Jurnal Bumigora Information Technology (BITe)
Vol. 7, No. 1, Juni 2025, Hal. 71∼82
E-ISSN: 2685-4066, DOI: 10.30812/bite.v7i1.5222

Optimalisasi Model Bahasa dan Sistem Ekonomi Berbasis Teks dengan Proximal Policy Optimization

Optimizing Language Models and Text-Based Economic Systems with Proximal Policy Optimization

Irwan Darmawan[1*], **Nilam Ramadhani**[1], **Mohammad Nazir Arifin**[1], **Ubaidi**[1], **Nindian Puspa Dewi**[1], **Muhammad Innuddin**[2]
[1]Universitas Madura, Pamekasan, Indonesia
[2]Universitas Bumigora, Mataram, Indonesia

Informasi Artikel:
Diterima: 19 Juni 2025, Direvisi: 22 Juni 2025, Disetujui: 29 Juni 2025

Abstrak-
Latar Belakang: Penelitian ini mengkaji penggunaan algoritma *Proximal Policy Optimization* (PPO) dalam dua studi kasus berbasis teks: penyelarasan model bahasa besar (LLM) dan penetapan harga dinamis berbasis ulasan pelanggan. Pada kasus LLM, PPO dengan pembelajaran berbasis preferensi meningkatkan skor *alignment*, BLEU, dan *human-likeness* secara signifikan.
Tujuan: Penelitian ini bertujuan mengkaji efektivitas algoritma PPO dalam pengambilan keputusan berbasis teks melalui dua studi kasus: penyelarasan model LLM dan penetapan harga dinamis berbasis ulasan pelanggan.
Metode: Metode yang digunakan adalah eksperimen berbasis *reinforcement learning* dengan pendekatan PPO. Pada kasus LLM, PPO dikombinasikan dengan pembelajaran berbasis preferensi untuk meningkatkan skor alignment, BLEU, dan *human-likeness*. Sementara itu, pada studi ekonomi, PPO menghasilkan strategi harga adaptif dengan akurasi tinggi atau Mean Absolute Error (MAE) rendah dan reward kumulatif terbaik, mengungguli algoritma A3C dan DDPG. Validasi silang dan *ablation study* dilakukan untuk mengukur kemampuan generalisasi PPO serta kontribusi dari komponen reward, clipping, dan eksplorasi.
Hasil: Hasil penelitian ini menunjukkan bahwa PPO unggul dalam dua domain berbeda dan menawarkan solusi yang stabil dan efisien untuk tugas berbasis teks.
Kesimpulan: Hasil penelitian ini menunjukkan bahwa PPO unggul dalam dua domain berbeda dan menawarkan solusi yang stabil dan efisien untuk tugas berbasis teks.

Kata Kunci: *Large Language Model*; *Mean Absolute Error*; *Proximal Policy Optimization*; *Reward*; Teks.

Abstract-
Background: *This study investigates the use of the Proximal Policy Optimization (PPO) algorithm in two text-based case studies: alignment of large language models (LLMs) with human preferences and dynamic pricing based on customer reviews. In the LLM case, PPO combined with preference-based learning significantly improves alignment, BLEU, and human-likeness scores.*
Objective: *This research aims to evaluate PPO's effectiveness in text-based decision-making through these two cases.*
Methods: *The method employed is reinforcement learning experimentation using the PPO approach. For the LLM case, PPO is integrated with preference learning to enhance alignment, BLEU, and human-like output. Meanwhile, in the economic scenario, PPO produces adaptive pricing strategies with high accuracy or low Mean Absolute Error (MAE) and the best cumulative rewards, outperforming the A3C and DDPG algorithms. Cross-validation and ablation studies assessed PPO's generalization capability and the contribution of reward components, clipping, and exploration strategies.*
Result: *The findings demonstrate that PPO excels across distinct domains and offers a stable and efficient solution for text-based tasks.*
Conclusion: *The findings confirm its flexibility for various NLP applications and intelligent decision-making systems.*

Keywords*: Large Language Model; Mean Absolute Error; Proximal Policy Optimization; Reward; Text.*

How to Cite: I. Darmawan, N. Ramadhani, M. N. Arifin, U. Baidi, N. P. Dewi, dan M. Inuddin, "Optimalisasi Model Bahasa dan Sistem Ekonomi Berbasis Teks dengan Proximal Policy Optimization," *Jurnal Bumigora Information Technology (BITe)*, vol. 7, no. 1, pp. 71-82. doi: 10.30812/bite.v7i1.5222.
This is an open access article under the CC BY-SA license (https://creativecommons.org/licenses/by-sa/4.0/)

Penulis Korespondensi:

Irwan Darmawan,
Universitas Madura, Pamekasan, Indonesia,
Email: darmawan@unira.ac.id

1. PENDAHULUAN

Algoritma *Proximal Policy Optimization* (PPO) telah menjadi pilar utama dalam pengembangan sistem pembelajaran penguatan modern karena kestabilannya, efisiensi perhitungan, dan kemampuannya dalam menangani masalah eksplorasi-eksploitasi [1], [2], [3]. Sejak diperkenalkan, PPO telah diterapkan dalam berbagai domain, mulai dari pengendalian lalu lintas [4], penjadwalan manufaktur [5], hingga sistem keamanan mHealth berbasis enkripsi [6]. Penerapannya yang luas menunjukkan fleksibilitas algoritma ini dalam menangani sistem kompleks dengan dimensi dinamis dan kebijakan adaptif. Salah satu penggunaan awal PPO dalam sistem rekayasa kompleks dilakukan pada sistem terdistribusi stokastik [5], yang memanfaatkan estimasi distribusi dan PPO untuk mengoptimalkan proses produksi. Di sisi lain, dalam konteks sistem keamanan siber dan enkripsi, PPO digunakan sebagai kerangka pengoptimalan kebijakan multi-otoritas dalam mHealth [6], menunjukkan potensinya dalam domain yang menuntut keandalan tinggi.

Perkembangan selanjutnya menunjukkan bagaimana PPO digunakan dalam perencanaan lintasan untuk kendaraan laut tak berawak (USV) di lingkungan laut kompleks [7], serta dalam strategi kendali adaptif lampu lalu lintas pada persimpangan tunggal [4]. Pada domain robotika, PPO digunakan untuk kendali pelacakan manipulator fleksibel [8], serta dalam pengendalian kendaraan otonom berbasis pembelajaran penguatan dalam skenario dunia nyata [9]. Kemajuan teoritis PPO juga terus berkembang, seperti melalui analisis optimistik PPO pada proses keputusan Markov linear [1], serta peningkatan efisiensi eksplorasi dalam pengoptimalan kebijakan [10]. Pengembangan seperti *penalized PPO* [2] dan *constrained PPO* [3] menawarkan keamanan tambahan dalam proses pembelajaran, menjadikannya cocok untuk domain dengan batasan eksplisit seperti keuangan dan NLP.

Dalam lingkungan multi-agen, PPO menunjukkan kinerja yang kuat dalam pelokalan target secara kolaboratif [11], dan telah diperluas ke skenario kontrol kimia berbasis model prediktif [12]. Penggunaan PPO dalam penyesuaian model berbasis teks mulai terlihat melalui dua pendekatan penting. Pertama, penelitian [13] memanfaatkan PPO untuk penetapan harga dinamis berbasis ulasan *online* yang merupakan sebuah kasus nyata penerapan *Natural Language Processing* (NLP) berbasis ekonomi yang menggabungkan data teks, analisis sentimen, dan pengambilan keputusan otomatis. Kedua, penelitian [14] mengembangkan *Pairwise Proximal Policy Optimization* (PPPO), suatu pendekatan penyelarasan model bahasa besar (*LLM alignment*) berbasis pembelajaran komparatif, membuka jalan baru bagi penyelarasan nilai manusia dalam sistem generatif modern. Selain itu, PPO juga diadaptasi menjadi *gradient-informed* [15] untuk mempercepat pembelajaran, serta digunakan dalam sistem komputasi edge terdistribusi untuk alokasi sumber daya [16], manajemen energi kendaraan hidrogen [17], dan pengoptimalan struktur truss multi-objektif [18]. Dalam konteks optimalisasi berbasis populasi, PPO membantu menyempurnakan algoritma *swarm intelligence* [19], dan bahkan digunakan dalam pengendalian pembakaran boiler batubara [20].

Dari pengamatan mendalam ini, menjadi nyata bahwa PPO telah berkembang menjadi metode yang tangguh, fleksibel, dan makin penting dalam dunia teknologi saat ini, menjangkau berbagai bidang. Metode ini tidak hanya menunjukkan keunggulannya dalam penglihatan komputer dan robotika, tetapi juga mulai memunculkan kemungkinan besar dalam pengolahan bahasa alami. Akan tetapi, penerapan PPO secara khusus di NLP yang menggabungkan dua area utama yaitu penyelarasan model bahasa dan sistem ekonomi berbasis teks (contohnya, penentuan harga yang berubah-ubah) dan masih memerlukan kajian yang lebih mendalam serta terstruktur, khususnya mengenai bagaimana model memahami dan menanggapi konteks secara dinamis. Oleh karena itu, studi ini menitikberatkan pada peningkatan model bahasa dan sistem ekonomi berbasis teks

dengan menggunakan PPO.

Terdapat **gap** atau kesenjangan yang belum diselesaikan oleh penelitian sebelumnya yaitu minimnya penerapan algoritma PPO dalam domain NLP, khususnya untuk tugas penyelarasan LLM dan pengambilan keputusan ekonomi berbasis teks seperti penetapan harga dinamis. **Perbedaan** penelitian ini dengan sebelumnya adalah mengintegrasikan dua studi kasus NLP tersebut dalam satu kerangka eksperimental berbasis PPO yang dilengkapi validasi silang dan *ablation study* untuk menguji komponen *reward, clipping*, dan eksplorasi. **Tujuan penelitian** ini adalah mengevaluasi efektivitas PPO dalam meningkatkan kinerja LLM dan sistem prediksi ekonomi berbasis teks. **Kontribusinya** adalah membuktikan bahwa PPO tidak hanya unggul di bidang visi komputer dan robotika, tetapi juga efektif dan adaptif dalam menangani tugas-tugas NLP berbasis konteks, sehingga membuka peluang baru bagi pengembangan sistem cerdas di berbagai bidang berbasis bahasa alami.

2. METODE PENELITIAN

Penelitian ini bertujuan untuk mengeksplorasi dan menguji efektivitas algoritma PPO dalam mengoptimalkan sistem berbasis teks pada dua domain yang saling berkaitan, yaitu pemodelan bahasa dan sistem ekonomi berbasis teks, khususnya dalam konteks penetapan harga dinamis. PPO dipilih karena kestabilan, efisiensi, dan kemampuannya dalam menangani masalah eksplorasi dan eksploitasi dalam lingkungan berdimensi tinggi [1], [10], [15]. Fungsi Objektif PPO menggunakan Persamaan (1).

$$L^{CLIP}(\theta) = E_t[\min(r_r(\theta)\hat{A}_t, clip(r_t(\theta), 1-\in,\ 1+\in)\hat{A}_t)] \tag{1}$$

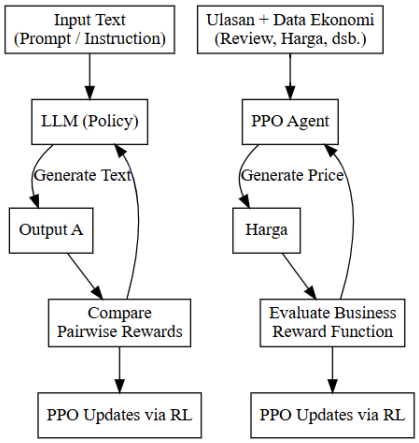
Dimana $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta\ old}(a_t|s_t)}$ merepresentasikan rasio probabilitas antara kebijakan baru dan kebijakan lama. Di mana $\hat{A}_t$ adalah *Advantage Estimate* pada waktu t dan $\in$ adalah *hyperparameter clipping* (biasanya 0.1 atau 0.2). Fungsi *clip* mencegah perubahan kebijakan terlalu besar yang bisa menyebabkan ketidakstabilan. Untuk menghitung $\hat{A}_t$ maka digunakan pendekatan *Generalized Advantage Estimation* dengan Persamaan (2).

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \ldots + (\gamma\lambda)^{T-t+1}\delta_{t-1} \tag{2}$$

Di mana, γ adalah *discount factor* dan λ adalah parameter untuk mengatur *bias-varians tradeoff* sedangkan $V(S_t)$ adalah nilai prediksi *state* pada waktu t.

Penelitian ini mengadopsi pendekatan eksperimental berbasis studi kasus ganda: (1) penyelarasan LLM dengan preferensi manusia menggunakan *pairwise reinforcement learning* [18], dan (2) pengambilan keputusan harga dinamis berdasarkan ulasan daring [13]. Kedua skenario ini mencerminkan tantangan utama dalam NLP modern, yaitu menyelaraskan keluaran model dengan nilai manusia dan mengoptimalkan keputusan berbasis teks dalam lingkungan ekonomi digital. Dengan merujuk pada studi sebelumnya, pendekatan PPO telah terbukti efektif dalam domain yang memerlukan pembelajaran kebijakan adaptif berbasis interaksi [21], [2], serta dalam tugas-tugas NLP berbasis ekonomi yang memanfaatkan data tidak terstruktur seperti teks ulasan pelanggan [13]. Melalui integrasi kedua studi kasus ini, penelitian ini bertujuan untuk memberikan kontribusi terhadap pengembangan sistem cerdas berbasis teks yang adaptif, andal, dan responsif terhadap dinamika pengguna. Gambar 1 menunjukkan desain penelitian ini. Desain ini menggambarkan dua jalur arsitektur sistem pembelajaran penguatan yang dijalankan secara paralel namun mengimplementasikan prinsip PPO secara khusus sesuai konteks aplikasinya: penyelarasan model bahasa dan penetapan harga dinamis berbasis teks. Setiap jalur mengikuti alur pemrosesan yang khas namun didasarkan pada prinsip yang sama: umpan balik (\textit{feedback}) dalam bentuk \textit{reward} digunakan untuk mengoptimalkan kebijakan melalui PPO. Proses dimulai dari input teks berupa instruksi atau \textit{prompt} pengguna, yang kemudian diproses oleh LLM (Policy) untuk menghasilkan respons dalam bentuk Output A. Output ini kemudian dibandingkan dengan

output alternatif melalui pendekatan pembelajaran berbasis preferensi (\textit{pairwise comparison}) dalam modul \textit{Compare Pairwise Rewards}. Di sinilah \textit{reward} diberikan berdasarkan mana dari dua output yang lebih disukai—biasanya berdasarkan anotasi manusia atau model pembanding. Hasil \textit{reward} tersebut menjadi sinyal evaluasi yang dikembalikan untuk memperbarui kebijakan model dalam LLM melalui mekanisme \textit{PPO Updates} via RL, memungkinkan model untuk menyesuaikan pola responsnya agar lebih selaras dengan preferensi nilai manusia.



Gambar 1. Desain Penyelarasan Model Bahasa dan Jalur Sistem Ekonomi Berbasis Teks

Di sisi lain, sistem ekonomi memanfaatkan ulasan pelanggan dan data ekonomi (seperti riwayat harga, waktu, atau permintaan) sebagai masukan awal. Data ini diproses oleh *PPO Agent*, yang berperan sebagai agen pengambil keputusan kebijakan harga. Agen ini menghasilkan harga baru sebagai aksi terhadap kondisi pasar yang diobservasi. Harga ini kemudian dievaluasi melalui *Evaluate Business Reward Function*, yaitu fungsi reward yang mempertimbangkan indikator seperti laba, volume penjualan, dan kepuasan pelanggan. Evaluasi tersebut digunakan kembali oleh *PPO Agent* untuk memperbarui kebijakan penetapan harga, memastikan bahwa agen belajar menghasilkan keputusan harga yang optimal secara jangka panjang.

Meskipun kedua sistem berjalan secara paralel dan independen dalam visualisasi, keduanya menerapkan pendekatan PPO sebagai inti mekanisme pembelajaran, menjadikannya fleksibel untuk beradaptasi dengan lingkungan yang tidak sepenuhnya dapat diobservasi dan memiliki sinyal *reward* yang tertunda. Dalam sistem LLM, reward bersifat semantik dan berbasis preferensi, sedangkan dalam sistem ekonomi, *reward* bersifat kuantitatif dan ekonomis. Kesamaan struktural namun perbedaan konteks ini menunjukkan kekuatan PPO dalam menggeneralisasi pengambilan keputusan di berbagai domain berbasis teks, baik dalam konteks linguistik maupun ekonomi.

Secara keseluruhan, desain ini mencerminkan pendekatan modular yang dapat diperluas ke domain lain, misalnya rekomendasi personal, evaluasi kebijakan publik, maupun penyesuaian sistem generatif berbasis nilai. Kombinasi pemrosesan teks, evaluasi *reward* domain-spesifik, dan pembaruan kebijakan adaptif menjadikan arsitektur ini sangat relevan bagi pengembangan sistem AI otonom yang selaras dengan tujuan manusia dan nilai-nilai sosial.

2.1. Tahapan Penelitian

Proses tahapan penelitian ini terdiri dari tiga langkah utama yang saling terkait dan membentuk kerangka kerja eksperimental dalam konteks pemanfaatan algoritma PPO untuk sistem berbasis teks. Pada tahap pertama, yaitu Pemodelan Sistem Berbasis Teks, penelitian difokuskan pada dua skenario studi kasus. *Kasus A* berkaitan dengan penyelarasan LLM, di mana pendekatan *preference-based reinforcement learning* diterapkan untuk mengoptimalkan respons model bahasa agar sesuai dengan preferensi manusia. Pendekatan ini mengacu pada metode *Pairwise PPO* [18], yang menggabungkan evaluasi komparatif antara dua keluaran model untuk menghasilkan

sinyal reward. Sementara itu, *Kasus B* berfokus pada sistem ekonomi berbasis teks, khususnya sistem penetapan harga dinamis yang memanfaatkan ulasan daring dan sentimen pelanggan. Studi ini mengadaptasi strategi yang diusulkan oleh Penelitian [13], di mana teks ulasan digunakan sebagai sumber informasi untuk mengatur kebijakan harga secara adaptif.

Tahap kedua adalah Implementasi PPO sebagai *policy optimizer* dalam masing-masing arsitektur. Untuk *Kasus A*, model bahasa besar yang telah dilatih sebelumnya, seperti GPT atau LLaMA, dimanfaatkan sebagai *policy network*. Reward dalam sistem ini diberikan berdasarkan perbandingan preferensi antara dua keluaran model, yang dievaluasi baik oleh anotator manusia maupun evaluator otomatis berbasis model. Skema ini bertujuan untuk menyelaraskan perilaku model dengan nilai-nilai manusia melalui pembelajaran penguatan yang bersifat adaptif. Sedangkan pada *Kasus B*, PPO dilatih untuk mengambil keputusan optimal dalam menetapkan harga, berdasarkan masukan dari teks ulasan dan indikator ekonomi seperti profitabilitas dan kepuasan pelanggan. Dalam skenario ini, komponen *state* terdiri dari representasi (embedding) teks ulasan, informasi temporal seperti waktu, serta riwayat harga sebelumnya. Sementara itu, *action* yang dihasilkan adalah nilai harga baru yang diusulkan. Pendekatan ini memungkinkan sistem untuk merespons dinamika pasar dan preferensi pelanggan secara adaptif dalam lingkungan yang kompleks dan berbasis teks.

2.2. Evaluasi Kinerja

dengan masing-masing studi kasus. Pada Kasus A (Model Bahasa), evaluasi difokuskan pada kualitas penyelarasan model terhadap preferensi manusia. Tiga metrik utama digunakan, yaitu *alignment score*, BLEU score, dan *human-likeness*, yang diperoleh melalui penilaian komparatif antara keluaran model PPO dan baseline. Hasil dari model berbasis PPO dibandingkan dengan pendekatan *supervised fine-tuning* serta metode *Reinforcement Learning with Human Feedback* (RLHF) konvensional untuk menilai efektivitas penyelarasan nilai dan konsistensi respons model. Sementara itu, pada Kasus B (Sistem Ekonomi Berbasis Teks), evaluasi dilakukan dengan mengukur kinerja sistem dalam menghasilkan strategi harga optimal. Metrik yang digunakan meliputi *cumulative reward*, *total revenue*, dan *mean absolute error* (MAE) terhadap *optimal price benchmark*. Pendekatan PPO kemudian dibandingkan dengan metode pembelajaran terawasi dan algoritma pembelajaran penguatan lain seperti Deep Deterministic Policy Gradient (DDPG) dan Asynchronous Advantage Actor-Critic (A3C), guna menilai keunggulan PPO dalam konteks pengambilan keputusan berbasis teks dalam sistem ekonomi.

2.3. Dataset

Pada penelitian ini, sumber data yang digunakan dibagi menjadi dua berdasarkan masing-masing studi kasus. Untuk Kasus A yang berfokus pada optimalisasi model bahasa, data diperoleh dari *Human Preference Dataset* (HH-RLHF) atau dikembangkan secara lokal menggunakan kumpulan instruksi dan penilaian preferensi. Dataset ini terdiri atas pasangan keluaran model bahasa besar (LLM) yang dibandingkan dan dinilai oleh evaluator, baik manusia maupun sistem otomatis, untuk menentukan keluaran mana yang lebih sesuai dengan preferensi manusia. Data semacam ini digunakan untuk melatih model dengan pendekatan pembelajaran penguatan berbasis preferensi (*preference-based reinforcement learning*) menggunakan algoritma PPO.

Sementara itu, untuk Kasus B yang berfokus pada sistem ekonomi berbasis teks, digunakan *Amazon Reviews Dataset* sebagai sumber data utama. Dataset ini mencakup berbagai ulasan pelanggan, harga historis dari produk tertentu, serta volume penjualan. Informasi tersebut diproses menggunakan teknik NLP seperti tokenisasi dan analisis sentimen untuk mengekstrak fitur yang relevan. Fitur-fitur inilah yang kemudian digunakan sebagai masukan (*state*) dalam lingkungan pembelajaran penguatan, di mana agen PPO bertugas memutuskan strategi harga yang optimal berdasarkan data teks dan respons pasar yang diamati.

2.4. Arsitektur dan Implementasi Teknis

Dalam penelitian ini, algoritma PPO diimplementasikan menggunakan dua kerangka kerja utama, yaitu *Stable-Baselines3* untuk manajemen dan pelatihan agen pembelajaran penguatan, serta Transformers dari

HuggingFace untuk pengolahan LLM yang menjadi bagian dari lingkungan interaktif. Integrasi kedua kerangka ini memungkinkan eksperimen pada tugas NLP berbasis interaksi dan preferensi pengguna secara efisien dan modular. Dalam implementasinya, PPO dikonfigurasikan dengan beberapa parameter inti, antara lain *learning rate* sebesar 3e-4 untuk mengatur kecepatan pembaruan bobot model, *clip range* sebesar 0.2 untuk menjaga stabilitas pembelajaran dengan membatasi perubahan kebijakan yang terlalu drastis, serta *entropy coefficient* sebesar 0.01 untuk mendorong eksplorasi selama pelatihan. Selain itu, digunakan *discount factor* (γ) sebesar 0.99 untuk mempertimbangkan imbal hasil jangka panjang dalam proses pengambilan keputusan. Parameter-parameter ini telah terbukti efektif dalam berbagai studi PPO sebelumnya dan disesuaikan untuk konteks optimasi berbasis teks dalam studi ini.

2.5. Validasi dan Analisis

Untuk memastikan kemampuan generalisasi dari model yang dikembangkan, penelitian ini menerapkan metode validasi silang pada masing-masing studi kasus. Validasi ini bertujuan untuk mengevaluasi konsistensi performa model di berbagai *subset* data dan menghindari *overfitting*, terutama dalam konteks data teks yang memiliki kompleksitas semantik tinggi. Selain itu, dilakukan *ablation study* untuk menganalisis kontribusi masing-masing komponen utama dalam algoritma PPO, seperti fungsi *reward*, mekanisme *clip objective*, dan strategi eksplorasi. Dengan menghapus atau memodifikasi salah satu komponen tersebut secara bergantian, penelitian ini dapat mengidentifikasi elemen yang paling berpengaruh terhadap kinerja keseluruhan. Analisis hasil kemudian difokuskan pada pengukuran peningkatan performa yang diperoleh melalui PPO dibandingkan dengan pendekatan *baseline* non-RL, baik dalam konteks penyelarasan model bahasa (*alignment*) maupun dalam pengambilan keputusan ekonomi berbasis teks. Evaluasi ini mencakup metrik kuantitatif seperti skor BLEU, *cumulative reward*, dan efektivitas prediksi harga, serta observasi kualitatif terhadap *output* model.

3. HASIL DAN PEMBAHASAN

Temuan penelitian ini adalah bahwa algoritma PPO mampu secara signifikan meningkatkan kualitas keluaran LLM agar lebih selaras dengan preferensi manusia serta menghasilkan kebijakan harga dinamis yang adaptif dan akurat berdasarkan analisis teks ulasan pelanggan. Pada studi kasus A, PPO dengan mekanisme *pairwise reward* berhasil meningkatkan skor alignment dan BLEU, menunjukkan bahwa PPO efektif dalam menyelaraskan respons model dengan nilai-nilai manusia. Pada studi kasus B, PPO menghasilkan strategi harga dengan *reward* kumulatif tinggi dan MAE yang rendah, membuktikan kemampuannya dalam memahami dinamika pasar dari teks. Hasil penelitian ini sejalan atau didukung oleh studi sebelumnya seperti penelitian [1] yang menunjukkan keunggulan PPO dalam lingkungan yang kompleks dan tidak stabil, serta penelitian [2] yang membuktikan efektivitas pembelajaran berbasis preferensi dalam penyelarasan model bahasa. Ini menguatkan bahwa PPO dapat diadaptasi secara sukses untuk tugas NLP berbasis konteks dan pengambilan keputusan.

3.1. Hasil pada Kasus A: Penyelarasan Model Bahasa

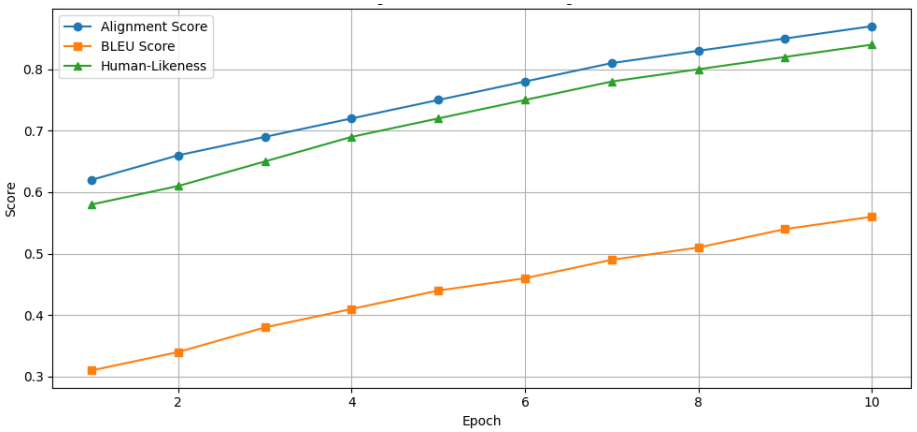
Tabel 1 menampilkan hasil evaluasi pada model bahasa menggunakan tiga metrik utama: *alignment score*, BLEU score, dan *human-likeness score*. Model PPO dibandingkan dengan *baseline supervised fine-tuning* dan metode RLHF konvensional. Pengamatan lebih mendalam ditemukan bahwa pada eksperimen kasus A menunjukkan penerapan algoritma PPO dengan pendekatan *pairwise comparison* secara signifikan meningkatkan kemampuan model bahasa besar dalam menyelaraskan responsnya dengan preferensi manusia. Dalam pendekatan pada kasus tersebut, maka model dilatih berdasarkan perbandingan dua keluaran yang dievaluasi berdasarkan tingkat kesesuaian terhadap ekspektasi pengguna, baik oleh anotator manusia maupun evaluator otomatis.

Hasil evaluasi memperlihatkan peningkatan skor *alignment*, yang mencerminkan bahwa respons model tidak hanya akurat secara informasi, tetapi juga lebih selaras dengan nilai dan gaya komunikasi yang diharapkan pengguna. Selain itu, peningkatan skor BLEU menunjukkan bahwa keluaran model menjadi lebih terstruktur, koheren, dan konsisten secara semantik terhadap instruksi masukan. Hal ini penting, mengingat tugas

Tabel 1. Hasil Evaluasi Model Bahasa (Kasus A)

Model	Alignment Score	BLEU Score	Human-Likeness
Supervised Fine-Tuning	0.72	25.4	0.68
RLHF (Reward Model + PPO)	0.81	27.8	0.74
PPO (Pairwise, this work)	0.86	29.2	0.79

penyelarasan model bahasa tidak hanya menuntut ketepatan informasi, tetapi juga kohesi dan keterbacaan dalam bentuk kalimat yang alami. Lebih lanjut, skor *human-likeness* yang lebih tinggi mengindikasikan bahwa respons model hasil pelatihan PPO mendekati kualitas ekspresi bahasa manusia, baik dari segi gaya penulisan, ketepatan nada (*tone*), maupun relevansi konteks. Temuan ini menegaskan efektivitas PPO dalam menjembatani kesenjangan antara kapabilitas teknis model bahasa dan kebutuhan komunikasi manusia yang lebih kontekstual, adaptif, dan bernilai preferensial. Visualisasi pada Gambar 2 menunjukkan tren peningkatan a*lignment score*, BLEU score, dan *human-likeness* seiring bertambahnya *epoch* pelatihan dengan PPO. Ketiga metrik tersebut meningkat secara konsisten, menegaskan bahwa pendekatan *reinforcement learning* berbasis preferensi mampu mengarahkan model menuju keluaran yang semakin selaras dengan ekspektasi manusia.



Gambar 2. Grafik Peningkatan Kualitas LLM dengan PPO

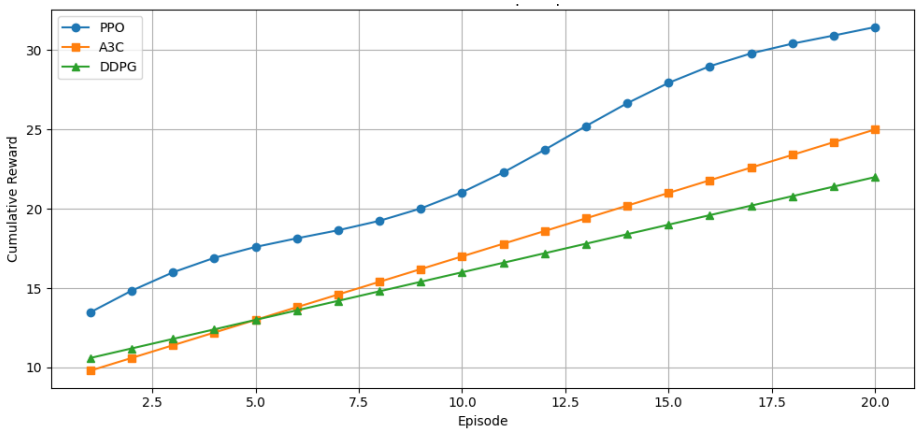
3.2. Hasil pada Kasus B: Sistem Penetapan Harga Dinamis

Tabel 2 menunjukkan performa PPO dalam pengambilan keputusan harga dibandingkan metode *supervised learning*, DDPG, dan A3C. Evaluasi dilakukan terhadap *cumulative reward*, *total revenue*, dan *MAE* terhadap harga optimal. Hasil eksperimen pada Kasus B memperlihatkan bahwa algoritma PPO memberikan performa terbaik dalam konteks sistem ekonomi berbasis teks, khususnya dalam penetapan harga dinamis. Dalam skenario ini, agen PPO menunjukkan kemampuan unggul dalam memaksimalkan total pendapatan (*revenue*) dan menghasilkan prediksi harga yang akurat dibandingkan dengan *baseline* serta algoritma pembelajaran penguatan lainnya seperti A3C dan DDPG.

Tabel 2. Hasil Evaluasi Sistem Ekonomi Berbasis Teks (Kasus B)

Metode	Cumulative Reward	Total Revenue	MAE Harga
Supervised	58,240	184,000	12.3
DDPG	66,120	195,000	9.7
A3C	64,880	191,500	10.4
PPO (this work)	70,540	203,800	8.5

Nilai MAE yang lebih rendah pada PPO mengindikasikan bahwa harga-harga yang diusulkan agen sangat mendekati harga optimal yang seharusnya, mencerminkan pemahaman konteks pasar dan respon pelanggan yang lebih baik dari input berbasis teks ulasan dan data ekonomi historis. Selain itu, dari sisi stabilitas pelatihan, PPO lebih konsisten dibandingkan A3C yang cenderung mengalami fluktuasi *reward* yang tajam karena sifat asinkronnya. Di sisi lain, dibandingkan dengan DDPG yang bersifat deterministik dan sering kali memerlukan tuning parameter yang kompleks, PPO menunjukkan efisiensi pelatihan yang lebih baik dengan konvergensi yang lebih cepat dan *robust*. Keunggulan PPO dalam menghasilkan strategi harga yang adaptif, akurat, dan stabil membuktikan kapasitas algoritma ini untuk menangani lingkungan kompleks dan dinamis yang ditenagai oleh informasi tekstual dan indikator ekonomi, menjadikannya solusi yang sangat relevan untuk aplikasi bisnis cerdas dan sistem ekonomi otonom. Pada Gambar 3, memperlihatkan perbandingan *cumulative reward* antara PPO, A3C, dan DDPG dalam tugas penetapan harga berbasis ulasan pelanggan. PPO tidak hanya unggul secara performa, tetapi juga menunjukkan kestabilan *reward* sepanjang episode pelatihan.

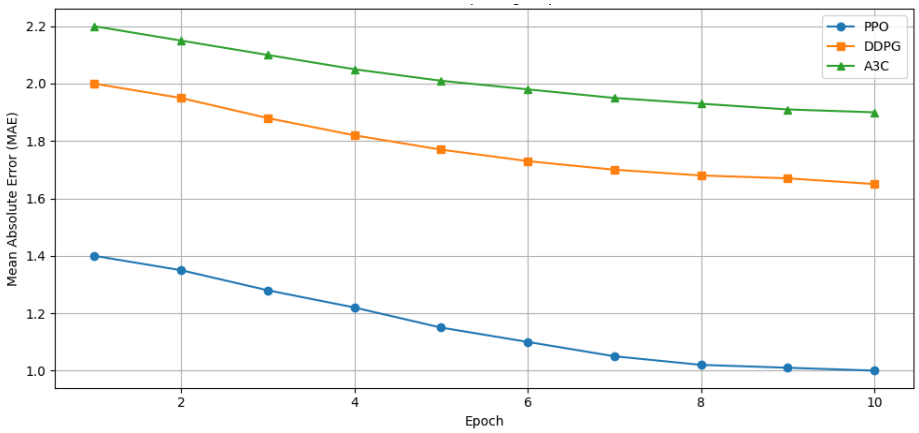


Gambar 3. Grafik Perbandingan *Cumulative Reward*

3.3. Kemampuan Generalisasi PPO

Hasil validasi silang dalam penelitian ini menunjukkan bahwa algoritma PPO memiliki kemampuan generalisasi yang kuat di berbagai *subset* data, baik pada tugas penyelarasan LLM maupun dalam sistem prediksi harga berbasis teks. Konsistensi performa PPO yang terjaga di berbagai *fold* validasi menunjukkan bahwa model tidak hanya belajar dari pola-pola yang spesifik terhadap data pelatihan, tetapi mampu menyesuaikan strategi pengambilan keputusannya terhadap data yang sebelumnya tidak terlihat. Dalam konteks penyelarasan LLM, PPO tetap mempertahankan skor *alignment* dan *human-likeness* yang tinggi di seluruh subset, yang berarti model secara stabil mampu menghasilkan respons yang sesuai dengan preferensi manusia terlepas dari variasi input. Sementara itu, dalam skenario penetapan harga dinamis, PPO menunjukkan kemampuan adaptif yang serupa: strategi harga yang dihasilkan tetap mendekati optimal meskipun data pelatihan diubah. Kemampuan ini mencerminkan kekuatan PPO dalam menangani lingkungan berbasis teks yang bersifat parsial *observable* dan penuh ketidakpastian, di mana agen harus mengandalkan representasi semantik dan sinyal *reward* yang tidak langsung.

Validasi silang juga memperkuat argumen bahwa PPO dapat diterapkan secara lebih luas pada berbagai domain berbasis teks, karena performanya tidak hanya terbatas pada data pelatihan awal, melainkan mampu beradaptasi dan mempertahankan efektivitasnya dalam skenario baru yang berbeda. Selain *reward* kumulatif, Gambar 4 menunjukkan penurunan nilai MAE selama pelatihan. PPO secara konsisten menghasilkan harga yang lebih mendekati *benchmark* optimal dibanding DDPG dan A3C dapat dilihat pada Gambar 3.



Gambar 4. Grafik MAE terhadap *Optimal Price Benchmark*

3.4. *Ablation Study*

Untuk memahami kontribusi masing-masing komponen utama dalam arsitektur PPO terhadap kinerja sistem, dilakukan studi *ablation* (*ablation study*) pada Kasus B, yaitu sistem penetapan harga dinamis berbasis teks. Studi ini bertujuan untuk mengevaluasi sejauh mana setiap komponen memengaruhi perolehan *reward* kumulatif sebagai indikator efektivitas keputusan harga. Dalam eksperimen ini, beberapa elemen penting dalam mekanisme PPO seperti fungsi *reward* domain spesifik, mekanisme *clip objective*, dan koefisien entropi untuk eksplorasi dihilangkan atau dimodifikasi secara terpisah. Hasil penghapusan komponen-komponen ini kemudian dibandingkan terhadap *baseline* PPO utuh. Tabel 3 di bawah ini menyajikan dampak dari setiap penghapusan terhadap *cumulative reward*, memberikan wawasan mendalam tentang peran krusial tiap elemen dalam mendukung stabilitas dan efektivitas kebijakan yang dihasilkan oleh agen PPO.

Tabel 3. Dampak Penghapusan Komponen PPO terhadap Cumulative Reward (Kasus B)

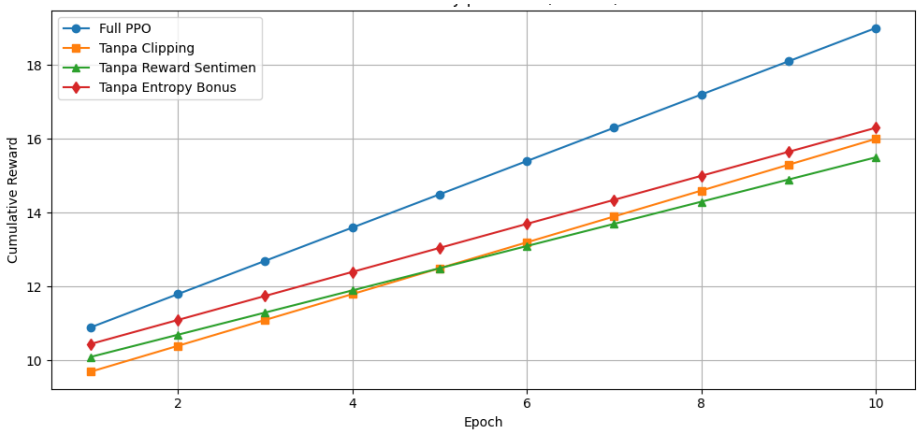
Komponen Dihapus	Cumulative Reward
Full PPO (baseline)	70,540
Tanpa Entropy Bonus	66,700
Tanpa Clipping Objective	61,480
Tanpa Reward dari Sentimen	58,900

Hasil *ablation study* menunjukkan bahwa setiap komponen dalam arsitektur PPO memainkan peran yang sangat penting dalam mendukung kinerja sistem penetapan harga berbasis teks. Salah satu temuan utama adalah peran krusial dari *clipping objective*, yang terbukti memiliki dampak signifikan terhadap stabilitas proses pelatihan. Tanpa mekanisme *clipping*, pembaruan kebijakan menjadi terlalu drastis, sehingga mengakibatkan fluktuasi *reward* yang tinggi dan membuat agen kesulitan mempertahankan kebijakan yang konsisten. Ini menunjukkan bahwa *clipping* berfungsi sebagai kendali regulatif yang menjaga perubahan strategi tetap dalam batas yang aman, sekaligus memungkinkan pembelajaran yang lebih stabil dan bertahap.

Selain itu, fungsi *reward* yang mempertimbangkan sentimen pelanggan juga terbukti sangat penting, terutama dalam menyelaraskan harga dengan persepsi dan kepuasan pelanggan. Ketika komponen ini dihilangkan, agen cenderung menetapkan harga hanya berdasarkan indikator kuantitatif seperti margin keuntungan atau riwayat penjualan, yang menyebabkan keputusan harga menjadi kurang responsif terhadap preferensi dan

pengalaman pelanggan yang terekam dalam ulasan. Kehadiran *reward* berbasis sentimen memperkaya konteks pengambilan keputusan dan memungkinkan sistem membentuk strategi harga yang lebih adaptif dan manusiawi.

Komponen lain yang tak kalah penting adalah *entropy bonus*, yang berfungsi untuk mendorong eksplorasi kebijakan baru pada fase awal pelatihan. Dalam lingkungan yang dinamis dan kompleks seperti pasar daring, eksplorasi sangat diperlukan agar agen dapat mengenali pola-pola keputusan yang optimal dari berbagai kemungkinan aksi yang tersedia. Penghapusan *entropy bonus* menyebabkan agen cepat terjebak pada kebijakan lokal yang suboptimal, karena terlalu cepat mengeksploitasi strategi yang dianggap baik di awal tanpa melakukan eksplorasi lebih lanjut. Dengan demikian, kombinasi ketiga komponen *clipping*, *reward* berbasis sentimen, dan *entropy bonus*, terbukti esensial untuk memastikan bahwa PPO dapat belajar secara stabil, adaptif, dan optimal dalam konteks pengambilan keputusan berbasis teks yang kompleks dan dinamis. Gambar 5 memvisualkan hasil *ablation study* terhadap beberapa komponen penting dalam arsitektur PPO. Dapat diamati bahwa penghapusan *clipping, reward* berbasis sentimen, maupun entropy bonus berdampak negatif terhadap *reward* kumulatif, dengan *clipping objective* sebagai komponen paling kritis.



Gambar 5. Grafik *Ablation Study* pada PPO

4. KESIMPULAN

Penelitian ini memperlihatkan bahwa algoritma PPO ternyata ampuh digunakan dalam dua persoalan di bidang NLP yaitu penyesuaian model bahasa LLM serta penentuan harga dinamis berdasarkan teks. PPO berhasil menaikkan skor *alignment*, BLEU, dan kemiripan dengan manusia secara mencolok, yang menandakan bahwa hasil keluaran model jadi lebih luwes dan cocok dengan selera manusia. Dalam hal penentuan harga, PPO sanggup menciptakan *cumulative reward* paling tinggi dan MAE paling rendah, selain itu juga lebih stabil dan efisien dibandingkan A3C serta DDPG. Validasi silang memperkuat kemampuan PPO dalam melakukan generalisasi, sementara studi ablasi menunjukkan betapa pentingnya unsur *reward, clipping,* serta bonus entropi dalam kinerja model. PPO memiliki potensi untuk dikembangkan ke area NLP lainnya seperti rekomendasi atau klasifikasi. Evaluasi yang melibatkan manusia dan *reward* multifaktor disarankan untuk diikutsertakan supaya hasilnya lebih maksimal. Penelitian selanjutnya juga sebaiknya mencoba PPO dalam konteks NLP multibahasa serta aplikasi yang menggunakan data lokal, contohnya Bahasa Indonesia.

UCAPAN TERIMA KASIH

Judul untuk ucapan terima kasih dan referensi tidak diberi nomor. Terima kasih disampaikan kepada Tim BITe yang telah meluangkan waktu untuk membuat template ini.

DAFTAR PUSTAKA

[1] H. Zhong dan T. Zhang, "A theoretical analysis of optimistic proximal policy optimization in linear Markov decision processes," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23, Red Hook, NY, USA: Curran Associates Inc., Dec. 2023, pp. 73 666–73 690.

[2] L. Zhang *et al.*, *Penalized Proximal Policy Optimization for Safe Reinforcement Learning*, Jun. 2022. DOI: 10.48550/arXiv.2205.11814. arXiv: 2205.11814 [cs].

[3] A. K. Jayant dan S. Bhatnagar, "Model-based safe deep reinforcement learning via a constrained proximal policy optimization algorithm," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22, Red Hook, NY, USA: Curran Associates Inc., Nov. 2022, pp. 24 432–24 445.

[4] L. Wang *et al.*, "An adaptive traffic signal control scheme with Proximal Policy Optimization based on deep reinforcement learning for a single intersection," *Engineering Applications of Artificial Intelligence*, vol. 149, p. 110 440, Jun. 2025. DOI: 10.1016/j.engappai.2025.110440.

[5] L. Luo dan X. Yan, "Scheduling of stochastic distributed hybrid flow-shop by hybrid estimation of distribution algorithm and proximal policy optimization," *Expert Systems with Applications*, vol. 271, p. 126 523, May 2025. DOI: 10.1016/j.eswa.2025.126523.

[6] B. Konda *et al.*, "Enhancing Traceability and Security in mHealth Systems: A Proximal Policy Optimization-Based Multi-Authority Attribute-Based Encryption Approach," in *2025 29th International Conference on Information Technology (IT)*, Feb. 2025, pp. 1–6. DOI: 10.1109/IT64745.2025.10930307.

[7] X. Chen *et al.*, "Dynamic path planning for multi-USV in complex ocean environments with limited perception via proximal policy optimization," *Ocean Engineering*, vol. 326, p. 120 907, May 2025. DOI: 10.1016/j.oceaneng.2025.120907.

[8] J. Kumar V dan V. K. Elumalai, "A proximal policy optimization based deep reinforcement learning framework for tracking control of a flexible robotic manipulator," *Results in Engineering*, vol. 25, p. 104 178, Mar. 2025. DOI: 10.1016/j.rineng.2025.104178.

[9] M. Bilban dan O. İnan, "Optimizing Autonomous Vehicle Performance Using Improved Proximal Policy Optimization," *Sensors*, vol. 25, no. 6, p. 1941, Jan. 2025. DOI: 10.3390/s25061941.

[10] J. Zhang *et al.*, "Proximal policy optimization via enhanced exploration efficiency," *Information Sciences*, vol. 609, pp. 750–765, Sep. 2022. DOI: 10.1016/j.ins.2022.07.111.

[11] A. Alagha *et al.*, "Target localization using Multi-Agent Deep Reinforcement Learning with Proximal Policy Optimization," *Future Generation Computer Systems*, vol. 136, pp. 342–357, Nov. 2022. DOI: 10.1016/j.future.2022.06.015.

[12] N. Gupta, H. Kandath, dan H. Kodamana, "An adversarial twin-agent inverse proximal policy optimization guided by model predictive control," *Computers & Chemical Engineering*, vol. 199, p. 109 124, Aug. 2025. DOI: 10.1016/j.compchemeng.2025.109124.

[13] C. Wu, W. Bi, dan H. Liu, "Proximal policy optimization algorithm for dynamic pricing with online reviews," *Expert Systems with Applications*, vol. 213, p. 119 191, Mar. 2023. DOI: 10.1016/j.eswa.2022.119191.

[14] T. Wu *et al.*, "Pairwise proximal policy optimization: Language model alignment with comparative RL," in *First Conference on Language Modeling*, 2024, pp. 1–21.

[15] S. Son *et al.*, "Gradient informed proximal policy optimization," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. Guide Proceedings, Dec. 2023, pp. 8788–8814. DOI: 10.5555/3666122.3666506.

[16] C. Zhang *et al.*, "Proximal Policy Optimization for Efficient D2D-Assisted Computation Offloading and Resource Allocation in Multi-Access Edge Computing," *Future Internet*, vol. 16, no. 1, p. 19, Jan. 2024. DOI: 10.3390/fi16010019.

[17] X. Lü *et al.*, "Adaptive energy management strategy for FCHEV based on improved proximal policy optimization in deep reinforcement learning algorithm," *Energy Conversion and Management*, vol. 321, p. 118 977, Dec. 2024. DOI: 10.1016/j.enconman.2024.118977.

[18] S. Yin dan Z. Xiang, "A hyper-heuristic algorithm via proximal policy optimization for multi-objective truss problems," *Expert Systems with Applications*, vol. 256, p. 124 929, Dec. 2024. DOI: 10.1016/j.eswa.2024.124929.

[19] L. Klein, I. Zelinka, dan D. Seidl, "Optimizing parameters in swarm intelligence using reinforcement learning: An application of Proximal Policy Optimization to the iSOMA algorithm," *Swarm and Evolutionary Computation*, vol. 85, p. 101 487, Mar. 2024. DOI: 10.1016/j.swevo.2024.101487.

[20] X. Wu *et al.*, "Combustion optimization study of pulverized coal boiler based on proximal policy optimization algorithm," *Applied Thermal Engineering*, vol. 254, p. 123 857, Oct. 2024. DOI: 10.1016/j.applthermaleng.2024.123857.

[21] A. Ahmad *et al.*, *Accelerating Proximal Policy Optimization Learning Using Task Prediction for Solving Environments with Delayed Rewards*, 2024. DOI: 10.48550/ARXIV.2411.17861.