

Integrasi BiLSTM-GRU dan Retrieval-Augmented Generation untuk Prediksi Harga Saham Perbankan

^{1*}Roni Halim Saputra, ²Eri Zuliarso

^{1*,2}Magister Teknologi Informasi, Universitas Stikubank (UNISBANK), Semarang, Indonesia

*Korespondensi : ronihalim0019@mhs.unisbank.ac.id

Submit : 15 Jun 2026 | Diterima : 06 Jul 2026 | Terbit : 27 Agust 2026

ABSTRACT

Stock price prediction models based on deep learning generally produce raw numerical outputs that are difficult for non-technical investors to interpret, and they rarely incorporate external contextual information that drives price movements. This study aimed to build and evaluate an integrated system that combines a Hybrid BiLSTM-GRU prediction model, Retrieval-Augmented Generation (RAG), and a Large Language Model (LLM) to produce interpretable stock price predictions for PT Bank Central Asia Tbk (BBCA). The deep learning component compared four architectures (LSTM, BiLSTM, GRU, and Hybrid BiLSTM-GRU) across univariate and multivariate scenarios and three prediction horizons. The RAG component indexed 121 contextual documents into a vector database, while three LLMs (DeepSeek-LLM 7B, GPT, and Grok) generated analytical narratives. Evaluation used RMSE, MAE, MAPE, and Directional Accuracy for prediction, and RAGAS and BERTScore for narrative quality. The results showed that the Hybrid BiLSTM-GRU univariate model at the t+1 horizon achieved the best performance with an RMSE of 280.17 IDR and MAPE of 2.77%. The RAG component achieved a Context Precision of 0.967, and the integrated system successfully transformed numerical predictions into contextual narratives. The study concluded that integrating deep learning, RAG, and LLM is a feasible approach to improve the interpretability of stock price prediction, although the locally-run LLM still exhibits quality limitations relative to commercial models.

Keywords: BBCA, BiLSTM-GRU, large language model, retrieval-augmented generation, stock prediction

ABSTRAK

Model prediksi harga saham berbasis *deep learning* umumnya menghasilkan keluaran numerik mentah yang sulit diinterpretasikan oleh investor non-teknis, serta jarang mempertimbangkan informasi kontekstual eksternal yang memengaruhi pergerakan harga. Penelitian ini bertujuan membangun dan mengevaluasi sebuah sistem terintegrasi yang memadukan model prediksi Hybrid BiLSTM-GRU, *Retrieval-Augmented Generation* (RAG), dan *Large Language Model* (LLM) untuk menghasilkan prediksi harga saham PT Bank Central Asia Tbk (BBCA) yang interpretable. Komponen deep learning yang terdiri dari empat arsitektur (LSTM, BiLSTM, GRU, dan *Hybrid* BiLSTM-GRU) dibandingkan pada skenario data univariat dan multivariat serta tiga horizon prediksi. Komponen RAG mengindeks 121 dokumen kontekstual ke dalam basis data vektor, sementara tiga LLM (DeepSeek-LLM 7B, GPT, dan Grok) menghasilkan narasi analitis. Evaluasi menggunakan RMSE, MAE, MAPE, dan *Directional Accuracy* untuk prediksi, serta RAGAS dan BERTScore untuk kualitas narasi. Hasil penelitian menunjukkan model *Hybrid* BiLSTM-GRU univariat pada horizon t+1 mencapai performa terbaik dengan RMSE 280,17 IDR dan MAPE 2,77%. Komponen RAG mencapai *Context Precision* 0,967, dan sistem terintegrasi berhasil mentransformasi prediksi numerik menjadi narasi kontekstual. Penelitian menyimpulkan bahwa integrasi *deep learning*, RAG, dan LLM merupakan pendekatan yang layak untuk meningkatkan interpretabilitas prediksi harga saham, meskipun LLM yang dijalankan secara lokal masih memiliki keterbatasan kualitas dibandingkan model komersial.

Kata Kunci: BBCA, BiLSTM-GRU, *Large Language Model*, Prediksi Saham, *Retrieval-Augmented Generation*

PENDAHULUAN

Pasar modal merupakan bagian penting dalam perekonomian suatu negara khususnya dalam mendorong pertumbuhan ekonomi dan investasi jangka panjang. Di Bursa Efek Indonesia (BEI), sektor perbankan menjadi salah satu sektor strategis yang memiliki kontribusi besar terhadap pergerakan Indeks Harga Saham Gabungan (IHSG). Salah satu saham yang menjadi perhatian investor adalah PT Bank Central Asia Tbk (BBCA), sebagai perusahaan dengan nilai kapitalisasi pasar terbesar di Indonesia, pergerakan harga saham BBCA menjadi salah satu indikator kondisi sektor keuangan nasional. Perubahan harga saham dipengaruhi oleh berbagai faktor internal maupun eksternal. Analisis fundamental melalui informasi keuangan perusahaan menjadi salah satu pendekatan penting bagi investor dalam memahami kinerja dan prospek perusahaan sebelum mengambil keputusan investasi (Dayanti et al., 2024). Oleh karena itu, prediksi harga saham yang akurat menjadi kebutuhan strategis dalam mendukung investasi dan manajemen risiko.

Perkembangan *Artificial Intelligence* (AI) mendorong pemanfaatan pendekatan *deep learning* dalam prediksi harga saham. *Deep learning* mampu mempelajari pola kompleks dan hubungan non-linear pada data historis sehingga banyak diterapkan pada data deret waktu (*time series*). Pendekatan yang banyak digunakan adalah *Recurrent Neural Network* (RNN), khususnya *Long Short-Term Memory* (LSTM), *Bidirectional LSTM* (BiLSTM), dan *Gated Recurrent Unit* (GRU). (Saberironaghi et al., 2025) menunjukkan bahwa metode *deep learning* saat ini mendominasi pendekatan pemodelan prediksi harga saham. (Satria, 2023) juga melakukan perbandingan RNN, LSTM, dan GRU pada saham sektor perbankan dan menunjukkan bahwa pendekatan berbasis RNN mampu menghasilkan prediksi dengan performa yang baik dibanding metode statistik konvensional.

Penelitian berkembang dari penerapan model tunggal menuju pendekatan *Hybrid* yang menggabungkan beberapa arsitektur untuk meningkatkan kemampuan ekstraksi pola non-linear. (Dwiandiyanta et al., 2025) membandingkan model RNN, LSTM, dan GRU pada beberapa saham unggulan BEI termasuk BBCA, sedangkan (Rahmadayan & Mustakim, 2024) menunjukkan bahwa GRU mampu menghasilkan nilai *error* prediksi lebih rendah dibandingkan LSTM. (Barua et al., 2024) menemukan bahwa kombinasi GRU dan CNN memberikan tingkat akurasi terbaik pada pasar internasional. Perbedaan hasil tersebut menunjukkan bahwa tidak terdapat satu model yang selalu unggul, sehingga pendekatan *Hybrid* menjadi relevan. (J. Zhang et al., 2023) mengungkapkan bahwa kombinasi CNN-BiLSTM-Attention mampu meningkatkan performa prediksi, sementara (Farhadi et al., 2025) menunjukkan bahwa kombinasi LSTM-GRU mampu menurunkan *error* dibandingkan model dasar. (Karim & Das, 2022) juga menunjukkan bahwa *Hybrid* BiLSTM-GRU memiliki performa lebih baik dibandingkan model LSTM, BiLSTM, maupun GRU secara individual. Namun, sebagian besar penelitian tersebut masih berfokus pada data historis harga saham seperti *Open*, *High*, *Low*, *Close*, dan *Volume* (OHLCV), sehingga belum sepenuhnya mempertimbangkan informasi eksternal yang bersifat dinamis seperti berita keuangan, sentimen pasar, dan kondisi makroekonomi. Untuk mengatasi keterbatasan tersebut, penelitian terbaru mulai mengeksplorasi penggunaan *Retrieval-Augmented Generation* (RAG) dan *Large Language Model* (LLM) dalam analisis finansial. (Iaroshev et al., 2024) mengevaluasi penerapan RAG pada tugas tanya jawab laporan keuangan menggunakan kerangka evaluasi RAGAS. (Yang et al., 2024) memperkenalkan TimeRAG untuk meningkatkan kemampuan prediksi deret waktu berbasis LLM melalui mekanisme *retrieval*. Selain itu, (Luckianto & Gunawan, 2026) mengusulkan arsitektur *multi-agent* berbasis RAG-LLM untuk mendukung keputusan investasi, sedangkan (Fatouros et al., 2025) mengintegrasikan GPT dan RAG dalam analisis harga saham.

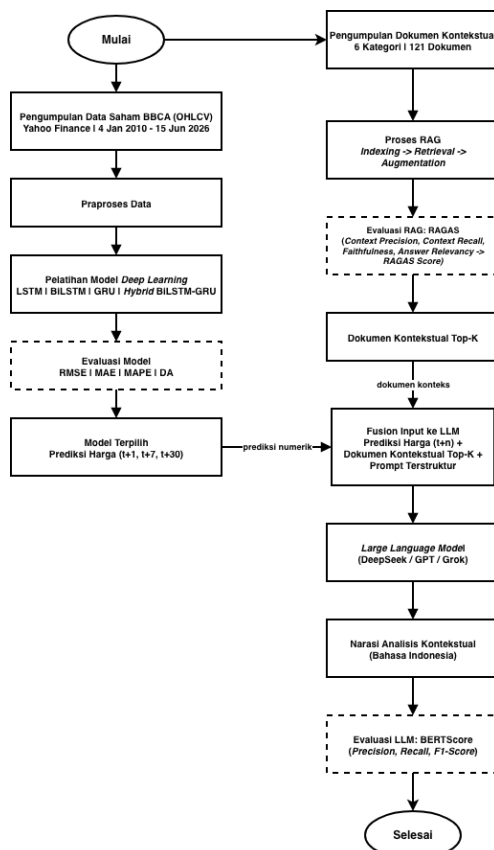
Meskipun RAG dan LLM memiliki potensi besar, penelitian terdahulu masih dilakukan secara terpisah dari model *deep learning* berbasis RNN. Model *deep learning* hanya menghasilkan prediksi numerik tanpa interpretasi kontekstual, sedangkan RAG-LLM lebih berfokus pada pemrosesan informasi tanpa mengintegrasikan kemampuan prediksi *time series*. (Lewis et al., 2020) menjelaskan bahwa RAG mampu mengurangi keterbatasan pengetahuan internal dan risiko halusinasi LLM melalui integrasi sumber eksternal. (Chen et al., 2023) dalam penelitiannya juga menunjukkan bahwa kualitas dokumen hasil *retrieval* berpengaruh terhadap kualitas keluaran sistem RAG. Dengan kemampuan LLM, hasil prediksi dapat dikonversi menjadi narasi analitis yang lebih mudah dipahami investor, dengan evaluasi menggunakan BERTScore (T. Zhang et al., 2020).

Berdasarkan permasalahan tersebut, penelitian ini mengusulkan integrasi *Hybrid* BiLSTM-GRU, *Retrieval-Augmented Generation* (RAG), dan *Large Language Model* (LLM) untuk prediksi

harga saham BBKA. Penelitian ini bertujuan menerapkan dan membandingkan performa beberapa arsitektur *deep learning* menggunakan metrik *Root Mean Square Error* (RMSE), *Mean Absolute Error* (MAE), *Mean Absolute Percentage Error* (MAPE), dan *Directional Accuracy*, RAGAS untuk evaluasi komponen RAG, serta BERTScore digunakan untuk mengevaluasi kualitas narasi hasil generasi LLM.

METODE PENELITIAN

Penelitian ini dilakukan melalui serangkaian tahapan yang terdiri dari tiga komponen inti yang saling terhubung, dimulai dari penerapan metode *deep learning*, penerapan RAG, dan integrasi LLM. Seluruh tahapan metode diimplementasikan menggunakan bahasa Python pada lingkungan Google Colaboratory. Arsitektur sistem secara keseluruhan diilustrasikan pada Gambar 1.



Gambar 1. Rancangan Arsitektur Sistem Terintegrasi

Penerapan metode *deep learning* meliputi tahapan pengolahan data harga saham BBKA mulai dari pengumpulan data, praproses, pelatihan model, hingga menghasilkan prediksi harga. Tahapan kedua merupakan penerapan metode RAG dengan membangun basis pengetahuan dari dokumen kontekstual dan menghasilkan dokumen relevan melalui proses *retrieval*. Tahapan ketiga merupakan tahapan integrasi LLM, dengan menggabungkan keluaran kedua tahapan sebelumnya menjadi sebuah *fusion input* yang menghasilkan narasi analisis. Pada diagram tersebut, evaluasi setiap tahapan metode ditandai dengan kotak bergaris putus-putus untuk membedakan dari proses utama, evaluasi model prediksi menggunakan metrik RMSE, MAE, MAPE, dan DA, evaluasi komponen RAG menggunakan RAGAS, serta evaluasi kualitas narasi hasil generasi LLM menggunakan BERTScore

Praproses Data

Data harga harian saham BBKA dikumpulkan dari sumber resmi Yahoo Finance sepanjang periode 4 Januari 2010 sampai 15 Juni 2026. Berdasarkan hari bursa aktif, diperoleh sejumlah 4.043 baris dengan lima variabel OHLCV. Data akan melalui tahap pemeriksaan validitas dan integritas untuk memastikan kelengkapan data. Selanjutnya dilakukan pembagian data secara kronologis (*time-based split*) menjadi tiga bagian, yaitu data latih periode 2010-2020 (2.732 baris), data uji periode 2021-2023 (732 baris), dan data validasi periode 2024-2026 (579 baris).

Pembagian data dilakukan sebelum proses normalisasi guna mencegah kebocoran data (*data leakage*). Normalisasi kemudian diterapkan untuk memetakan rentang nilai seluruh fitur ke dalam skala 0 hingga 1. Tahap berikutnya adalah pembentukan *sliding window* sepanjang 60 hari untuk mengubah data deret waktu menjadi format sekuensial, dengan target prediksi pada horizon (hari prediksi) $t+1$, $t+7$, dan $t+30$. Data akan diuji pada dua skenario fitur, yaitu skenario univariat yang hanya menggunakan fitur harga penutupan (*close price*) dan skenario multivariat dengan menggunakan seluruh fitur data variabel yaitu OHLCV.

Arsitektur Deep Learning

Penelitian ini menerapkan dan membandingkan empat arsitektur *deep learning*, yakni LSTM, BiLSTM, GRU, dan *Hybrid* BiLSTM-GRU. Arsitektur *Hybrid* memadukan lapisan BiLSTM, yang menangkap konteks temporal dari dua arah, dengan lapisan GRU yang menawarkan efisiensi komputasi. Demi menjaga objektivitas perbandingan, sebagian *hyperparameter* model diseragamkan sebagaimana dirinci pada Tabel 1.

Tabel 1. Konfigurasi *Hyperparameter*

<i>Hyperparameter</i>	<i>LSTM</i>	<i>BiLSTM</i>	<i>GRU</i>	<i>Hybrid</i>
<i>Units</i>	64	128	64	128
<i>Layers</i>	2	2	2	2
<i>Dropout</i>	0,2	0,2	0,2	0,2
<i>Batch size</i>	32	32	32	32
<i>Learning rate</i>	0,001	0,001	0,001	0,001
<i>Timestep</i>	60	60	60	60
<i>Optimizer</i>	Adam	Adam	Adam	Adam
<i>Epoch</i>	100	100	100	100
<i>Early stopping</i>	15	15	15	15

Penerapan RAG

RAG dibangun dari 121 dokumen tekstual sebagai *knowledge base* yang dikumpulkan dari berbagai sumber resmi, jumlah dokumen ditentukan berdasarkan ruang lingkup penelitian yang dianggap relevan, yaitu dokumen terkait BBKA yang dikelompokkan ke dalam enam kategori, yaitu berita keuangan BBKA, laporan keuangan BBKA, makroekonomi, laporan analis saham, sentimen pasar, dan dokumen penelitian terkait topik prediksi saham. Setiap dokumen diubah menjadi potongan (*chunk*) sepanjang 800 karakter, sehingga diperoleh 1.288 *chunk*. Tiap *chunk* direpresentasikan menggunakan model *embedding multilingual* yang digunakan dalam penelitian ini yaitu model paraphrase-multilingual-MiniLM-L12-v2 berdimensi 384 dan disimpan pada basis data vektor ChromaDB. Kualitas *retrieval* dinilai menggunakan metrik *Recall@3* dan *Precision@3*.

Integrasi dan Evaluasi LLM

Integrasi dilakukan dengan menggabungkan tiga komponen menjadi satu *fusion input*, yaitu prediksi numerik dari model *Hybrid* BiLSTM-GRU, tiga dokumen teratas hasil *retrieval*, dan *prompt template* terstruktur. *Fusion input* diproses oleh tiga model LLM, yakni DeepSeek-LLM 7B yang dijalankan secara lokal melalui Ollama sebagai model utama, serta GPT dan Grok melalui API sebagai model perbandingan, dengan nilai temperature 0,3. Hasil narasi dinilai melalui dua pendekatan, yaitu kerangka evaluasi RAGAS yang mengukur empat metrik utama, yakni *context precision*, *context recall*, *faithfulness*, dan *answer relevancy* yang kemudian diagregasi menjadi RAGAS *Score*, serta BERTScore yang mencakup *precision*, *recall*, dan *F1-score* dengan dokumen konteks sebagai acuan. Penilaian dilakukan atas lima kueri yang diuji.

HASIL DAN PEMBAHASAN

Performa Model Deep Learning

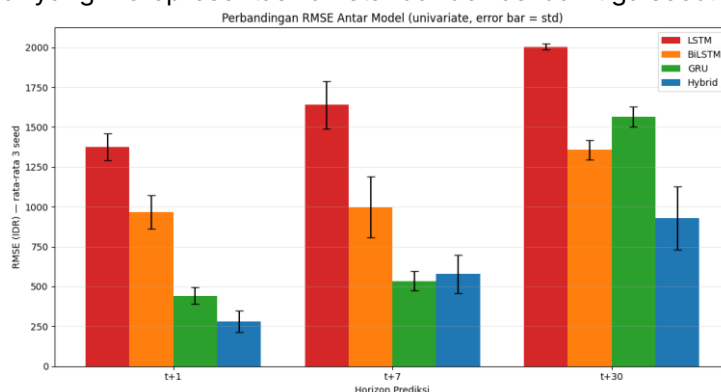
Keempat model dievaluasi pada dua periode pengujian. Tabel 2 memuat hasil evaluasi skenario univariat pada periode validasi data 2024-2026 yang dipandang paling relevan karena paling mendekati kondisi terkini. Model *Hybrid* BiLSTM-GRU menghasilkan kinerja terbaik pada horizon $t+1$ dengan RMSE 280,17 IDR dan MAPE 2,77%, jauh melampaui BiLSTM yang memperoleh RMSE 966,46 dan LSTM dengan RMSE 1375,75. Namun, terdapat pengecualian pada horizon $t+7$, ketika GRU dengan RMSE 534,56 IDR sedikit lebih unggul dibandingkan

Hybrid yang mencatat 579,03 IDR, sehingga keduanya dapat dipandang setara pada horizon menengah.

Tabel 2. Hasil Evaluasi Periode Validasi (Univariat)

<i>Horizon</i>	<i>Model</i>	<i>RMSE</i>	<i>MAE</i>	<i>MAPE (%)</i>	<i>DA (%)</i>
t+1	Hybrid	280,17	234,23	2,77	39,06
t+1	GRU	442,89	365,56	4,28	45,11
t+1	BiLSTM	966,46	870,40	10,14	39,83
t+1	LSTM	1375,75	1245,43	14,52	41,51
t+7	GRU	534,56	429,01	5,04	47,27
t+7	Hybrid	579,03	484,50	5,69	43,23
t+7	BiLSTM	997,99	883,31	10,30	43,36
t+7	LSTM	1638,60	1467,99	17,06	46,22
t+30	Hybrid	929,09	764,30	8,95	44,31
t+30	BiLSTM	1357,58	1161,37	13,43	44,10
t+30	GRU	1563,71	1371,80	15,93	45,06
t+30	LSTM	2003,43	1798,45	20,97	45,40

Untuk memperjelas perbandingan antarmodel secara visual, Gambar 2 menampilkan diagram batang nilai RMSE dari keempat model pada skenario univariat untuk ketiga horizon, dilengkapi *error bar* yang merepresentasikan *standar deviasi* dari tiga *seed*.



Gambar 2. Perbandingan RMSE

Pada Gambar 2 terlihat bahwa di horizon t+1 model *Hybrid* BiLSTM-GRU memiliki RMSE terendah dan diikuti GRU, sementara BiLSTM dan LSTM jauh lebih besar. Di horizon t+7 GRU dan *Hybrid* relatif setara, sedangkan di horizon t+30 *Hybrid* kembali menjadi yang terbaik. *Error bar* yang relatif sempit pada model *Hybrid* mengindikasikan kinerja model stabil. Skenario multivariat memperlihatkan kecenderungan yang serupa namun dengan *error* yang lebih besar, sebagaimana dirinci pada Tabel 3. Model *Hybrid* BiLSTM-GRU tetap menjadi yang terbaik pada skenario ini, tetapi seluruh nilai RMSE-nya lebih tinggi dibandingkan skenario univariat. Keadaan ini memperkuat temuan bahwa penambahan variabel *Open*, *High*, *Low*, dan *Volume* tidak memperbaiki kualitas prediksi.

Tabel 3. Hasil Evaluasi Periode Validasi (*Multivariat*)

<i>Horizon</i>	<i>Model</i>	<i>RMSE</i>	<i>MAE</i>	<i>MAPE (%)</i>	<i>DA (%)</i>
t+1	Hybrid	347,77	289,34	3,37	42,54
t+1	GRU	953,73	851,83	9,92	45,88
t+1	BiLSTM	1444,75	1315,70	15,36	43,11
t+1	LSTM	1623,52	1462,94	17,02	44,66
t+7	Hybrid	659,13	550,33	6,41	43,42
t+7	GRU	906,23	771,35	8,92	46,48
t+7	BiLSTM	1431,67	1283,31	14,93	42,51
t+7	LSTM	1802,94	1634,81	19,07	44,99
t+30	Hybrid	897,13	733,36	8,60	42,47
t+30	BiLSTM	1616,92	1418,84	16,47	42,88
t+30	GRU	1753,52	1550,21	18,01	43,97
t+30	LSTM	2231,25	2018,61	23,59	44,99

Dari kedua tabel tersebut tampak dua pola yang konsisten. Pertama, skenario univariat menghasilkan ketepatan yang lebih tinggi daripada multivariat pada hampir seluruh horizon. Hal ini terjadi karena variabel *Open*, *High*, dan *Low* bergerak sangat berdekatan dengan harga

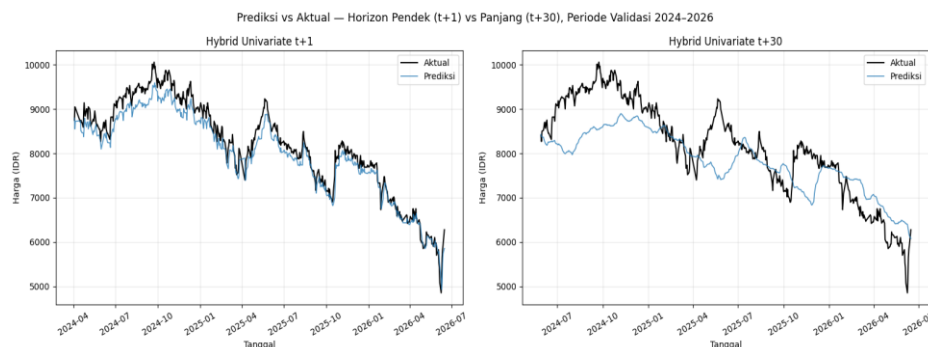
penutupan sehingga nyaris tidak memberikan informasi baru, sedangkan *Volume* memiliki pola yang berlainan dan cenderung berperan sebagai derau. Kedua, ketepatan prediksi menurun seiring bertambah panjangnya horizon, sebab pengaruh pola historis kian melemah pada rentang waktu yang lebih jauh. Nilai *Directional Accuracy* seluruh model berkisar pada 38-48%, yang menandakan keterbatasan dalam memprediksi arah pergerakan harga meskipun besaran harga dapat diramalkan dengan cukup akurat.

Guna menilai generalisasi, model terbaik diuji pada dua periode sebagaimana disajikan pada Tabel 4. Nilai RMSE pada periode data 2024-2026 secara konsisten lebih tinggi daripada periode data 2021-2023, dengan kenaikan berkisar antara 48% hingga 99%. Peningkatan nilai *error* ini bersumber dari jarak waktu terhadap data pelatihan serta sifat data harga saham yang non-stasioner, sehingga perlu dilakukan pembaruan model secara berkala.

Tabel 4. Perbandingan RMSE *Hybrid* BiLSTM-GRU pada Dua Periode

Skenario	Horizon	Uji (2021-2023)	Validasi (2024-2026)	Kenaikan
Univariat	t+1	156,97	280,17	+78%
Univariat	t+7	345,14	579,03	+68%
Univariat	t+30	629,86	929,09	+48%
Multivariat	t+1	174,39	347,77	+99%
Multivariat	t+7	381,84	659,13	+73%
Multivariat	t+30	580,88	897,13	+54%

Gambar 3 memvisualisasikan perbandingan harga prediksi model *Hybrid* BiLSTM-GRU terhadap harga aktual pada periode validasi untuk horizon t+1 dan t+30. Pada horizon t+1, garis prediksi mengikuti harga aktual dengan sangat rapat, sedangkan pada horizon t+30 penyimpangan tampak lebih nyata, khususnya pada awal periode.



Gambar 3. Perbandingan Harga Prediksi dan Aktual pada Periode Validasi

Pola pada Gambar 3 memperkuat temuan kuantitatif sebelumnya. Kerapatan antara prediksi dan aktual pada horizon t+1 sejalan dengan RMSE yang rendah sebesar 280,17 IDR, sementara penyimpangan yang lebih lebar pada horizon t+30 selaras dengan meningkatnya RMSE menjadi 929,09 IDR. Dengan demikian, model terbukti sangat andal untuk peramalan jangka pendek, namun ketepatannya menurun pada horizon yang lebih panjang.

Kualitas *Retrieval*/RAG

Penilaian komponen RAG atas kueri uji menghasilkan *Recall@3* sebesar 1,000, yang menunjukkan bahwa sistem secara konsisten mampu menarik dokumen relevan ke dalam tiga hasil teratas. Capaian ini menegaskan keandalan komponen *retrieval* dalam menyediakan konteks yang relevan dengan dokumen *knowledge base* yang digunakan dalam penelitian ini.

Kualitas Narasi LLM

Kualitas narasi ketiga model LLM dinilai melalui metrik RAGAS (Tabel 5) dan metrik BERTScore (Tabel 6).

Tabel 5. Hasil Evaluasi RAGAS (rata-rata 5 kueri)

Model	Context Precision	Context Recall	Faithfulness	Answer Rel.	RAGAS Score
DeepSeek-LLM 7B	0,967	0,640	0,442	0,785	0,708

Model	Context Precision	Context Recall	Faithfulness	Answer Rel.	RAGAS Score
GPT	0,967	0,640	0,611	0,783	0,750
Grok	0,967	0,640	0,622	0,769	0,750

Berdasarkan evaluasi metrik RAGAS diperoleh nilai *Context Precision* sebesar 0,967 dan *Context Recall* sebesar 0,640 identik antarmodel karena keduanya mengukur hasil *retrieval* berdasarkan konteks yang sama. Perbedaan justru muncul pada *Faithfulness*, ketika GPT dengan nilai 0,611 dan Grok dengan nilai 0,622 lebih tinggi daripada DeepSeek yang hanya 0,442. Secara agregat, GPT dan Grok memperoleh RAGAS *Score* yang setara sebesar 0,750, sementara DeepSeek berada pada 0,708.

Tabel 6. Hasil Evaluasi BERTScore

Model	Precision	Recall	F1-Score
DeepSeek-LLM 7B	0,7746	0,7548	0,7646
GPT	0,7578	0,7204	0,7386
Grok	0,7231	0,6567	0,6883

BERTScore memperlihatkan pola yang berlainan sebagaimana terlihat pada Tabel 6. DeepSeek meraih *F1-score* tertinggi sebesar 0,7646, diikuti GPT dengan 0,7386 dan Grok dengan 0,6883. Perbedaan arah hasil ini berakar pada karakter kedua metrik, yakni BERTScore mengukur kemiripan leksikal-semantik terhadap dokumen sumber, sehingga narasi yang lebih dekat dengan rincian dokumen memperoleh nilai lebih tinggi. DeepSeek cenderung lebih merapatkan teksnya pada sumber, sementara Grok yang menghasilkan narasi paling ringkas memperoleh nilai lebih rendah karena lebih banyak melakukan parafrase.

Contoh Narasi Interpretatif

Guna memperlihatkan kemampuan sistem dalam meningkatkan interpretabilitas, berikut disajikan contoh narasi yang dihasilkan untuk kueri mengenai prospek harga dan kondisi terkini saham BBKA. Prediksi numerik model, yakni penurunan harga dari Rp 6.275 menjadi Rp 6.139 atau sekitar 2,17%, berhasil ditransformasikan menjadi narasi yang menjelaskan konteks di balik angka tersebut.

Narasi GPT:

"Berdasarkan hasil prediksi model Hybrid BiLSTM-GRU, harga penutupan saham BBKA pada 15 Juni 2026 adalah Rp 6,275, dengan prediksi penurunan harga sebesar 2,17% menjadi Rp 6,139 dalam satu hari ke depan. Hal ini sejalan dengan tren penurunan yang telah terjadi sejak awal tahun 2026, di mana saham BBKA tercatat turun sekitar 19% hingga awal April 2026. Meskipun terdapat tekanan jual, terutama dari investor asing, analis mencatat bahwa tekanan tersebut mulai berkurang menjelang laporan keuangan full year 2025. Kinerja keuangan BBKA tetap menunjukkan pertumbuhan meskipun harga sahamnya melemah. Penurunan harga saham ini juga sejalan dengan pergerakan IHSG yang turun 15,79% secara year to date. Disclaimer: Analisis ini bukan merupakan rekomendasi investasi."

Narasi Grok:

"Model prediksi Hybrid BiLSTM-GRU memperkirakan harga saham BBKA akan turun dari Rp6.275 (15 Juni 2026) menjadi Rp6.139 keesokan harinya, atau sekitar 2,17%. Prediksi ini konsisten dengan tren pelemahan saham BBKA sejak awal 2026 yang telah mencapai minus 19% hingga April 2026, disertai tekanan jual investor asing dan penurunan di bawah level Rp7.000. Meskipun kinerja keuangan perseroan tetap tumbuh, sentimen pasar masih negatif. Analisis ini bukan rekomendasi investasi."

Narasi DeepSeek-LLM 7B:

"Berdasarkan hasil prediksi model Hybrid BiLSTM-GRU, saham PT Bank Central Asia Tbk (BBKA) diprediksi akan turun sekitar 2.17% dan mencapai harga penutup Rp6,139 pada satu hari ke depan. Prediksi ini berkaitan dengan beberapa faktor, termasuk tekanan jual yang semakin berkurang terkait kebijakan perseroan yang tetap agresif dalam membagikan keuntungan kepada pemegang saham; perdagangan harga BBKA yang semakin melemah sejak awal tahun

2026, dan pergerakan saham di pasar yang berbeda dengan indikasi penurunan. Dalam konteks informasi tersebut, dokumen 1 menjelajahi laporan keuangan PT Bank Central Asia Tbk (BBCA) ... [narasi berhenti sebelum selesai pada bagian penutup disclaimer]."

Analisis Hasil Integrasi

Contoh tersebut memperlihatkan bahwa hasil keluaran sistem tidak lagi sebatas angka prediksi, melainkan prediksi yang disertai penjelasan dengan konteks faktual seperti tren penurunan, sentimen investor, dan pergerakan IHSG. Inilah esensi peningkatan interpretabilitas yang menjadi tujuan penelitian.

Kedua metode penilaian narasi menyajikan sudut pandang yang saling melengkapi. BERTScore menunjukkan DeepSeek paling dekat secara tekstual dengan sumber, sedangkan RAGAS menunjukkan GPT dan Grok lebih cermat dalam menghasilkan klaim yang benar-benar sesuai konteks. Dengan demikian, kemiripan tekstual tidak serta-merta mencerminkan ketepatan klaim, sehingga penilaian kualitas narasi menuntut lebih dari satu metrik. Selain itu, GPT dan Grok secara konsisten menghasilkan narasi yang utuh, sedangkan DeepSeek-LLM 7B memperlihatkan keterbatasan berupa keluaran yang terpotong atau mengulang sebagian format masukan, sejalan dengan nilai *Faithfulness* terendah.

Apabila dibandingkan dengan penelitian terdahulu, Karim dan Das (2022) memperlihatkan keunggulan *Hybrid* BiLSTM-GRU pada data NIFTY-50, dan penelitian ini mengukuhkan pola serupa pada saham perbankan Indonesia. Berbeda dengan penelitian terdahulu yang berhenti pada prediksi numerik, penelitian ini memperluas *pipeline* melalui RAG dan LLM untuk menjawab kesenjangan interpretabilitas yang diidentifikasi Saberironaghi et al. (2025). Pemaduan ini memungkinkan sistem menutup keterbatasan model numerik yang hanya menghasilkan angka tanpa penjelasan, dengan menghadirkan narasi yang menerangkan bukan sekadar besaran prediksi harga, tetapi juga faktor yang melatarinya berdasarkan dokumen yang relevan.

Pemilihan DeepSeek-LLM 7B sebagai model utama didasarkan pada pertimbangan kemandirian, biaya, dan privasi data, mengingat sifatnya yang dapat dijalankan secara lokal tanpa bergantung pada layanan berbayar maupun server pihak ketiga, pertimbangan privasi ini bermakna penting dalam konteks keuangan. Beberapa keterbatasan penelitian mencakup penurunan ketepatan model secara temporal, kualitas model lokal yang masih di bawah model komersial, serta penggunaan lima kueri uji sebagai studi awal pada evaluasi narasi. Keterbatasan tersebut sekaligus menjadi pijakan bagi pengembangan lanjutan, seperti pembaruan model secara berkala, pemanfaatan model lokal berkapasitas lebih besar, atau penyempurnaan (*fine-tuning*) pada domain keuangan berbahasa Indonesia.

KESIMPULAN

Penelitian ini berhasil menerapkan serta mengevaluasi sebuah sistem terintegrasi yang memadukan model *Hybrid* BiLSTM-GRU, RAG, dan LLM untuk menghasilkan prediksi harga saham BBCA yang *interpretable*. Model *Hybrid* BiLSTM-GRU univariat pada horizon $t+1$ terpilih sebagai model terbaik dengan RMSE 280,17 IDR dan MAPE 2,77%, dengan skenario univariat yang pada umumnya melampaui multivariat dan ketepatan yang menurun seiring bertambahnya horizon. Komponen RAG mencapai *retrieval* yang andal dengan Recall@3 sebesar 1,000 dan Context Precision sebesar 0,967, sedangkan komponen LLM berhasil mengubah prediksi numerik menjadi narasi kontekstual sehingga sasaran peningkatan interpretabilitas terpenuhi. Penilaian narasi mengungkap adanya kontras yang bermakna antara BERTScore dan RAGAS, yang menandakan bahwa kedua metrik menyoroti dimensi kualitas yang berbeda. Model lokal DeepSeek-LLM 7B ditetapkan sebagai model utama atas dasar pertimbangan kemandirian, biaya, dan privasi data, meskipun secara kualitas masih berada di bawah model komersial. Secara keseluruhan, pemaduan *deep learning*, RAG, dan LLM terbukti merupakan pendekatan yang layak untuk menghasilkan prediksi harga saham yang tidak hanya akurat secara numerik pada jangka pendek, tetapi juga dilengkapi penjelasan kontekstual yang mudah dipahami para investor. Penelitian lanjutan dapat diarahkan pada pematuration model secara berkala, pemanfaatan model lokal yang lebih besar atau *fine-tuning*, serta perluasan evaluasi narasi dengan melibatkan pakar keuangan.

REFERENSI

Barua, M., Kumar, T., Raj, K., & Roy, A. M. (2024). Comparative Analysis of Deep Learning Models for Stock Price Prediction in the Indian Market. *FinTech*, 3(4), 551–568. <https://doi.org/10.3390/fintech3040029>

- Chen, J., Lin, H., Han, X., & Sun, L. (2023). *Benchmarking Large Language Models in Retrieval-Augmented Generation*. <http://arxiv.org/abs/2309.01431>
- Dayanti, Primmasatya, & Fernando. (2024). *View of Pengaruh Rasio Keuangan Terhadap Harga Saham Perusahaan Sektor Teknologi Di Bei*. 8, 1.
- Dwiandiyanta, B. Y., Hartanto, R., & Ferdiana, R. (2025). Harnessing Deep Learning and Technical Indicators for Enhanced Stock Predictions of Blue-Chip Stocks on the Indonesia Stock Exchange (IDX). *Engineering, Technology and Applied Science Research*, 15(1), 20348–20357. <https://doi.org/10.48084/etasr.9850>
- Farhadi, A., Zamanifar, A., Alipour, A., Taheri, A., & Asadolahi, M. (2025). A Hybrid LSTM-GRU Model for Stock Price Prediction. *IEEE Access*, 13, 117594–117618. <https://doi.org/10.1109/ACCESS.2025.3586558>
- Fatouros, G., Metaxas, K., Soldatos, J., & Kyriazis, D. (2025). Can Large Language Models beat wall street? Evaluating GPT-4's impact on financial decision-making with MarketSenseAI. *Neural Computing and Applications*, 37(30), 24893–24918. <https://doi.org/10.1007/s00521-024-10613-4>
- Iaroshev, I., Pillai, R., Vaglietti, L., & Hanne, T. (2024). Evaluating Retrieval-Augmented Generation Models for Financial Report Question and Answering. *Applied Sciences (Switzerland)*, 14(20). <https://doi.org/10.3390/app14209318>
- Karim, M. E., & Das, S. (2022). *Stock Price Prediction using Bi-LSTM and GRU based Hybrid Deep Learning Approach*. https://doi.org/10.1007/978-981-19-3148-2_60
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. <http://arxiv.org/abs/2005.11401>
- Marvin Luckianto, & Alexander Agung Santoso Gunawan. (2026). MARAG-Fin: An Intelligent Multi-agent RAG-LLM Architecture Integrating Financial News Sentiment and Time Series Data for Data-driven Trading Decision-making. *International Journal of Intelligent Engineering and Systems*, 19(2), 738–753. <https://doi.org/10.22266/ijies2026.0228.46>
- Rahmadeyan, A., & Mustakim. (2024). Long Short-Term Memory and Gated Recurrent Unit for Stock Price Prediction. *Procedia Computer Science*, 234, 204–212. <https://doi.org/10.1016/j.procs.2024.02.167>
- Saberironaghi, M., Ren, J., & Saberironaghi, A. (2025). Stock Market Prediction Using Machine Learning and Deep Learning Techniques: A Review. In *AppliedMath* (Vol. 5, Number 3). Multidisciplinary Digital Publishing Institute (MDPI). <https://doi.org/10.3390/appliedmath5030076>
- Satria, D. (2023). PREDICTING BANKING STOCK PRICES USING RNN, LSTM, AND GRU APPROACH. *Applied Computer Science*, 19(1), 82–94. <https://doi.org/10.35784/acs-2023-06>
- Yang, S., Wang, D., Zheng, H., & Jin, R. (2024). *TimeRAG: BOOSTING LLM Time Series Forecasting via Retrieval-Augmented Generation*. <http://arxiv.org/abs/2412.16643>
- Zhang, J., Ye, L., & Lai, Y. (2023). Stock Price Prediction Using CNN-BiLSTM-Attention Model. *Mathematics*, 11(9). <https://doi.org/10.3390/math11091985>
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT*. <http://arxiv.org/abs/1904.09675>