# LLM pada OpenAI untuk Mengevaluasi Kinerja Pegawai di PT Javan Cipta Solusi

**Novi Setiani**
Informatika, Fakultas Teknologi Industri, Universitas Islam Indonesia
Email: novi.setiani@uii.ac.id

***Abstract.*** *Evaluating the performance of the best employees in a company is a common problem faced by the human resources division. Assessing subjectively and comprehensively is a challenge that must be overcome in building sustainable business growth. For organizations that are already able to manage employee assignment data in a structured and systematic form, this can be done by processing quantitative data obtained from the task management system. In this study, a large language model-based approach is proposed to identify the best employees based on assignment data for one year. The main contribution of this study is a comparative framework for evaluating six OpenAI LLM versions using structured task management data and a three-stage prompting design validated by domain experts. A comparison was made of six versions of LLM on OpenAI to measure employee performance based on a predetermined prompting design. As a result, 50% of respondents considered that the criteria proposed by GPT-4o-mini were more appropriate to their needs, and 60% of respondents considered that the employee ranking results produced by GPT-4 were more relevant to the reality. This study acknowledges that LLM-based evaluation carries inherent ethical implications, including potential bias and fairness concerns, which are discussed as part of the study limitations.*
***Keywords:*** *employee performance; LLM; evaluation; comparation.*

***Abstrak.*** *Evaluasi kinerja pegawai terbaik di sebuah perusahaan merupakan permasalahan umum yang dihadapi oleh divisi sumber daya manusia. Menilai secara subjektif dan komprehensif merupakan tantangan yang harus diselesaikan dalam membangun pertumbuhan bisnis yang berkelanjutan. Bagi organisasi yang sudah mampu mengelola data penugasan pegawai secara terstruktur dan sistematis, hal tersebut dapat dilakukan dengan mengolah data kuantitatif yang diperoleh dari sistem manajemen tugas (task management system). Pada penelitian ini, diusulkan pendekatan berbasis large language model untuk mengidentifikasi pegawai terbaik berdasarkan data penugasan selama satu tahun. Kontribusi utama penelitian ini adalah kerangka perbandingan enam versi LLM pada OpenAI menggunakan data manajemen tugas terstruktur dengan desain prompting tiga tahap yang divalidasi oleh pakar domain. Dilakukan perbandingan terhadap enam versi LLM pada OpenAI untuk mengukur kinerja pegawai berdasarkan rancangan prompting yang sudah ditentukan. Hasilnya, sebanyak 50% responden menilai bahwa kriteria yang diusulkan oleh GPT4o-mini lebih sesuai dengan kebutuhan, serta 60% responden (4 orang) menilai bahwa hasil pemeringkatan pegawai yang dihasilkan oleh GPT-4 lebih relevan dengan kenyataan di lapangan**.** Penelitian ini juga mengakui adanya implikasi etis dalam penggunaan LLM untuk evaluasi kinerja, termasuk potensi bias dan isu keadilan, yang dibahas sebagai bagian dari keterbatasan penelitian.*
***Kata Kunci:*** *employee performance; LLM; evaluation; comparation.*

## 1. Pendahuluan

Manajemen kinerja pegawai merupakan salah satu aspek krusial dalam manajemen sumber daya manusia (*human resource management*). Kinerja individu dan tim sangat berpengaruh terhadap pencapaian tujuan organisasi, baik dalam sektor publik maupun swasta.

Namun, organisasi sering menghadapi tantangan dalam mengelola dan mengevaluasi kinerja pegawai secara objektif, efektif, dan efisien. Tantangan tersebut mencakup subjektivitas dalam penilaian kinerja, kurangnya standar yang jelas dalam sistem evaluasi, serta keterbatasan dalam mengidentifikasi faktor-faktor yang mempengaruhi produktivitas pegawai. Oleh karena itu, diperlukan pendekatan yang lebih sistematis dan berbasis data untuk meningkatkan akurasi dan keadilan dalam analisis kinerja pegawai. Penelitian ini hadir untuk mengisi kesenjangan tersebut dengan mengusulkan pemanfaatan Large Language Model (LLM) sebagai alat bantu analisis kinerja berbasis data tugas kuantitatif yang terstruktur, sebuah pendekatan yang belum banyak dieksplorasi dalam literatur manajemen SDM berbasis AI.

Meskipun berbagai teknik seperti analisis multikriteria (AHP, DEMATEL) dan penilaian 360 derajat telah menunjukkan efektivitas dalam meningkatkan pengelolaan SDM, pendekatan berbasis AI Assistant yang didukung oleh *Large Language Model (LLM)*, menawarkan potensi baru dalam analisis kinerja pegawai. LLM seperti GPT-4 (OpenAI), Claude, dan Deepseek dapat membantu organisasi dalam mengolah data kinerja pegawai secara lebih efisien melalui analisis sentimen, pemrosesan umpan balik, serta identifikasi tren produktivitas. Penelitian (Brown, 2020) menunjukkan bahwa model berbasis Transformer, seperti GPT-3, mampu memahami dan mengolah teks dalam berbagai konteks bisnis dengan akurasi tinggi. Sementara itu, penelitian (Devlin, 2019) tentang BERT menunjukkan keunggulan model ini dalam memahami hubungan semantik dalam teks, yang dapat berguna dalam menganalisis laporan kerja dan evaluasi karyawan.

Namun, penggunaan AI Assistant dalam analisis kinerja pegawai juga menghadapi sejumlah tantangan. Tantangan tersebut mencakup keterbatasan dalam memahami konteks spesifik organisasi, potensi bias dalam data pelatihan, serta ketiadaan *ground truth* yang terstandarisasi untuk mengvalidasi hasil evaluasi secara objektif. Di samping itu, aspek etis seperti keadilan peringkat, transparansi kriteria, dan privasi data pegawai menjadi isu kritis yang perlu diperhatikan. Oleh karena itu, penelitian ini bertujuan untuk menjawab pertanyaan: (1) Bagaimana performa berbagai versi LLM OpenAI dalam mengusulkan kriteria dan meranking kinerja pegawai berdasarkan data tugas terstruktur? (2) Model LLM mana yang paling relevan menurut penilaian pakar domain (pimpinan divisi)? Penelitian ini bersifat eksploratif kualitatif awal, dan hasilnya dimaksudkan sebagai wawasan awal yang perlu divalidasi lebih lanjut pada skala yang lebih besar.

## 2. Tinjauan Pustaka

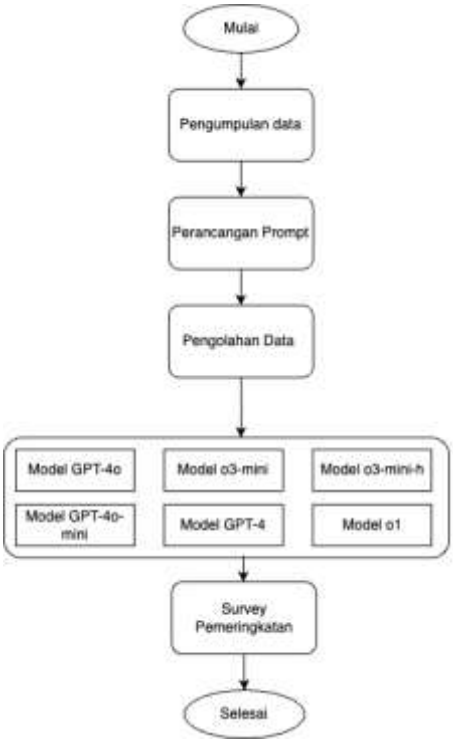
Penelitian sebelumnya telah menunjukkan bahwa penggunaan metode berbasis analisis data dapat membantu mengatasi tantangan dalam HRM (Shi, 2023). Misalnya, (Aksakal, 2014) menggunakan pendekatan *fuzzy Analytic Hierarchy Process* (AHP) untuk mengevaluasi faktor-faktor insentif dalam sumber daya manusia di bidang teknologi tinggi, salah satunya faktor motivasi. Sejalan dengan itu, [3] menggunakan metode *AHP dan Decision Making Trial and Evaluation Laboratory (DEMATEL)* untuk menganalisis manajemen penghargaan berdasarkan empat kriteria utama, yaitu lingkungan kerja, pembelajaran dan pengembangan, tunjangan, serta gaji. Temuan mereka menunjukkan bahwa keseimbangan antara faktor-faktor tersebut sangat berpengaruh terhadap tingkat kepuasan dan kinerja pegawai. Penelitian kualitatif dari (Kharub, 2025) mengidentifikasi faktor yang mempengaruhi kinerja pegawai yaitu motivasi, kepemimpinan dan lingkungan kerja. Secara lebih mendalam, penelitian dilakukan pada beberapa perusahaan berskala kecil, menengah dan besar. Secara kuantitatif, telah dilakukan analisis faktor yang memvalidasi skala kinerja pegawai.

Selain itu, pemilihan metode yang tepat dalam *Performance Appraisal (PA)* juga menjadi tantangan tersendiri bagi organisasi modern. Menurut penelitian (Ijadi, 2018), proses seleksi metode PA yang sesuai dalam lingkungan bisnis yang dinamis dan *agile* merupakan permasalahan kompleks karena melibatkan berbagai faktor dan skala pembiayaan. Studi tersebut menemukan bahwa metode *360-degree feedback* adalah salah satu pendekatan terbaik dalam mengevaluasi kinerja pegawai secara komprehensif. Dengan mempertimbangkan berbagai perspektif, metode ini mampu mengurangi bias dalam penilaian dan memberikan hasil yang lebih objektif.

## 3. Metode Penelitian

Penelitian ini mengambil subjek penelitian di PT Javan Cipta Solusi, sebuah perusahaan yang bergerak di bidang pengembangan perangkat lunak dengan jumlah pegawai sebanyak 90 orang. Untuk mengelola pekerjaan yang ditugaskan kepada setiap pegawai, digunakan piranti lunak manajemen tugas Active Collab. Metode yang digunakan dalam penelitian ini mengikuti tahapan pada Gambar 1, yaitu terdiri dari tahapan: i) pengumpulan data; ii) pembuatan prompt awal; iii) pengolahan data dengan AI Assistant; iv) analisis dan perbandingan hasil; dan v) penentuan ranking kinerja pegawai. Penelitian ini bersifat eksploratif kualitatif awal (pilot study) dengan keterbatasan jumlah responden ahli (n=6) yang merupakan seluruh pimpinan divisi aktif di perusahaan. Ukuran sampel ini dipilih karena penelitian berfokus pada penilaian pakar domain internal, bukan pada generalisasi statistik. Hasil penelitian ini perlu dipandang sebagai wawasan awal yang memerlukan validasi lebih lanjut dengan sampel yang lebih besar dan metrik objektif.

Pengumpulan data bertujuan untuk mengumpulkan data tugas dari Active Collab sejak 1 Januari sampai dengan 31 Desember 2024, sebanyak 22.091 data pengerjaan tugas. Data tersebut melibatkan 90 pegawai yang tersebar di beberapa divisi, dengan rata-rata 245 tugas per pegawai per tahun. Dari 22.091 tugas, sebanyak 18.734 tugas (84,8%) memiliki data completed_on yang terisi, sedangkan sisanya berupa tugas yang belum selesai atau dibatalkan. Statistik deskriptif menunjukkan bahwa rata-rata keterlambatan penyelesaian tugas adalah 2,3 hari dari due_on, dengan standar deviasi 5,1 hari. Selanjutnya, dilakukan perancangan prompt versi awal berdasarkan tujuan evaluasi kinerja pegawai yaitu mencari tiga pegawai terbaik di PT Javan Cipta Solusi. Prompt yang sudah dirancang akan dijalankan pada setiap model OpenAI. Hasil luaran dari setiap model akan dianalisis untuk mengidentifikasi efisiensi, akurasi dan relevansi dengan kebutuhan organisasi. Untuk menentukan peringkat kinerja pegawai, dilakukan survey yang melibatkan pemimpin setiap divisi (n=6) mengenai hasil luaran dari setiap model. Perlu dicatat bahwa penelitian ini tidak memiliki ground truth (kebenaran dasar) yang terstandarisasi untuk validasi objektif hasil pemeringkatan, sehingga validasi dilakukan secara kualitatif melalui penilaian pakar.




**Gambar 1. Metode Penelitian**

Ada 3 tabel berkorelasi yang digunakan yaitu Tasks, Users, dan Projects.

 ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

1. *Tasks* adalah tabel yang digunakan untuk menyimpan tugas yang harus dikerjakan oleh pegawai, dengan detail kolom dijelaskan pada Tabel 1.

**Tabel 1. Kolom pada Tabel *Tasks***

| *No* | *Nama Kolom* | *Keterangan* |
|---|---|---|
| 1 | project_id | Reference column ke table projects |
| 2 | name | Judul tugas yang harus dikerjakan |
| 3 | assignee_id | Reference column ke table users |
| 4 | due_on | Tanggal batas waktu pengerjaan task |
| 5 | completed_on | Tanggal kapan task selesai dikerjakan |
| 6 | completed_by_name | Nama pegawai yang menyelesaikan task tersebut |

2. *Users* adalah tabel yang digunakan untuk menyimpan daftar pegawai yang bekerja di perusahaan, dengan detail kolom dijelaskan pada Tabel 2.

**Tabel 2. Kolom pada Tabel *Users***

| *No* | *Nama Kolom* | *Keterangan* |
|---|---|---|
| 1 | id | Referenced column dari table tasks kolom assignee_id |
| 2 | first_name | Nama depan pegawai |
| 3 | last_name | Nama belakang pegawai |

3. *Projects* adalah tabel yang digunakan untuk menyimpan daftar pegawai yang bekerja di perusahaan, dengan detail kolom dijelaskan pada Tabel 3.

**Tabel 3. Kolom pada Tabel *Projects***

| *No* | *Nama Kolom* | *Keterangan* |
|---|---|---|
| 1 | id | Referenced column dari table tasks kolom project_id |
| 2 | name | Judul proyek yang harus dikerjakan |

Data dari database disimpan sebagai file CSV dengan menampilkan kolom-kolom yang penting untuk mencari pegawai berkinerja tinggi. Query SQL yang digunakan untuk mengambil data dari ketiga tabel yang berkorelasi terdapat pada Kode 1.

**Kode 1. Query pengambilan data**

```
select p.name, concat(u.first_name," ", u.last_name) as assignee, t.name as task,
t.due_on, t.completed_on, t.completed_by_name
from tasks t
join projects p on p.id = t.project_id
left join users u on u.id = t.assignee_id
where year(t.created_on) = 2024
```

Kolom yang diekspor dari database terdiri dari enam kolom, dijelaskan pada Tabel 4.

**Tabel 4. Kolom pada hasil ekspor data**

| *No* | *Kolom* | *Keterangan* |
|---|---|---|
| 1 | project | Nama atau kode proyek tempat tugas ini berada. |
| 2 | assignee | Orang yang ditugaskan untuk mengerjakan tugas ini. |
| 3 | task | Nama atau deskripsi tugas yang harus dikerjakan. |
| 4 | due_on | Tanggal batas akhir tugas harus diselesaikan. |
| 5 | completed_on | Tanggal tugas ini selesai dikerjakan. |
| 6 | completed_by_name | Nama orang yang menyelesaikan tugas ini. |

## 4. Hasil dan Pembahasan

### 4.1 Perancangan Prompt

Prompt yang digunakan untuk melakukan evaluasi ditentukan berdasarkan tiga tujuan yang ingin dicapai yaitu: AI Assistant mampu membaca data, AI Assistant mampu memahami apa yang diperlukan, dan AI Assistant mampu menghasilkan tiga orang pegawai dengan kinerja terbaik. Detail hasil rancangan prompt ditunjukkan pada Tabel 5.

**Tabel 5. Perancangan Prompt**

| *No* | *Goal* | *Prompt* |
|---|---|---|
| 1 | AI Assistant membaca data | Saya memiliki data pengerjaan tugas dalam format CSV dengan delimiter ';'. Data ini berasal dari Active Collab dengan kriteria dibuat pada tahun 2024.<br>Ada 6 kolom data tersedia:<br>1. project, Nama atau kode proyek tempat tugas ini berada.<br>2. assignee, Orang yang ditugaskan untuk mengerjakan tugas ini.<br>3. task, Nama atau deskripsi tugas yang harus dikerjakan.<br>4. due_on, Tanggal batas akhir tugas harus diselesaikan.<br>5. completed_on, Tanggal tugas ini selesai dikerjakan.<br>6. completed_by_name, Nama orang yang menyelesaikan tugas ini. |
| 2 | AI assistant memahami apa yang diinginkan | Saya ingin mengetahui siapa yang memiliki kinerja terbaik berdasarkan data tersebut. Buatlah kriteria serta alasan mengapa kriteria tersebut dapat dievaluasi berdasarkan data yang ada dan perlu digunakan untuk menilai kinerja terbaik. |
| 3 | AI assistant memberikan data top performer | Sebutkan tiga orang dengan kinerja terbaik beserta alasannya |

### 4.2 Perbandingan Luaran

Pada Tabel 6 ditunjukkan rangkuman perbandingan respon dari setiap versi LLM pada open AI terhadap prompt yang diajukan.

**Tabel 6. Perbandingan Respon**

| *No* | *LLM* | *Respon 1* | *Respon 2* | *Respon 3* |
|---|---|---|---|---|
| 1 | GPT-4o | Analisis data | Usulan kriteria | 3 nama pegawai |
| 2 | o3-mini | Ucapan Terima kasih | Usulan kriteria | Nama dummy |
| 3 | o3-mini-high | Mengulang informasi dan ucapan terima kasih | Usulan kriteria | Nama dummy |
| 4 | o1 | Pertanyaan informasi apa yang ingin digali | Usulan kriteria | Langkah mengolah data |
| 5 | GPT-4o mini | Mengulang informasi | Usulan Kriteria | 3 nama pegawai |
| 6 | GPT-4 | Ucapan terima kasih dan pertanyaan informasi apa yang akan digali | Usulan Kriteria | 3 nama pegawai |

Beberapa contoh hasil kriteria yang diusulkan oleh LLM adalah: (1) Jumlah tugas yang diselesaikan; (2) Persentase penyelesaian tepat waktu; (3) Rata-rata selisih hari antara tanggal penyelesaian dan tenggat *(Due Date*); (4) Konsistensi dalam penyelesaian tugas ; dan (5)Tugas yang memiliki dampak positif terhadap proyek.

### 4.3 Evaluasi Pemeringkatan Pegawai

Luaran dari setiap model dievaluasi kepada pimpinan divisi sebanyak enam orang responden dengan menggunakan survey *online*. Pertanyaan yang diajukan dalam survey adalah sebagai berikut: (1) Kriteria mana yang paling sesuai untuk mengevaluasi kinerja pegawai?; dan (2) Hasil pemeringkatan pegawai manakah yang paling sesuai?

Responden mengisi survey tanpa mengetahui jenis versi LLM yang digunakan untuk membangkitkan kriteria dan meranking pegawai. Berdasarkan hasil survey tersebut, sebanyak 50% responden (3 orang) menilai bahwa kriteria yang diusulkan oleh GPT4o-mini lebih sesuai dengan kebutuhan. Sedangkan untuk pertanyaan kedua, 60% responden (4 orang) menilai bahwa hasil pemeringkatan pegawai yang dihasilkan oleh GPT-4 lebih relevan dengan kenyataan di lapangan. Secara kualitatif, GPT-4 cenderung mempertimbangkan faktor kontekstual seperti konsistensi lintas proyek dan kompleksitas relatif tugas, sementara GPT-4o-mini lebih unggul dalam menyusun kriteria yang ringkas dan mudah dipahami oleh manajer. Adapun model o3-mini, o3-mini-high, dan o1 menghasilkan nama dummy atau langkah pengolahan data, yang mengindikasikan perbedaan perilaku model dalam merespons prompt berbasis dokumen CSV. Perlu dicatat bahwa hasil evaluasi ini bersifat kualitatif dan subjektif karena didasarkan pada persepsi enam responden, bukan metrik performa objektif. Tidak adanya *ground truth* yang terstandarisasi merupakan keterbatasan utama yang membuat hasil penelitian ini tidak dapat digeneralisasikan secara langsung. Terdapat pula potensi bias konfirmasi di mana responden mungkin menyukai hasil yang sesuai persepsi awal mereka. Implikasi etis penggunaan LLM dalam evaluasi kinerja pegawai meliputi: transparansi kriteria penilaian, pemahaman potensi bias algoritmik, serta *human oversight* sebelum hasil evaluasi AI digunakan dalam keputusan SDM resmi.

## KESIMPULAN

Perbandingan terhadap penggunaan enam versi LLM pada OpenAI untuk mengevaluasi kinerja pegawai memberikan hasil yang berbeda. Tiga dari enam versi LLM mampu melakukan pemeringkatan terhadap kinerja pegawai berdasarkan kriteria yang dirancang. Secara efisiensi respon, model o3-mini, o3-mini-high, dan GPT-4o-mini banyak mengulang informasi dan hanya menyampaikan terimakasih. Berdasarkan survey terhadap enam pimpinan divisi, hasil pemeringkatan oleh GPT-4 dinilai paling relevan dengan kenyataan di lapangan. Namun, kriteria yang diusulkan oleh GPT4o-mini dinilai paling sesuai dengan kebutuhan mereka. Penelitian ini memiliki sejumlah keterbatasan yang perlu diakui secara eksplisit: (1) ukuran sampel responden yang kecil (n=6) membatasi generalisasi hasil; (2) tidak adanya ground truth yang terstandarisasi menyebabkan validasi hanya bersifat kualitatif; (3) data yang digunakan hanya berasal dari satu perusahaan sehingga hasilnya tidak mewakili konteks organisasi yang berbeda; dan (4) evaluasi hanya dilakukan pada model-model OpenAI tanpa membandingkan dengan LLM lain seperti Claude atau Gemini. Oleh karena itu, temuan penelitian ini harus dipandang sebagai wawasan kualitatif awal (pilot study), bukan sebagai hasil yang definitif. Penelitian lanjutan disarankan untuk menggunakan sampel responden yang lebih besar, melibatkan ground truth formal (seperti data KPI resmi), serta mempertimbangkan aspek etis dan keadilan algoritmik dalam merancang sistem evaluasi kinerja berbasis AI.

## UCAPAN TERIMA KASIH

Penulis mengucapkan terimakasih kepada pimpinan divisi di PT Javan Cipta Solusi yang telah berpartisipasi dalam penelitian ini dengan mengisi survey evaluasi hasil analisis kinerja pegawai.

## Referensi

Aksakal, E., & Dağdeviren, M. (2014). Analyzing reward management framework with multi criteria decision making methods. *Procedia-Social and Behavioral Sciences*, *147*, 147-152.

Apriadi, D., Susena, K. C., & Irwanto, T. (2020). Analisis Kinerja Pegawai Pada Kantor Kesbangpol Kabupaten Kaur Performance Analysis Of Employees In Kesbangpol Office Kaur District. *Journal Bima (Business, Management And Accounting)*, *1*(2), 97-105.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Ijadi Maghsoodi, A., Abouhamzeh, G., Khalilzadeh, M., & Zavadskas, E. K. (2018). Ranking and selecting the best performance appraisal method using the MULTIMOORA approach integrated Shannon's entropy. *Frontiers of Business Research in China*, *12*(1), 2.

Kharub, M., Mondal, S., Singh, S., & Gupta, H. (2025). Evaluation of competency dimensions for employee performance assessment: evidence from micro, small, and medium enterprises. *International Journal of Productivity and Performance Management*, *74*(1), 107-138.

Na-Nan, K., Chaiprasit, K., & Pukkeeree, P. (2018). Factor analysis-validated comprehensive employee job performance scale. *International Journal of Quality & Reliability Management*, *35*(10), 2436-2449.

Shi, J. L., & Lai, W. H. (2023). Fuzzy AHP approach to evaluate incentive factors of high-tech talent agglomeration. *Expert systems with applications*, *212*, 118652.

ISSN: 2579-566X (Online)
ISSN: 2477-5274 (Print)

[halaman ini sengaja dikosongkan]