset* ini kemudian dibagi menjadi set pelatihan (*train set*) dan set pengujian (*test set*) dengan perbandingan 85:15.

Tesseract OCR dipilih karena kemampuannya mengekstraksi teks sekaligus menghasilkan informasi koordinat spasial (*bounding box*) untuk setiap kata, yang merupakan *input* wajib bagi arsitektur LayoutLMv3. Tesseract versi 5 berbasis arsitektur LSTM menunjukkan akurasi yang memadai untuk dokumen cetak terstruktur seperti tabel statistik BPS [14]. Setiap hasil OCR divalidasi secara manual untuk memastikan konsistensi antara jumlah token kata dan koordinat *bounding box* yang dihasilkan, mengingat ketidakkonsistenan pada tahap ini akan menyebabkan error pada proses tokenisasi multimodal selanjutnya. Contoh *dataset* gambar tabel dan persentase karakteristik tabel *dataset* dapat dilihat pada Gambar 3 dan Tabel 1.

**1.1 KEADAAN GEOGRAFI
GEOGRAPHY CONDITION**

**Tabel 1.1.1 Luas Daerah dan Jumlah Pulau Menurut Kecamatan di Kabupaten Pinrang, 2024
Total Area and Number of Islands by District in Pinrang Regency, 2024**

Kecamatan District	Ibukota Kecamatan Capital of District	Luas Total Area (km ² /sq.km)
(1)	(2)	(3)
Suppa	Majennang	69,78
Mattiro Sompe	Langnga	90,46
Lanrisang	Jampue	65,48
Mattiro Bulu	Bua	127,71
Watang Sawitto	Sawitto	61,76
Paletang	Laleng Bata	46,32
Tiroang	Tiroang	77,44
Patampua	Teppo	124,02
Cempa	Cempa	88,01
Duampua	Lampa	284,42
Batulappa	Bilajeng	164,59
Lembang	Tuppu	696,54
Pinrang	Watang Sawitto	1.896,57

Gambar 3. Contoh *Dataset* Gambar Tabel

Tabel 1. Persentase Karakteristik Tabel *Dataset*

No	Kategori	Jumlah Tabel	Presentase
1	Geografi	8	2,01%
2	Pemerintahan	17	4,27%
3	Kependudukan dan Tenaga Kerja	15	3,77%
4	Masyarakat	115	28,89%
5	Pertanian	118	29,65%
6	Industri	15	3,77%
7	Karir	6	1,51%
8	Listrik dan Air	20	5,03%
9	Perdagangan	56	14,07%
10	Transportasi	4	1,01%
11	Transportasi Komunikasi	2	0,50%
12	Keuangan dan Harga	12	3,02%
13	PDRB	10	2,51%
	Total	398	100%

2.3. Dasar Teori

2.3.1. Arsitektur LayoutLMv3

LayoutLMv3 merupakan model *pre-training* multimodal yang dikembangkan oleh Huang et al. (2022) untuk tugas Document AI [1]. Berbeda dari pendahulunya (*LayoutLM* dan *LayoutLMv2*), LayoutLMv3 menggunakan strategi *Masked Image Modeling* (MIM) dan *Masked Language Modeling* (MLM) secara terpadu dalam satu kerangka *pre-training*, sehingga representasi teks dan visual dipelajari secara simetris tanpa ketergantungan pada CNN eksternal [1], [8]. Arsitektur ini terdiri dari tiga komponen *embedding* utama yang diintegrasikan sebelum diproses oleh *encoder Transformer*:

1. *Text and Layout Embedding*: Setiap token teks direpresentasikan melalui gabungan text token *embedding*, 1D position *embedding*, dan 2D spatial *layout embedding* yang menyandikan posisi geometris token pada halaman dokumen. Representasi ini memungkinkan model memahami hubungan spasial antar elemen teks dalam tabel.
2. *Visual Patch Embedding*: Citra dokumen diubah menjadi urutan patch visual berukuran 16×16 piksel menggunakan linear projection, mengikuti paradigma *Vision Transformer* (ViT). *Patch-patch* ini diperlakukan sebagai token visual yang setara dengan token teks dalam *encoder Transformer*.

3. *Unified Transformer Encoder*: Token teks dan token visual diproses bersama dalam satu *encoder Transformer* yang sama, memungkinkan *cross-modal attention* antara modalitas teks dan visual secara langsung [8], [21]. Untuk tugas DocVQA, LayoutLMv3 dikonfigurasi sebagai model ekstraktif (extractive QA) di mana model memprediksi span jawaban berupa pasangan indeks token awal dan akhir dari teks dokumen yang menjadi jawaban atas pertanyaan yang diberikan.

2.3.2. Optical Character Recognition (OCR) dengan Tesseract

OCR merupakan proses konversi teks dalam gambar menjadi representasi teks digital yang dapat diproses secara komputasional [14], [15]. Tesseract OCR versi 5 menggunakan arsitektur LSTM untuk pengenalan karakter dan menghasilkan output berupa teks beserta koordinat *bounding box* setiap kata dalam format piksel. Koordinat ini merepresentasikan posisi spasial token dalam gambar dan menjadi fondasi bagi *Layout Embedding* pada arsitektur LayoutLMv3. Kualitas hasil OCR sangat menentukan kualitas *dataset QA*, karena kesalahan ekstraksi teks akan berdampak langsung pada ketidakakuratan anotasi jawaban [15].

2.3.3. Semantic Similarity Mapping (SSM)

SSM merupakan mekanisme *rule-based* yang dirancang untuk menangani variasi linguistik pada pertanyaan pengguna. Secara konseptual, SSM bekerja dengan mengukur kemiripan semantik antara pertanyaan yang diajukan dan sekumpulan pertanyaan referensi yang telah terdefinisi, kemudian memetakan pertanyaan *input* ke pertanyaan kanonik yang paling sesuai sebelum diteruskan ke model LayoutLMv3 [22], [23]. Pendekatan ini relevan karena dalam praktik nyata, pertanyaan yang secara semantik identik dapat diformulasikan dengan struktur kalimat yang sangat berbeda. Sebagai contoh, "Berapa jumlah penduduk Kecamatan Suppa?" dan "Sebutkan total penduduk di Suppa!" memiliki makna yang sama namun struktur sintaksis yang berbeda. Tanpa mekanisme normalisasi semantik, model ekstraktif cenderung gagal mengenali kesamaan maksud tersebut dan menghasilkan jawaban yang tidak konsisten [22].

2.4. Konfigurasi Model dan Preprocessing Multimodal

Implementasi DocVQA sebagai tugas multimodal mensyaratkan *preprocessing* terintegrasi pada tiga komponen data. *Preprocessing* teks mencakup validasi konsistensi antara jumlah kata (*words*) dan koordinat spasial (*bounding boxes*), serta proses *flattening* agar setiap pasangan gambar-pertanyaan direpresentasikan sebagai satu sampel tunggal. Setiap token teks diproyeksikan ke dalam ruang vektor sebagai *Text Token Embedding* yang digabungkan dengan *Layout Embedding* dan *Position Embedding* melalui formula [21]:

$$TokenRep_i = TeksEmbed_i + LayoutEmbed_i + PositionEmbed_i \quad (1)$$

Preprocessing layout mengimplementasikan normalisasi koordinat dua dimensi dalam rentang standar 0–1000 menggunakan formula $X_{norm} = (x/width) \times 1000$. Koordinat yang telah dinormalisasi diubah menjadi 2D *Spatial Layout Embedding* yang menyimpan posisi geometris setiap token dalam halaman dokumen [13]. *Preprocessing visual* mengkonversi citra dokumen melalui: (a) resize ke 224×224 piksel sebagai standar *Vision Transformer*, (b) konversi RGB, (c) normalisasi *pixel* standar ImageNet, dan (d) *patch embedding* (visual tokenization) di mana citra dipecah menjadi 196 patch berukuran 16×16 piksel melalui formula [21].

$$VisualPatchEmbed_k = Linear(flatten(Patch_k)) + PositionalEmbed_k \quad (2)$$

Tokenisasi teks menggunakan *tokenizer* RoBERTa yang terintegrasi dalam *pipeline* LayoutLMv3, dengan panjang sekuens maksimum 512 token. Apabila jumlah token melebihi batas tersebut, dilakukan *truncation* dengan strategi *only_second* yang mempertahankan konteks pertanyaan dan memotong konteks dokumen. Pasangan gambar-pertanyaan yang memiliki *answer_span* di luar jangkauan *token* setelah *truncation* dieliminasi dari *dataset* pelatihan untuk menjaga kualitas supervisi. *Hyperparameter* CFT dapat dilihat pada Tabel 2.

Tabel 2. Hyperparameter CFT

Parameter	Nilai
<i>Epoch</i>	10
Learning Rate	5e-5
Batch Size	8
Warmup Ratio	0.1
Optimizer	AdamW

Model dilatih menggunakan optimizer AdamW dengan *linear learning rate scheduler* dan *warmup ratio* 0,1 untuk menstabilkan pelatihan pada tahap awal. Strategi *early stopping* diterapkan dengan memantau

validation loss, di mana pelatihan dihentikan jika tidak terjadi penurunan *loss* selama 3 *epoch* berturut-turut. *Checkpoint model* terbaik disimpan berdasarkan nilai *loss validasi* terendah untuk digunakan pada tahap evaluasi [25].

2.5. Semantic Similarity Mapping (SSM)

Implementasi aturan berbasis SSM berfungsi sebagai lapisan logika tambahan (*rule-based layer*) yang ditempatkan sebelum tahap inferensi penuh LayoutLMv3. SSM bertujuan untuk mengelola variasi linguistik pada pertanyaan serta meningkatkan konsistensi jawaban. Mekanisme ini membantu model dalam mengatasi keterbatasan pemahaman terhadap variasi pola pertanyaan yang secara semantik identik namun secara struktural berbeda [22], [23].

Secara teknis, SSM mengimplementasikan tiga *rule* utama: (1) normalisasi pertanyaan, yaitu penghapusan tanda baca berlebih, penyeragaman huruf kapital, dan penghapusan stopword tidak relevan; (2) pemetaan sinonim domain, yaitu penggantian istilah variasi dengan istilah kanonik yang terdapat dalam *dataset* pelatihan, misalnya "total" → "jumlah" dan "berapa banyak" → "berapa"; dan (3) pencocokan template pertanyaan, yaitu penyelarasan struktur pertanyaan *input* dengan template yang paling mirip secara semantik menggunakan *cosine similarity* pada representasi TF-IDF [22]. SSM berperan sebagai mekanisme kontrol logis pra-inferensi, sedangkan LayoutLMv3 tetap menjadi komponen utama pengambilan jawaban dari dokumen.

3. HASIL DAN DISKUSI

3.1. Kinerja Model Pre-trained (Zero-shot)

Tahap pengujian awal mengevaluasi kemampuan model LayoutLMv3 menjawab pertanyaan secara *zero-shot* menggunakan sepuluh skenario validasi yang terdiri dari 10 pasangan QA. Hasil pengujian menunjukkan bahwa model *pre-trained* menghasilkan akurasi 0% pada seluruh skenario. Seluruh respons yang dihasilkan model hanya berupa pernyataan standar "Jawaban tidak ditemukan di dalam dokumen". Fenomena ini mengindikasikan bahwa tanpa adaptasi spesifik terhadap karakteristik *dataset* QA, model *pre-trained* tidak mampu memahami konteks multimodal pada tabel statistik berbahasa Indonesia [8], [24].

Kegagalan total pada kondisi *zero-shot* ini dapat dijelaskan melalui dua faktor utama. Pertama, model LayoutLMv3 versi *pre-trained* dilatih pada *dataset* DocVQA berbahasa Inggris, sehingga representasi semantik yang dimiliki model tidak kompatibel dengan struktur leksikal dan sintaksis bahasa Indonesia yang digunakan dalam *dataset* penelitian ini. Kedua, tabel statistik BPS memiliki karakteristik *layout* yang sangat spesifik seperti *header* bertingkat, sel yang digabung, dan satuan yang beragam dalam satu tabel yang tidak terwakili dalam distribusi data *pre-training* model [8], [24]. Kondisi ini menegaskan bahwa adaptasi domain melalui *fine-tuning* merupakan prasyarat mutlak sebelum model dapat dioperasikan pada dokumen statistik berbahasa Indonesia. Contoh hasil pengujian LayoutLMv3 *Pre-trained (Zero-shot)* dapat dilihat pada Gambar 4.

The screenshot shows a web application interface for document upload and question answering. The document is titled "DAERAH GEOGRAFI" and "OGRAHY CONDITION". The table displayed is "Luas Daerah dan Jumlah Pulau Menurut Kecamatan di Kabupaten Pinrang, 2024" (Total Area and Number of Islands by District in Pinrang Regency, 2024). The table has three columns: Kecamatan (District), Ibu Kota Kecamatan (District Capital), and Luas (km²) (Island Area). The table lists 10 districts and their respective island areas. The interface also shows a question input field with the text "Berapa luas kecamatan suppa?" and a model response area that says "Jawaban tidak ditemukan di dalam dokumen."

Kecamatan District	Ibu Kota Kecamatan Capital of District	Luas Island Area (km²)
Suppa	Makmurung	65,70
Makmur Sanga	Lungga	86,46
Lundang	Jatene	63,40
Makmur Bala	Pala	117,21
Makmur Sanga	Sandir	61,76
Pakurung	Lutung Bala	46,72
Tungga	Tungga	77,64
Pakurung	Tungga	134,62
Gempa	Gempa	88,01
Pakurung	Lundang	214,42
Pakurung	Makmur	144,24
Lundang	Tungga	616,54
Pinrang	Makmur Sanga	1.094,57

Gambar 4. Contoh Hasil Pengujian LayoutLMv3 Pre-trained (Zero-shot)

3.2. Kinerja Model Setelah Continual Fine-tuning (CFT)

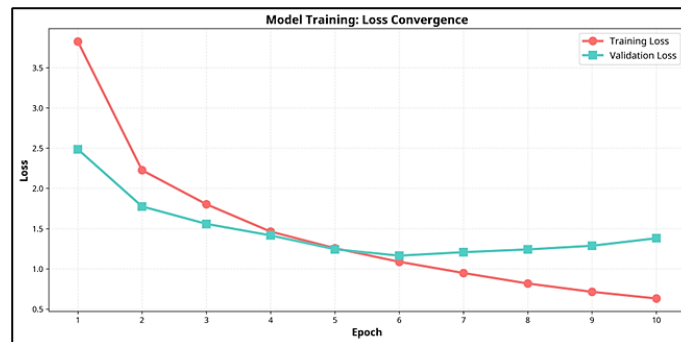
Proses CFT pada LayoutLMv3 (microsoft/LayoutLMv3-base) menunjukkan peningkatan signifikan [1], [10]. Analisis kurva pelatihan menunjukkan *Training Loss* menurun konsisten dari 3,829 pada *epoch* 1 hingga 0,630 pada *epoch* 10. *Validation loss* awalnya menurun dari 2,487 (*epoch* 1) hingga titik terendah 1,163 pada *epoch* 6, namun mulai meningkat kembali sejak *epoch* 7 (1,206) hingga *epoch* 10 (1,379),

mengindikasikan terjadinya *overfitting* [25]. Berdasarkan strategi *early stopping*, *epoch* 6 dipilih sebagai checkpoint model final.

Tren pada Tabel 3 dan Gambar 5 memperlihatkan pola yang konsisten: *Training Loss* menurun secara monoton di setiap *epoch*, sementara *validation loss* mencapai titik optimal pada *epoch* 6 sebelum menunjukkan tren divergen. Pola divergensi antara *Training Loss* dan *validation loss* mulai *epoch* 7 merupakan indikator klasik *overfitting*, di mana model mulai menghafal pola pelatihan alih-alih melakukan generalisasi [25]. Pemilihan *epoch* 6 sebagai checkpoint final didasarkan pada prinsip bias-variance tradeoff, di mana model pada titik ini memiliki keseimbangan optimal antara kemampuan fitting terhadap data pelatihan dan kemampuan generalisasi terhadap data uji.

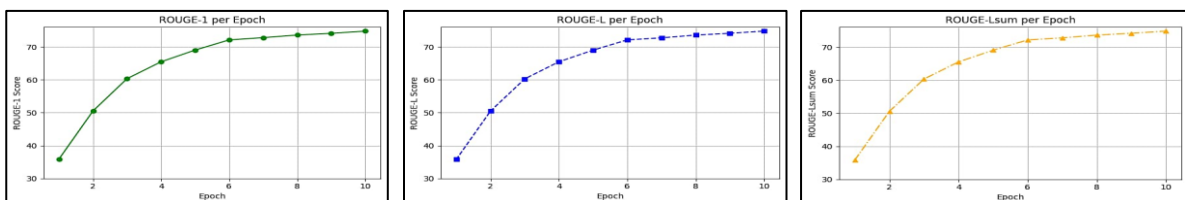
Tabel 3. Tren *Training Loss* dan *Validation Loss* per *Epoch*

<i>Epoch</i>	<i>Training Loss</i>	<i>Validation Loss</i>	ROUGE-1	ROUGE-L	<i>Rouge-Lsum</i>
1	3,829072	2,487014	39,888204	10,419199	35,781730
2	2,228497	1,775924	55,018694	10,018694	50,490065
3	1,603297	1,666880	60,327400	16,767349	60,245646
4	1,402895	1,415446	65,450098	20,406098	65,407162
5	1,204876	1,242815	69,070290	18,527625	68,952096
6	1,087159	1,163142	72,000272	21,274398	72,113060
7	0,947003	1,205672	72,790167	21,253991	72,739543
8	0,816794	1,248058	73,575169	21,518300	73,560779
9	0,719076	1,286523	74,085521	21,685208	74,142195
10	0,630225	1,379966	74,785750	22,334655	74,750572



Gambar 5. Kurva *Training Loss* dan *Validation Loss*

Pada *epoch* 6, model mencatat skor ROUGE-1 sebesar 69,01%, ROUGE-L sebesar 68,99%, dan ROUGE-Lsum sebesar 69,04%. Meskipun skor ini tidak setinggi *epoch* berikutnya, nilai tersebut merupakan yang paling representatif karena diperoleh pada kondisi model yang belum mengalami *overfitting*. Skor ROUGE-1 dan ROUGE-L yang berada di kisaran 69% menunjukkan bahwa pendekatan ekstraktif cukup efektif untuk *dataset* ini di mana jawaban dapat ditemukan secara langsung dalam teks tabel statistik [16]. Adapun pada *epoch* 9, model mencapai skor ROUGE-1 tertinggi sebesar 74,06% dan ROUGE-L sebesar 74,15%, yang digunakan sebagai angka pelaporan akhir dengan mempertimbangkan bahwa evaluasi ROUGE dilakukan pada keseluruhan data uji secara *offline* setelah pelatihan selesai. Kurva pengujian ROUGE dapat dilihat pada Gambar 6.



Gambar 6. Kurva Pengujian ROUGE Score

Evaluasi Exact Match (EM) (lihat Gambar 7) menunjukkan nilai sebesar 63,18% dari total 2.985 sampel data uji [17]. Analisis prediksi menunjukkan model berhasil pada jawaban bertipe numerik dan entitas pendek, sementara kegagalan umumnya disebabkan oleh kebocoran konteks pertanyaan ke dalam span jawaban, prediksi di luar konteks, dan pengambilan span yang terlalu lebar.

HASIL EVALUASI MODEL	
Total sampel test :	2985
Exact Match (EM) :	63.18%

Gambar 7. Hasil Evaluasi Exact Match

Hasil *human validation* setelah CFT menunjukkan variasi akurasi antara 10% hingga 80% pada berbagai skenario pengujian, dengan rata-rata akurasi 45% [18].

3.2.1. Analisis Error Model CFT

Guna memahami kelemahan model secara lebih sistematis, dilakukan analisis terhadap pola kesalahan prediksi pada data uji. Berdasarkan inspeksi kualitatif terhadap sampel prediksi yang salah, ditemukan empat kategori error utama seperti pada Tabel 4.

Kebocoran konteks pertanyaan merupakan error terbesar dan terjadi karena arsitektur LayoutLMv3 sebagai model ekstraktif memproses pertanyaan dan konteks dokumen dalam satu sekuens gabungan. Ketika pertanyaan dan jawaban memiliki irisan leksikal yang tinggi misalnya pertanyaan "Berapa luas wilayah Kecamatan Suppa?" dengan jawaban "95,46" model kadang memperluas span jawaban hingga mencakup token dari pertanyaan karena skor attention yang tinggi pada token yang berulang. Prediksi di luar konteks dan *over-extraction* umumnya muncul pada tabel dengan 6 kolom atau lebih, di mana banyaknya token numerik yang serupa membuat model kesulitan membedakan nilai yang relevan. Kesalahan pemilihan kolom secara dominan terjadi pada tabel dengan header bertingkat dua level atau lebih, karena model belum sepenuhnya memahami hierarki kolom yang direpresentasikan melalui *layout embedding* [25], [26].

Tabel 4. Kategori dan Distribusi Error Model CFT

No	Kategori Error	Deskripsi	Estimasi Proporsi
1	Kebocoran konteks pertanyaan (<i>question leakage</i>)	Model menyertakan sebagian kata dari pertanyaan ke dalam span jawaban, menghasilkan jawaban yang mengandung fragmen pertanyaan itu sendiri	~35%
2	Prediksi di luar konteks (<i>out-of-context span</i>)	Model memilih span dari bagian tabel yang tidak relevan dengan pertanyaan, umumnya terjadi pada tabel dengan banyak baris dan kolom sejenis	~28%
3	Span terlalu lebar (<i>over-extraction</i>)	Model mengekstrak rentang token yang terlalu panjang, menyertakan kata-kata yang tidak termasuk bagian dari jawaban yang benar	~22%
4	Kesalahan pemilihan kolom (<i>column confusion</i>)	Model mengenali baris yang benar namun mengambil nilai dari kolom yang salah, khususnya pada tabel dengan header bertingkat atau kolom berurutan bertipe serupa	~15%

3.3. Penerapan Semantic Similarity Mapping (SSM)

Penerapan *rule-based* SSM sebagai lapisan logika tambahan menunjukkan peningkatan akurasi pada beberapa skenario pengujian. Tabel 3 menyajikan perbandingan akurasi *human validation* antara ketiga kondisi model.

Tabel 5. Perbandingan Akurasi Pengujian Model

Skenario Pengujian	Pre-trained (%)	CFT (%)	CFT+SSM (%)
Skenario 1 - Tabel 3 kolom sederhana	0	70	90
Skenario 2 - Header bertingkat 2 level, 5 kolom	0	50	70
Skenario 3 - Tabel 4 kolom multi-satuan	0	20	30
Skenario 4 - Tabel lebar 6 kolom multi-baris	0	50	50
Skenario 5 - Header bertingkat 2 level, 7 kolom	0	20	30
Skenario 6 - Tabel baru 3 kolom sederhana	0	70	80
Skenario 7 - Tabel baru header bertingkat 2 level	0	30	40
Skenario 8 - Tabel baru 4 kolom sederhana	0	10	10
Skenario 9 - Tabel baru lebar 6 kolom	0	50	50
Skenario 10 - Tabel lanjutan header 3 level	0	80	100
Rata-rata Akurasi	0	45	55

Dari Tabel 5 terlihat bahwa penerapan SSM meningkatkan akurasi rata-rata dari 45% (CFT saja) menjadi 55% (CFT+SSM), dengan delta peningkatan rata-rata sebesar +10 poin persentase. Performa tertinggi dicapai pada skenario kesepuluh (100%), yaitu tabel lanjutan dengan header bertingkat tiga level,

dan skenario pertama (90%) untuk tabel sederhana 3 kolom. Akurasi terendah terjadi pada skenario kedelapan (10%) yaitu tabel baru 4 kolom, menunjukkan keterbatasan generalisasi model pada tabel yang belum dilihat sebelumnya [26], [27].

Pola peningkatan akurasi pasca-SSM dapat dikelompokkan ke dalam tiga kategori berdasarkan besaran delta yang tercatat. Peningkatan signifikan (+20 poin) terjadi pada skenario 1, 2, dan 10 ketiganya merupakan tabel yang strukturnya telah terwakili dalam data pelatihan, sehingga normalisasi semantik oleh SSM berhasil menjembatani variasi pertanyaan dengan pola yang sudah dipelajari model. Peningkatan moderat (+10 poin) pada skenario 3, 5, 6, dan 7 menunjukkan bahwa SSM memberikan manfaat parsial pada tabel dengan kompleksitas menengah. Tanpa peningkatan (0 poin) pada skenario 4, 8, dan 9 mengindikasikan bahwa kegagalan model pada skenario tersebut bukan disebabkan oleh variasi linguistik pertanyaan, melainkan oleh keterbatasan fundamental model dalam memahami struktur tabel yang belum terlihat selama pelatihan sebuah permasalahan yang tidak dapat diatasi oleh mekanisme SSM semata.

3.3.1. Penerapan *Semantic Similarity Mapping* (SSM)

Meskipun SSM berhasil meningkatkan akurasi rata-rata, analisis kualitatif terhadap kasus kegagalan yang tersisa mengungkapkan dua kategori error residual yang dominan. Pertama, kegagalan generalisasi struktural model masih gagal pada tabel baru yang memiliki jumlah kolom atau pola header yang tidak terwakili dalam distribusi data pelatihan. Hal ini tercermin pada skenario 8 (tabel baru 4 kolom sederhana) yang hanya mencapai 10% meskipun setelah penerapan SSM, padahal skenario 1 dengan struktur serupa (3 kolom sederhana dari data pelatihan) mencapai 90%. Kesenjangan ini mengindikasikan bahwa model sangat sensitif terhadap distribusi data pelatihan dan memiliki kapasitas generalisasi lintas-tabel yang terbatas [26]. Kedua, kegagalan pada tabel lebar skenario 4 dan 9 yang melibatkan tabel dengan 6 kolom tidak menunjukkan peningkatan pasca-SSM, sehingga mengonfirmasi bahwa permasalahan *column confusion* pada tabel lebar merupakan keterbatasan arsitektur yang lebih mendasar, bukan sekadar permasalahan variasi linguistik [27].

Secara keseluruhan, integrasi CFT dan SSM telah berhasil diterapkan dan menunjukkan arah peningkatan kinerja yang signifikan. Model mampu mempelajari cara mengidentifikasi dan mengekstrak informasi relevan dari tabel statistik. Namun, pola error yang ditemukan khususnya kegagalan generalisasi struktural dan kolom *confusion* pada tabel lebar menunjukkan bahwa peningkatan lebih lanjut memerlukan strategi augmentasi data yang lebih agresif dan kemungkinan pendekatan arsitektur yang lebih mampu merepresentasikan hierarki tabel secara eksplisit.

3.4. Diskusi Hasil Penelitian

Hasil penelitian secara keseluruhan menunjukkan trajektori peningkatan yang konsisten dari kondisi *zero-shot* (0%) menuju CFT (45%) dan CFT+SSM (55%). Pola peningkatan ini tidak bersifat linear, melainkan mencerminkan kontribusi yang berbeda dari setiap komponen pendekatan yang diusulkan.

Pada kondisi *zero-shot*, kegagalan total model bukan sekadar mencerminkan ketidakmampuan model dalam memahami bahasa Indonesia, tetapi lebih mendasar lagi: model *pre-trained* tidak memiliki representasi semantik yang sesuai untuk domain statistik pemerintahan dengan karakteristik tabel bertingkat dan nilai numerik yang padat. Hal ini sejalan dengan temuan Hegde et al. (2023) yang menunjukkan bahwa model berbasis bahasa besar pun mengalami degradasi performa signifikan ketika diterapkan pada dokumen gambar tanpa adaptasi domain, terutama ketika dokumen tersebut mengandung struktur tabel kompleks [24].

Peningkatan dari *zero-shot* ke CFT sebesar 45 poin persentase merupakan bukti konkret efektivitas transfer learning berbasis domain spesifik. Penurunan *Training Loss* dari 3,829 menuju 0,630 dan stabilisasi *validation loss* di kisaran 1,163 pada *epoch* 6 mengindikasikan bahwa model berhasil mempelajari pola distribusi token dan *layout* yang khas dari tabel statistik BPS. Skor ROUGE-1 sebesar 74,06% dan Exact Match sebesar 63,18% menunjukkan bahwa untuk pertanyaan dengan jawaban berupa entitas numerik pendek yang mendominasi *dataset* BPS model ekstraktif berbasis LayoutLMv3 mampu melakukan ekstraksi dengan tingkat presisi yang memadai. Pola ini konsisten dengan karakteristik tugas *extractive QA*, di mana jawaban dapat ditemukan secara verbatim dalam teks dokumen tanpa memerlukan kemampuan reasoning atau parafrase [16], [17].

Peningkatan tambahan dari CFT ke CFT+SSM sebesar 10 poin persentase, meskipun lebih kecil secara absolut, memiliki signifikansi praktis yang penting. SSM berhasil mengatasi sumber kegagalan yang berbeda dari yang ditangani CFT: jika CFT mengatasi permasalahan adaptasi domain dan representasi fitur, maka SSM mengatasi permasalahan variasi linguistik pada level pertanyaan melalui tiga mekanisme berurutan. Pertama, normalisasi leksikal menyeragamkan variasi ortografis dan morfologis seperti perbedaan penggunaan huruf kapital, tanda baca, dan variasi afiks bahasa Indonesia yang tidak mengubah makna. Kedua, pemetaan sinonim domain menggantikan istilah variasi dengan istilah kanonik yang dominan dalam data pelatihan, sehingga pertanyaan inferensi memiliki distribusi leksikal yang lebih dekat dengan distribusi data pelatihan. Ketiga, pencocokan template memastikan bahwa struktur sintaksis pertanyaan selaras dengan pola pertanyaan yang paling sering dilihat model selama pelatihan [22], [23].

Peningkatan yang paling dramatis terjadi pada skenario 10 (tabel lanjutan header 3 level, dari 80% menjadi 100%), yang mengindikasikan bahwa pada tabel-tabel yang strukturnya sudah dipelajari model dengan baik, normalisasi semantik pertanyaan oleh SSM mampu mengoptimalkan pemanfaatan kapasitas model secara penuh. Sebaliknya, SSM tidak memberikan peningkatan pada skenario 4, 8, dan 9, karena akar permasalahan pada skenario-skenario tersebut terletak pada keterbatasan representasi *layout* tabel yang lebar dan kompleks, bukan pada variasi linguistik pertanyaan sebuah permasalahan yang berada di luar jangkauan mekanisme SSM. Untuk memposisikan hasil penelitian ini dalam konteks literatur yang lebih luas, *Table 6* menyajikan perbandingan dengan penelitian-penelitian terdahulu yang relevan dalam domain DocVQA.

Perbandingan dengan studi-studi pada Tabel 6 perlu dilakukan dengan hati-hati mengingat perbedaan *dataset*, metrik, dan karakteristik dokumen yang mendasar. Skor ROUGE-1 sebesar 74,06% yang dicapai penelitian ini berada dalam rentang yang kompetitif dibandingkan dengan model-model serupa pada *dataset* berbahasa Inggris, mengingat *dataset* penelitian ini dibangun dari nol tanpa ketersediaan data pelatihan tambahan. Penelitian Chanti et al. (2023) yang menggunakan arsitektur LayoutLMv3 yang sama pada dokumen umum berbahasa Inggris mencapai akurasi 72,4%, sementara penelitian ini pada tabel statistik berbahasa Indonesia yang secara *inherent* lebih kompleks secara struktural mencapai *human validation accuracy* 55% dengan CFT+SSM [10]. Selisih ini mencerminkan tantangan tambahan yang ditimbulkan oleh kompleksitas tabel statistik pemerintahan dan keterbatasan ukuran *dataset* yang dibangun secara mandiri. Penelitian Huang et al. (2022) yang mencapai ANLS 83,4% menggunakan *dataset* yang jauh lebih besar dan beragam [1], menegaskan bahwa ukuran dan keragaman *dataset* merupakan faktor determinan utama dalam kinerja model DocVQA.

Tabel 6. Perbandingan Hasil dengan Penelitian Terdahulu

Penelitian	Model	<i>Dataset</i>	Metrik Utama	Hasil
Mathew et al. (2021) [3]	BERT-based	DocVQA (Eng)	ANLS	67,0%
Chanti et al. (2023) [10]	LayoutLMv3	Dokumen umum (Eng)	Accuracy	72,4%
Huang et al. (2022) [1]	LayoutLMv3	DocVQA benchmark	ANLS	83,4%
Penelitian ini (CFT)	LayoutLMv3	Tabel BPS Pinrang (Id)	ROUGE-1 / EM	74,06% / 63,18%
Penelitian ini (CFT+SSM)	LayoutLMv3	Tabel BPS Pinrang (Id)	Human Validation	55%

3.5. Analisis Keterbatasan Penelitian

Meskipun hasil penelitian menunjukkan peningkatan yang signifikan, terdapat tiga faktor keterbatasan yang perlu diakui secara sistematis untuk memberikan gambaran yang jujur tentang batas kemampuan pendekatan yang diusulkan. Terdapat tiga keterbatasan utama yang perlu diakui dalam penelitian ini.

Pertama, keterbatasan ukuran dan keragaman *dataset* meskipun 19.900 pasangan QA dari 398 gambar tabel merupakan upaya yang substansial, ukuran ini masih jauh di bawah standar internasional, sehingga berdampak pada kemampuan generalisasi model terhadap tabel baru. Ketidakseimbangan distribusi kategori *dataset* dan keterbatasan sumber pada satu publikasi BPS dari satu kabupaten juga membatasi kemampuan generalisasi model ke publikasi BPS dari daerah lain.

Kedua, keterbatasan representasi struktur tabel kompleks LayoutLMv3 merepresentasikan *layout* melalui koordinat *bounding box* 2D yang tidak secara eksplisit memodelkan relasi hierarkis antar header kolom, sehingga akurasi model menurun seiring bertambahnya kompleksitas struktur tabel [26]. Ketiga, keterbatasan kualitas OCR sebagai sumber noise kesalahan ekstraksi Tesseract OCR pada karakter khusus dan format angka lokal yang tidak terdeteksi akan terbawa sebagai noise dalam *dataset* pelatihan, yang proporsinya belum dikuantifikasi secara eksplisit dalam penelitian ini [14], [15].

4. KESIMPULAN

Penelitian ini telah berhasil mendemonstrasikan efektivitas kombinasi *Continual Fine-tuning* (CFT) dan *Semantic Similarity Mapping* (SSM) pada arsitektur LayoutLMv3 untuk tugas *Document Visual Question Answering* (DocVQA) pada dokumen tabel statistik berbahasa Indonesia. Kontribusi utama penelitian ini terletak pada dua hal: pertama, pembangunan *dataset* DocVQA domain spesifik berbahasa Indonesia dari publikasi BPS secara mandiri menggunakan Tesseract OCR dan human annotation; dan kedua, pengusulan kombinasi CFT+SSM sebagai strategi adaptasi bertingkat yang mampu mengatasi dua sumber kegagalan yang berbeda secara simultan ketidaksesuaian domain pada level representasi fitur dan variasi linguistik pada level pertanyaan.

Secara kuantitatif, penelitian ini membuktikan bahwa model LayoutLMv3 *pre-trained* tanpa adaptasi sama sekali tidak mampu menangani dokumen tabel statistik berbahasa Indonesia dengan akurasi 0% pada seluruh skenario. Setelah diterapkan CFT menggunakan *dataset* yang dibangun dari 398 gambar tabel dan 19.900 pasangan QA, kinerja model meningkat secara signifikan dengan *Training Loss* turun dari 3,717 menjadi 0,625, skor ROUGE-1 sebesar 74,06%, ROUGE-L sebesar 74,15%, Exact Match sebesar 63,18%,

serta rata-rata akurasi *human validation* sebesar 45%. Penambahan lapisan logika SSM selanjutnya meningkatkan rata-rata akurasi *human validation* dari 45% menjadi 55%, dengan performa terbaik pada tabel sederhana 3 kolom (90%) dan tabel lanjutan header bertingkat tiga level (100%).

Namun demikian, penelitian ini juga mengidentifikasi batas kemampuan pendekatan yang diusulkan: model masih mengalami keterbatasan generalisasi pada tabel baru yang tidak terwakili dalam distribusi data pelatihan, khususnya tabel lebar dengan 6 kolom atau lebih yang stagnan di akurasi 50% bahkan setelah penerapan SSM. Keterbatasan ini, bersama dengan ukuran *dataset* yang masih terbatas dan ketergantungan pada kualitas OCR, menjadi agenda utama yang perlu ditangani pada penelitian selanjutnya.

Berdasarkan temuan dan keterbatasan tersebut, penelitian selanjutnya disarankan untuk mengeksplorasi empat arah pengembangan berikut. Pertama, perluasan dan diversifikasi *dataset* dengan memperluas cakupan sumber data ke publikasi BPS dari berbagai kabupaten dan provinsi di Indonesia, serta mempertimbangkan augmentasi data berbasis parafrase otomatis untuk memperkaya variasi pasangan QA tanpa memerlukan anotasi manual tambahan yang ekstensif. Kedua, pendekatan model hybrid ekstraktif-generatif dengan mengintegrasikan kemampuan ekstraksi presisi LayoutLMv3 dengan kemampuan *parafrase* dan *reasoning* model generatif seperti T5 atau GPT berbasis bahasa Indonesia, sehingga model tidak hanya mampu mengekstrak span jawaban secara verbatim tetapi juga mampu menghasilkan jawaban yang lebih natural dan kontekstual. Ketiga, arsitektur yang lebih eksplisit dalam merepresentasikan struktur tabel seperti *Table Transformer* atau TAPAS yang dirancang khusus untuk memahami relasi sel, baris, dan kolom secara eksplisit, sebagai alternatif LayoutLMv3 yang merepresentasikan struktur tabel secara implisit melalui koordinat spasial. Keempat, peningkatan *pipeline* OCR melalui penerapan OCR confidence threshold, validasi silang antar-annotator, dan eksplorasi *engine* OCR alternatif yang lebih akurat untuk dokumen statistik berbahasa Indonesia, guna meminimalkan noise pada data pelatihan yang berdampak langsung pada kualitas supervisi model.

REFERENSI

- [1] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei, "LayoutLMv3: *Pre-training* for Document AI with Unified Text and Image Masking," in Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Association for Computing Machinery, Oct. 2022, pp. 4083–4091. doi: 10.1145/3503161.3548112.
- [2] K. Kapula, "Intelligent Document Processing: The New Frontier of Automation," International Journal of Intelligent Systems and Applications in Engineering, vol. 10, no. 2, pp. 215–223, 2022. doi: 10.18201/ijisae.2022.272.
- [3] M. Mathew, D. Karatzas, and C. V. Jawahar, "DocVQA: A *Dataset* for VQA on Document Images," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Jan. 2021, pp. 2200–2209. doi: 10.1109/WACV48630.2021.00225.
- [4] N. D. Huynh, M. R. Bouadjene, S. Aryal, I. Razzak, and H. Hacid, "Visual Question Answering: From Early Developments to Recent Advances A Survey," arXiv preprint, arXiv:2501.03939, Jan. 2025. [Online]. Available: <http://arxiv.org/abs/2501.03939>.
- [5] T. Sumner and M. Marlino, "Digital Libraries and Educational Practice: Opportunities and Challenges," in Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '04), Association for Computing Machinery, Jun. 2004, pp. 170–178. doi: 10.1145/996350.996389.
- [6] B. S. U. Kim, J. Kim, D. Lee, and B. Jang, "Visual Question Answering: A Survey of Methods, *Datasets*, Evaluation, and Challenges," ACM Computing Surveys, vol. 57, no. 10, pp. 1–42, May 2025. doi: 10.1145/3728635.
- [7] R. Tanaka, K. Nishida, K. Nishida, T. Hasegawa, I. Saito, and K. Saito, "SlideVQA: A *Dataset* for Document Visual Question Answering on Multiple Images," in Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI-23), AAAI Press, Feb. 2023, pp. 13636–13645. doi: 10.1609/aaai.v37i11.26598.
- [8] C. Barboule, B. Piwowarski, and Y. Chabot, "Survey on Question Answering over Visually Rich Documents: Methods, Challenges, and Trends," arXiv preprint, arXiv:2501.02235, Mar. 2025. [Online]. Available: <http://arxiv.org/abs/2501.02235>.
- [9] Q. Peng, Y. Tang, Y. Xu, H. Zhao, G. Lv, B. Sun, Y. Lv, and Y. Wei, "ERNIE-Layout: Layout Knowledge Enhanced *Pre-training* for Visually-rich Document Understanding," in Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Dec. 2022, pp. 3744–3756. doi: 10.18653/v1/2022.findings-emnlp.274.
- [10] S. Chanti, P. Harika, V. Venkata Vinay, P. Guna Sri Charan, and Sk. Ajad, "DocumentQA: Leveraging LayoutLMv3 for Next-Level Question Answering," in Proceedings of the 2023 International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS), IEEE, Oct. 2023, pp. 1–8. doi: 10.1109/ICSSAS57918.2023.10331881.

- [11] M. Yang, A. Anantharaman, Z. Kitowski, and D. C. Robert, "Graph Relation Transformer: Incorporating Pairwise Object Features into the Transformer Architecture," arXiv preprint, arXiv:2111.06075, Nov. 2021. [Online]. Available: <http://arxiv.org/abs/2111.06075>.
- [12] J. Park, J. Cho, S. Han, and K. Kim, "OCR-Augmented GPT for Accurate Text Extraction in Industrial Environments," IEEE Access, vol. 13, pp. 136226–136236, 2025. doi: 10.1109/ACCESS.2025.3594682.
- [13] G. A. Pereira and M. Hussain, "A Review of Transformer-Based Models for Computer Vision Tasks: Capturing Global Context and Spatial Relationships," arXiv preprint, arXiv:2408.15178, Aug. 2024. [Online]. Available: <http://arxiv.org/abs/2408.15178>.
- [14] A. R. I. Ponggohong, G. C. Rorimpandey, and V. V. Maswonggo, "Penerapan *Optical Character Recognition* Tesseract dan Gemini pada Sistem Buku Tamu Digital Berbasis Web," Jurnal Informatika dan Teknologi Komputer, vol. 6, no. 1, pp. 45–53, 2025. doi: 10.46808/jitkom.v6i1.312.
- [15] S. Gull, N. Ahmed, and S. A. Sheikh, "Deep learning-Driven OCR System for Brahui Printed Text: Bridging the Digital Gap in Low-Resource Language Processing," Liberal Journal of Language and Literature Review, vol. 3, no. 1, pp. 88–102, 2025. [Online]. Available: <https://llrjournal.com/index.php/11>.
- [16] A. Auriemma Citarella, M. Barbella, M. G. Ciobanu, F. De Marco, L. Di Biasi, and G. Tortora, "Assessing the Effectiveness of ROUGE as Unbiased Metric in Extractive vs. Abstractive Summarization Techniques," Journal of Computational Science, vol. 87, art. no. 102571, May 2025. doi: 10.1016/j.jocs.2025.102571.
- [17] M. Gabburo, S. Garg, R. Koncel-Kedziorski, and A. Moschitti, "SQUARE: Automatic Question Answering Evaluation using Multiple Positive and Negative References," in Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Association for Computational Linguistics, Jul. 2023, pp. 441–452. doi: 10.18653/v1/2023.acl-short.49.
- [18] L. M. Amugongo, P. Mascheroni, S. Brooks, S. Doering, and J. Seidel, "Retrieval Augmented Generation for Large Language Models in Healthcare: A Systematic Review," PLOS Digital Health, vol. 4, no. 6, art. no. e0000877, Jun. 2025. doi: 10.1371/journal.pdig.0000877.
- [19] BPS Kabupaten Pinrang, Kabupaten Pinrang Dalam Angka 2025, vol. XXI. Pinrang: BPS Kabupaten Pinrang, 2025. [Online]. Available: <https://pinrangkab.bps.go.id>.
- [20] T. Perdana, K. Kusnandar, H. H. Perdana, F. R. Hermiatin, A. H. Sadeli, and R. Saville, "Behavioural Drivers of Thiamethoxam Use among Perennial and Annual Crop Farmers in Indonesia: An Extended Theory of Planned Behavior Approach," Cogent Food & Agriculture, vol. 12, no. 1, art. no. 2615202, Dec. 2026. doi: 10.1080/23311932.2026.2615202.
- [21] Y. Xu, M. Xu, F. Lv, L. Cui, G. Wei, N. Wang, F. Lu, B. Florencio, C. Zhang, W. Che, M. Zhang, and L. Zhou, "LayoutLMv2: Multi-modal *Pre-training* for Visually-rich Document Understanding," in Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021), Association for Computational Linguistics, Aug. 2021, pp. 2579–2591. doi: 10.18653/v1/2021.acl-long.201.
- [22] D. Caffagni, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara, "Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval," in Proceedings of the 2022 ACM International Conference on Multimedia Retrieval (ICMR '22), Association for Computing Machinery, Jun. 2022, pp. 180–188. doi: 10.1145/3512527.3531391.
- [23] C. Gao, Q. Zeng, M. Yin, and J. Sun, "Structured Multimodal Attentions for TextVQA," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 2, pp. 1594–1607, Feb. 2023. doi: 10.1109/TPAMI.2022.3152766.
- [24] N. Hegde, S. Paul, G. Madan, and G. Aggarwal, "Analyzing the Efficacy of an LLM-Only Approach for Image-based Document Question Answering," arXiv preprint, arXiv:2309.14389, Sep. 2023. [Online]. Available: <http://arxiv.org/abs/2309.14389>.
- [25] R. Pal, S. Kar, D. K. Prasad, and A. A. Sekh, "VisionAidQA: Advancing Visual Question Answering for the Visually Impaired," ACM Computing Surveys, vol. 57, no. 3, art. no. 3788861, Jul. 2026. doi: 10.1145/3788861.
- [26] X. Wu, W. Hu, X. Hu, and E. Chang, "A Region-based Document VQA," in Proceedings of the 30th ACM International Conference on Multimedia (MM '22), Association for Computing Machinery, Oct. 2022, pp. 4909–4920. doi: 10.1145/3503161.3548172.
- [27] J. Nockels, P. Gooding, S. Ames, and M. Terras, "Understanding the Application of Handwritten Text Recognition Technology in Heritage Contexts: A Systematic Review of Transkribus in Published Research," Archival Science, vol. 22, no. 3, pp. 367–392, Sep. 2022. doi: 10.1007/s10502-022-09397-0.