

## **PENGUKURAN KUALITAS LLM MENGGUNAKAN GLUE**

**Raihan Nabeel Alfarisi<sup>1</sup>, Andreas Hadiyono<sup>2</sup>**

Universitas Gunadarma

E-mail: [raihannabeel72@gmail.com](mailto:raihannabeel72@gmail.com)<sup>1</sup>,

[andre\\_hadiyono@gmail.com](mailto:andre_hadiyono@gmail.com)<sup>2</sup>

### **Abstrak**

Penelitian ini mengevaluasi kualitas Large Language Models (LLMs) menggunakan tolok ukur General Language Understanding Evaluation (GLUE), dengan fokus pada model LLaMA dan dataset Semantic Textual Similarity Benchmark (STS-B). Evaluasi dilakukan menggunakan Cosine Similarity, Pearson Correlation, dan Spearman Correlation untuk mengukur kesamaan semantik antara kalimat yang dihasilkan model dengan referensi manusia. Hasil penelitian menunjukkan bahwa model all-mpnet-base-v2 unggul dalam tugas kesamaan semantik, dengan skor Pearson dan Spearman Correlation sebesar 0,88, mengindikasikan kemampuannya dalam memahami hubungan antar kalimat. Sementara itu, dalam evaluasi chatbot, model LLaMA memperoleh skor Cosine Similarity sebesar 0,79, menunjukkan kinerja yang baik dalam menjawab pertanyaan umum, tetapi kurang stabil dalam domain yang lebih kompleks seperti sains dan teknologi. Hasil ini menunjukkan bahwa pemilihan model yang tepat sangat penting untuk tugas NLP tertentu. Selain itu, penelitian ini menyarankan bahwa fine-tuning pada model LLaMA dapat meningkatkan akurasi dan relevansi responsnya. Penelitian ini berkontribusi dalam pengembangan kerangka evaluasi LLM yang lebih komprehensif, dengan menekankan pentingnya metrik kesamaan dalam menilai efektivitas model bahasa.

**Kata Kunci** — LLM, GLUE, STS-B, Cosine Similarity, Pearson Correlation, Spearman Correlation, LLaMA.

### **Abstract**

This research evaluates the quality of Large Language Models (LLMs) using the General Language Understanding Evaluation (GLUE) benchmark, focusing on the LLaMA model and the Semantic Textual Similarity Benchmark (STS-B) dataset. The evaluation employs Cosine Similarity, Pearson Correlation, and Spearman Correlation to measure the semantic similarity between model-generated sentences and human-labeled references. The result shows that the all-mpnet-base-v2 model performs best in semantic similarity tasks, achieving Pearson and Spearman Correlation scores of 0.88, demonstrating its ability to understand relationships between sentences. Meanwhile, in chatbot evaluations, the LLaMA model achieved a Cosine Similarity score of 0.79, showing strong performance in answering general questions but exhibiting variability in more complex domains such as science and technology. This result highlights the importance of selecting the right model for specific NLP tasks. Additionally, this research suggests that fine-tuning the LLaMA model can further enhance its accuracy and response relevance. This study contributes to the development of a more comprehensive LLM evaluation framework, emphasizing the role of similarity metrics in assessing model effectiveness.

**Keywords:** LLM, GLUE, STS-B, Cosine Similarity, Pearson Correlation, Spearman Correlation, LLaMA.

## **PENDAHULUAN**

Artificial Intelligence (AI) telah membuat kemajuan yang signifikan, khususnya di

bidang Natural Language Processing (NLP). Large Language Models (LLMs), seperti seri LLaMA dari Meta AI, merupakan contoh dari kemajuan ini dengan kemampuannya untuk memahami dan menghasilkan teks yang mirip dengan manusia. Model-model ini dilatih pada kumpulan data yang sangat besar menggunakan arsitektur berbasis transformer, memungkinkan mereka untuk melakukan tugas-tugas mulai dari penerjemahan dan peringkasan hingga AI percakapan (Brown et al., 2020).

Perkembangan teknologi NLP telah melahirkan berbagai model bahasa besar seperti GPT, BERT, dan lainnya yang memiliki peran penting dalam berbagai aplikasi, mulai dari chatbot hingga sistem pencarian informasi. Model-model ini dirancang untuk memahami konteks dengan lebih baik, menghasilkan respons yang lebih relevan, serta mengoptimalkan interaksi antara manusia dan mesin.

Dalam konteks chatbot, cosine similarity dan pearson correlation telah menjadi metrik penting untuk mengukur relevansi respons. Cosine similarity mengevaluasi kosinus sudut antara dua vektor, yang biasanya digunakan untuk membandingkan kemiripan penyematan kata dalam tugas NLP. Di sisi lain, kesamaan Pearson mengukur korelasi linier antara dua variabel, yang juga dapat berguna dalam menilai relevansi respons chatbot dengan membandingkan kueri pengguna dan respons berdasarkan representasi vektornya. Kedua metrik tersebut sering diimplementasikan dalam sistem chatbot untuk menyempurnakan akurasi respons, memastikan bahwa respons secara kontekstual selaras dengan pertanyaan penggunaan.

Selain itu, Spearman correlation mengukur hubungan monoton antara dua variabel dengan mengandalkan peringkat daripada nilai absolut. Metrik ini berguna dalam skenario di mana hubungan antara kueri dan respons mungkin tidak linier tetapi tetap menunjukkan keterkaitan yang konsisten dalam peringkatnya. Dengan menggabungkan Cosine Similarity, Pearson Correlation, dan Spearman Correlation, sistem chatbot dapat menyempurnakan akurasi respons, memastikan bahwa respons secara kontekstual selaras dengan pertanyaan pengguna dan mempertimbangkan berbagai aspek relevansi.

Pentingnya evaluasi kualitas model bahasa menjadi aspek krusial dalam pengembangan LLM. Evaluasi ini diperlukan untuk memastikan keandalan dan akurasi model dalam berbagai tugas NLP. Salah satu tolok ukur standar yang sering digunakan adalah GLUE (General Language Understanding Evaluation), yang mengukur kemampuan model dalam memahami bahasa secara umum. GLUE mencakup berbagai tugas NLP seperti analisis sentimen, inferensi teks, dan pemahaman semantik, sehingga memberikan gambaran menyeluruh mengenai performa suatu model bahasa.

Seri LLaMA menonjol karena sifatnya yang open-source (Perrigo, 2024), sehingga dapat diakses oleh para peneliti, pengembang, dan institusi (Touvron et al., 2023). Desainnya menekankan efisiensi dalam pelatihan dan inferensi, sehingga cocok untuk beragam aplikasi, termasuk sistem waktu nyata dan bahasa yang kurang terwakili. Sementara model-model berpaten seperti GPT-3 dan GPT-4 telah menetapkan tolok ukur yang tinggi, LLaMA bertujuan untuk memberikan performa yang kompetitif dengan tuntutan komputasi yang lebih rendah, memfasilitasi adopsi yang lebih luas di seluruh domain akademis, komersial, dan komunitas (Bender et al., 2021).

Salah satu upaya untuk meningkatkan evaluasi LLM adalah dengan menggunakan benchmark seperti GLUE, yang telah menjadi acuan utama dalam mengukur pemahaman bahasa. Namun, masih terdapat kesenjangan dalam penggunaannya untuk menilai kualitas model dalam konteks sistem informasi. Oleh karena itu, mengintegrasikan metrik seperti Cosine similarity, Pearson Correlation, dan Spearman Correlation dalam evaluasi LLM dapat memberikan wawasan lebih mendalam mengenai efektivitas model dalam memberikan jawaban yang relevan. Penelitian ini berfokus pada penilaian relevansi menggunakan metrik tersebut.

## METODE PENELITIAN

### Metode Pengumpulan Data

Dalam penelitian ini, data yang digunakan berasal dari dataset General Language Understanding Evaluation (GLUE) yang tersedia secara publik. Dataset GLUE merupakan kumpulan berbagai tugas pemahaman natural language yang dirancang untuk mengevaluasi kemampuan model bahasa dalam memahami dan menganalisis teks. Dataset ini dipilih karena merupakan salah satu tolok ukur standar dalam evaluasi model bahasa alami, terutama dalam mengukur kesesuaian semantik dan relevansi jawaban yang diberikan oleh model.

Proses pengumpulan data dilakukan dengan mengunduh dataset STS-B dari repositori Huggingface. Setelah itu, dataset dikonversi ke dalam format CSV agar lebih mudah diolah dan dianalisis dalam Google Colab. Selanjutnya, data tersebut diproses menggunakan metode pemrosesan teks seperti tokenisasi, normalisasi, dan pemetaan ke representasi numerik menggunakan model Sentence Transformers ('all-mpnet-base-v2'). Hasil pengolahan ini kemudian digunakan dalam perhitungan skor relevansi menggunakan cosine similarity dan Pearson correlation, yang menjadi bagian utama dalam analisis evaluasi model bahasa dalam penelitian ini.

## HASIL DAN PEMBAHASAN

### 1. Analisis Hasil

Bagian ini mengevaluasi performa model berdasarkan eksperimen yang telah dijalankan, membandingkan hasil dengan baseline, serta memberikan interpretasi mendalam terhadap metrik evaluasi.

### 2. Performa Model pada Tugas GLUE dan Chatbot

Hasil eksperimen menunjukkan bahwa model yang digunakan memiliki performa yang bervariasi dalam tugas kesamaan teks STS-B serta dalam chatbot. Tabel berikut merangkum skor evaluasi dari masing-masing model:

Tabel 1 Perbandingan Rata-rata Performa Model pada Tugas STS-B

Model	Cosine Similarity	Pearson Correlation	Spearman Correlation
all-mpnet-base-v2	0,57	0,88	0,88
paraphrase-MiniLM-L6-v2	0,55	0,87	0,87
all-MiniLM-L6-v2	0,57	0,87	0,87

Model	Cosine Similarity	Pearson Correlation	Spearman Correlation
multi-qa-mpnet-base-dot-v1	0,71	0,75	0,73
multi-qa-MiniLM-L6-cos-v1	0,62	0,79	0,80

Tabel 1 menyajikan perbandingan lima model embedding berdasarkan evaluasi Semantic Textual Similarity Benchmark (STS-B) pada 1.500 pasang kalimat dalam set validasi (STS-dev). Hasilnya dihitung berdasarkan rata-rata skor dari pasangan kalimat yang diuji. Evaluasi dilakukan menggunakan tiga metrik utama:

- **Cosine Similarity:** Mengukur kesamaan berbasis vektor antara dua teks. Nilai yang lebih tinggi menunjukkan embedding yang lebih mirip.
- **Pearson Correlation:** Mengukur hubungan linear antara skor kesamaan model dengan skor referensi manusia. Semakin tinggi nilainya, semakin baik model dalam menyesuaikan dengan penilaian manusia.
- **Spearman Correlation:** Mengukur hubungan monoton antara skor model dan skor

referensi manusia. Nilai tinggi menunjukkan model mampu mempertahankan peringkat kesamaan yang sesuai dengan manusia.

Berikut adalah analisis performa setiap model berdasarkan hasil dalam tabel:

- 1) all-mpnet-base-v2
  - Memiliki Pearson Correlation (0,88) dan Spearman Correlation (0,88) tertinggi di antara semua model, menunjukkan bahwa model ini memiliki pemahaman yang sangat baik terhadap hubungan semantik antara teks.
  - Cosine Similarity-nya (0,57) tidak terlalu tinggi dibanding model lain, tetapi tetap kompetitif.
  - Model ini cocok untuk tugas STS-B yang memerlukan pemahaman kesamaan semantik yang akurat.
- 2) paraphrase-MiniLM-L6-v2
  - Memiliki skor Pearson (0,87) dan Spearman (0,87) yang hampir sama dengan all-mpnet-base-v2, menunjukkan bahwa model ini juga dapat menangkap hubungan semantik dengan baik.
  - Cosine Similarity-nya (0,55) sedikit lebih rendah dibanding all-mpnet-base-v2, tetapi masih dalam rentang yang baik.
  - Model ini merupakan pilihan yang baik untuk pemrosesan bahasa alami yang memerlukan pemahaman parafrase.
- 3) all-MiniLM-L6-v2
  - Skor Cosine Similarity (0,57) sama dengan all-mpnet-base-v2, tetapi Pearson (0,87) dan Spearman (0,87) sama dengan paraphrase-MiniLM-L6-v2.
  - Performa keseluruhan mirip dengan paraphrase-MiniLM-L6-v2, sehingga model ini juga dapat digunakan untuk tugas STS-B dengan hasil yang serupa.
  - Model ini lebih ringan dibandingkan all-mpnet-base-v2, yang bisa menjadi keuntungan dalam lingkungan dengan sumber daya terbatas.
- 4) multi-qa-mpnet-base-dot-v1
  - Memiliki skor Cosine Similarity tertinggi (0,71), yang berarti model ini menghasilkan embedding yang sangat dekat untuk teks yang dianggap mirip.
  - Namun, skor Pearson (0,75) dan Spearman (0,73) lebih rendah dibanding model lain, menunjukkan bahwa prediksinya kurang sesuai dengan skor manusia.
  - Model ini lebih dioptimalkan untuk tugas retrieval (misalnya pencarian dokumen atau sistem tanya-jawab) dibandingkan STS-B murni.
- 5) multi-qa-MiniLM-L6-cos-v1
  - Skor Cosine Similarity-nya (0,62) lebih tinggi dibanding model STS-B lainnya, tetapi lebih rendah dibanding multi-qa-mpnet-base-dot-v1.
  - Skor Pearson (0,79) dan Spearman (0,80) lebih tinggi dibanding multi-qa-mpnet-base-dot-v1, menunjukkan bahwa model ini lebih baik dalam menyesuaikan dengan skor manusia.
  - Model ini dapat menjadi alternatif jika membutuhkan keseimbangan antara pemahaman semantik dan kemampuan retrieval.

Dari hasil ini, all-mpnet-base-v2 menjadi pilihan terbaik untuk tugas STS-B karena memiliki keselarasan tertinggi dengan penilaian manusia. Jika membutuhkan model yang lebih ringan, paraphrase-MiniLM-L6-v2 dan all-MiniLM-L6-v2 bisa menjadi alternatif yang cukup akurat. Namun, jika fokus utama adalah pencarian berbasis embedding, maka multi-qa-mpnet-base-dot-v1 lebih unggul dalam menemukan pasangan teks yang mirip, meskipun kurang akurat dalam peringkat kesamaannya. Model multi-qa-MiniLM-L6-cos-v1 dapat menjadi pilihan yang seimbang antara retrieval dan pemahaman semantik.

Untuk chatbot, skor evaluasi menunjukkan sejauh mana model mampu memahami pertanyaan pengguna:

Tabel 2 Evaluasi Performa Chatbot dengan Skor Similaritas

Kategori Pertanyaan	Cosine Similarity	Pearson Correlation	Spearman Correlation	STS-B (Score)
Teknologi	0,74	0,75	0,73	4,5
Sains	0,68	0,69	0,67	4,4
Sosial	0,74	0,75	0,73	4,4
Umum	0,79	0,80	0,78	4,6

Berikut adalah penjelasan hasil perhitungan:

- Kategori "Umum" memiliki performa tertinggi, dengan Cosine Similarity (0,79), Pearson Correlation (0,80), Spearman Correlation (0,78), dan STS-B Score (4,6). Hal ini menunjukkan bahwa chatbot lebih akurat dalam menjawab pertanyaan umum, kemungkinan karena jenis pertanyaan ini lebih sering ditemukan dalam data pelatihan dan memiliki jawaban yang lebih jelas serta tidak terlalu bergantung pada konteks yang kompleks.
  - Kategori "Teknologi" dan "Sosial" memiliki skor yang hampir sama dengan Cosine Similarity (0,74) dan STS-B Score (4,4–4,5). Ini menunjukkan bahwa chatbot dapat menangani pertanyaan dalam domain ini dengan baik. Skor Pearson dan Spearman yang juga tinggi mengindikasikan bahwa chatbot cukup konsisten dalam memberikan jawaban yang relevan untuk pertanyaan di bidang ini.
  - Kategori "Sains" memiliki performa terendah, dengan Cosine Similarity (0,68) dan STS-B Score (4,4). Ini menunjukkan bahwa chatbot memiliki sedikit kesulitan dalam memahami dan menjawab pertanyaan sains dibandingkan kategori lainnya. Performa yang lebih rendah ini kemungkinan disebabkan oleh kompleksitas terminologi ilmiah atau kurangnya data pelatihan yang mencakup berbagai konsep sains secara mendalam. Selain itu, jawaban dalam bidang sains sering kali membutuhkan pemahaman yang lebih spesifik, sehingga chatbot mungkin mengalami kesulitan dalam memberikan respons yang benar-benar sesuai dengan maksud pertanyaan.
  - Selisih skor antar kategori tidak terlalu besar, yang menunjukkan bahwa chatbot masih memiliki performa yang cukup baik di semua kategori pertanyaan. Namun, dengan adanya perbedaan kecil ini, dapat disimpulkan bahwa chatbot lebih unggul dalam menjawab pertanyaan umum dibandingkan pertanyaan yang lebih spesifik seperti dalam kategori sains.
3. Interpretasi Hasil Berdasarkan Metrik Evaluasi
1. Analisis Kesamaan Kalimat Menggunakan Dataset STS-B
- Pada tahap ini, dilakukan evaluasi terhadap kesamaan semantik antar kalimat menggunakan dataset STS-B (Semantic Textual Similarity Benchmark). Dataset ini menyediakan pasangan kalimat dengan skor kesamaan semantik dari 0 hingga 5, di mana:
- Skor 5 menunjukkan kalimat yang hampir identik secara makna.
  - Skor 0 menunjukkan kalimat yang tidak memiliki kesamaan makna.
  - Untuk mengukur tingkat kemiripan antar kalimat, digunakan metrik cosine, pearson dan spearman:
  - Cosine Similarity: Mengukur kemiripan berbasis vektor representasi kata.
  - Pearson Correlation: Mengukur hubungan linier antara vektor kalimat.
  - Spearman Correlation: Mengukur hubungan peringkat antara vektor kalimat.

sentence1	sentence2	cosine_similarity	pearson_correlation	spearman_correlation	true_score	similarity_label
A man with a hard hat is dancing.	A man wearing a hard hat is dancing.	0.99670535	0.9967052873249665	0.99578639481675	5.0	Highly Similar
A young child is riding a horse.	A child is riding a horse.	0.9509652	0.9509655657150905	0.9461867426753451	4.75	Highly Similar
A man is feeding a mouse to a snake.	The man is feeding a mouse to the snake.	0.8543391	0.8543387406263035	0.8373455627790033	5.0	Highly Similar
A woman is playing the guitar.	A man is playing guitar.	0.5948903	0.5948579921741448	0.5577261366869413	2.400000095367432	Moderately Similar
A woman is playing the flute.	A man is playing a flute.	0.735644	0.735633080532547	0.6907523528244914	2.75	Moderately Similar
A woman is cutting an onion.	A man is cutting onions.	0.7216712	0.7216648145170685	0.6840275069809079	2.615000009536743	Moderately Similar
A man is erasing a chalk board.	The man is erasing the chalk board.	0.9559711	0.9559695888018813	0.9507990108897076	5.0	Highly Similar
A woman is carrying a boy.	A woman is carrying her baby.	0.6623735	0.6623805810129311	0.6342693062155934	2.332999944668896	Moderately Similar
Three men are playing guitars.	Three men are on stage playing guitars.	0.88425547	0.8842557790851926	0.8787928486596827	3.75	Highly Similar
A woman peels a potato.	A woman is peeling a potato.	0.9432427	0.9432426689063434	0.9285325057262942	5.0	Highly Similar

Gambar 2 Hasil Perhitungan Metrik dengan STS-B

Penjelasan dari hasil perhitungan metrik dengan STS-B:

1. Konsistensi antara Metrik Cosine Similarity, Pearson, dan Spearman menunjukkan pola konsisten dengan True Score.

Ketiga metrik menunjukkan pola yang konsisten dengan nilai true score dari STS-B:

- Jika true score tinggi, cosine similarity, pearson, dan spearman juga tinggi.
- Jika terdapat sedikit perbedaan dalam makna atau struktur, skor akan menurun tetapi masih berada dalam rentang cukup tinggi (~0.6-0.85), menunjukkan bahwa kedua kalimat masih memiliki kesamaan yang kuat meskipun tidak identik.

Sebagai contoh:

- Kalimat: "A young child is riding a horse." vs. "A child is riding a horse."
  - True Score: 4.75
  - Cosine Similarity: 0.9509652
  - Pearson Correlation: 0.9509656
  - Spearman Correlation: 0.9461867
- Karena kedua kalimat hanya memiliki perbedaan kecil ("young child" vs. "child"), skor tetap sangat tinggi.
2. Highly Similar kalimat memiliki perbedaan minimal dalam kata atau struktur.

Kalimat dalam kategori Highly Similar memiliki perbedaan minimal dalam kata-kata atau struktur. Meskipun ada sedikit perbedaan dalam kata, makna tetap sangat mirip, sehingga skor kesamaan tinggi.

Contoh:

- Kalimat: "A man with a hard hat is dancing." vs. "A man wearing a hard hat is dancing."
  - True Score: 5.0
  - Cosine Similarity: 0.99670535
  - Pearson Correlation: 0.99670528
  - Spearman Correlation: 0.99578639
- Perbedaan antara "with a hard hat" dan "wearing a hard hat" sangat kecil sehingga tidak mempengaruhi makna.
3. Moderately Similar kalimat memiliki perbedaan pada subjek, objek, atau kata kerja utama.

Kalimat dalam kategori Moderately Similar memiliki perbedaan lebih signifikan, terutama dalam subjek, objek, atau kata kerja utama. Perbedaan ini menyebabkan skor cosine similarity, pearson, dan spearman lebih rendah (~0.6-0.75), menunjukkan bahwa kalimat masih memiliki kesamaan tetapi dengan makna yang sedikit berbeda.

Contoh:

- Kalimat: "A woman is carrying a boy." vs. "A woman is carrying her baby."
  - True Score: 2.33
  - Cosine Similarity: 0.6623735
  - Pearson Correlation: 0.6623805
  - Spearman Correlation: 0.6342890
- Perbedaan antara "boy" dan "baby" membuat kalimat ini tidak identik, meskipun tindakan yang dilakukan sama.
4. Spearman dan Pearson Correlation cenderung mengikuti pola Cosine Similarity.

Spearman dan Pearson correlation cenderung mengikuti pola yang sama dengan cosine similarity:

- Jika cosine similarity tinggi, Spearman dan Pearson juga tinggi.
  - Jika cosine similarity sedang atau rendah, Spearman dan Pearson juga menurun.
- Hal ini menunjukkan bahwa ketiga metrik dapat digunakan secara bersamaan untuk mengevaluasi kesamaan semantik antar kalimat dengan hasil yang serupa.
- Sebagai contoh:
- Kalimat: "Three men are playing guitars." vs. "Three men are on stage playing guitars."
  - Cosine Similarity: 0.88425547
  - Pearson Correlation: 0.88425577
  - Spearman Correlation: 0.87879248
- Ketiga metrik memiliki nilai yang sangat mirip, menunjukkan kesamaan yang tinggi antara dua kalimat.

## 2. Evaluasi Jawaban Chatbot terhadap True Answer

Chatbot yang diuji memberikan respons berdasarkan input yang diterima, kemudian hasilnya dibandingkan dengan true answer menggunakan metrik evaluasi seperti Cosine Similarity, Pearson Correlation, dan Spearman Correlation. Berikut adalah hasil berdasarkan skornya.

Input	Chatbot Response	True Answer	Cosine Similarity	Pearson Correlation	Spearman Correlation	True Label
A man with a hard hat is dancing.	A man with a hard hat is dancing. He's wearing a bright	A man wearing a hard hat is dancing.	0.9606	0.9606	0.9538	5.0
A young child is riding a horse.	A young child is riding a horse. The child is wearing a	A child is riding a horse.	0.8948	0.8948	0.8852	4.75
A man is feeding a mouse to a snake.	A man is feeding a mouse to a snake. The snake is hungry and	The man is feeding a mouse to the snake.	0.7324	0.7324	0.7041	5.0
A woman is playing the guitar.	A woman is playing the guitar. She's wearing a pair	A man is playing guitar.	0.4766	0.4766	0.4334	2.4000000953674316
A woman is playing the flute.	A woman is playing the flute. She is sitting on a	A man is playing a flute.	0.6382	0.6381	0.6043	2.75
A woman is cutting an onion.	A woman is cutting an onion. She is crying, and	A man is cutting onions.	0.5740	0.5740	0.5376	2.615000009536743
A man is erasing a chalk board.	A man is erasing a chalk board. As he erases,	The man is erasing the chalk board.	0.8010	0.8010	0.7754	5.0
A woman is carrying a boy.	A woman is carrying a boy. The boy is 6	A woman is carrying her baby.	0.5554	0.5555	0.5166	2.3329999446868896
Three men are playing guitars.	Three men are playing guitars. One man is a beginner	Three men are on stage playing guitars.	0.7706	0.7706	0.7529	3.75
A woman peels a potato.	A woman peels a potato. She takes a bite,	A woman is peeling a potato.	0.7857	0.7857	0.7441	5.0

Gambar 3 Evaluasi Jawaban Chatbot terhadap True Answer

Penjelasan dari Evaluasi Jawaban Chatbot terhadap True Answer:

### 1. Kinerja chatbot sangat baik untuk jawaban dengan skor tinggi (True Label 5.0).

Chatbot memberikan jawaban yang sangat akurat, hampir identik dengan True Answer

Sebagai contoh:

- Input: "A man with a hard hat is dancing."
- Chatbot Response: "A man wearing a hard hat is dancing. He's wearing a bright."
- True Answer: "A man wearing a hard hat is dancing."
- Skor: Cosine (0.9606), Pearson (0.9606), dan Spearman (0.9538).

Hasil ini menunjukkan jawaban chatbot sangat mirip dengan jawaban yang benar, menunjukkan pemahaman konteks yang baik.

### 2. Untuk kasus dengan skor sedang (~2.4 - 3.75), jawaban chatbot mengandung beberapa variasi tetapi masih cukup relevan.

Meskipun masih berkaitan dengan konteks pertanyaan, chatbot

Sebagai contoh:

- Input: "A woman is playing the flute."
- Chatbot Response: "A woman is playing the flute. She is sitting on a",
- True Answer: "A man is playing the flute."
- Skor: Cosine (0.6382), Pearson (0.6381), dan Spearman (0.6043).

Chatbot berhasil menangkap aktivitas yang dilakukan (bermain seruling) tetapi tidak membedakan subjek dengan benar (pria vs wanita).

### 3. Kinerja Chatbot untuk Skor rendah (<2.5) menunjukkan jawaban yang kurang sesuai dengan True Answer.

Jawaban chatbot sering kali mengandung tambahan informasi yang tidak relevan atau salah memahami konteks.

Sebagai contoh:

- Input: "A woman is playing the guitar."
- Chatbot Response: "A woman is playing the guitar. She's wearing a pair"
- True Answer: "A man is playing guitar."
- Skor: Cosine (0.4766), Pearson (0.4766), dan Spearman (0.4344).

Dalam kasus ini, chatbot tidak hanya salah mengenali subjek (wanita vs pria), tetapi juga menambahkan informasi yang tidak ada dalam True Answer, sehingga membuat jawaban kurang relevan.

#### 4. Keterbatasan Metode

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan:

##### 1. Tidak dilakukan fine-tuning pada model LLaMA

Fine-tuning berpotensi meningkatkan akurasi model dalam memahami dan menghasilkan respons yang lebih relevan. Namun, proses ini membutuhkan sumber daya komputasi yang lebih besar, yang tidak tersedia dalam penelitian ini.

##### 2. Dataset yang digunakan hanya STS-B dan chatbot

Evaluasi model hanya dilakukan menggunakan dataset STS-B dari GLUE dan data chatbot. Hal ini membatasi generalisasi hasil penelitian ke tugas Natural Language Processing (NLP) lainnya, seperti klasifikasi teks, summarization, atau machine translation.

##### 3. Keterbatasan platform eksekusi

Seluruh eksperimen dilakukan di Google Colab yang memiliki keterbatasan dalam eksekusi skala besar, terutama dalam hal daya komputasi dan memori. Hal ini dapat membatasi eksplorasi model yang lebih kompleks atau dataset yang lebih besar.

#### 5. Implikasi Hasil Penelitian terhadap Pengembangan Sistem Informasi

Hasil penelitian ini menunjukkan bahwa model all-mpnet-base-v2 memiliki performa superior dalam tugas kesamaan teks STS-B, sementara model LLaMA, meskipun menunjukkan potensi dalam aplikasi chatbot, masih memerlukan optimasi lebih lanjut. Implikasinya bagi pengembangan sistem informasi adalah pentingnya pemilihan model bahasa yang sesuai dengan tugas spesifik yang dihadapi. Integrasi model yang telah terlatih dengan baik, seperti all-mpnet-base-v2, dapat meningkatkan akurasi dan efisiensi sistem informasi yang memerlukan pemahaman kesamaan teks. Namun, untuk aplikasi yang lebih interaktif seperti chatbot, diperlukan penyesuaian dan fine-tuning model seperti LLaMA agar respons yang dihasilkan lebih relevan dan kontekstual. Selain itu, penelitian ini menyoroti pentingnya evaluasi metrik seperti cosine similarity, Pearson correlation, dan Spearman correlation dalam menilai performa model, yang dapat menjadi acuan dalam pengembangan dan peningkatan sistem informasi berbasis teks.

#### 6. Perbandingan Hasil dengan Penelitian Sebelumnya

Penelitian ini membandingkan performa model dalam mengukur kesamaan teks menggunakan cosine similarity, Pearson correlation, dan Spearman correlation. Untuk menilai relevansi hasil, dilakukan perbandingan dengan penelitian sebelumnya dari jurnal "Analisis Perbandingan Pearson Correlation dan Cosine Similarity pada Rekomendasi Musik berbasis Collaborative Filtering".

##### 1. Persamaan Pendekatan

- Kedua penelitian menggunakan Pearson Correlation dan Cosine Similarity untuk mengukur kesamaan antar entitas.
- Fokus penelitian ini adalah kesamaan teks pada chatbot menggunakan dataset GLUE STS-B, sedangkan jurnal sebelumnya menggunakan rating pengguna dalam sistem rekomendasi musik.
- Kedua penelitian menganalisis perbedaan hasil antara Pearson dan Cosine dalam menilai hubungan antar data.

##### 2. Perbedaan Konteks dan Metodologi

Aspek	Penelitian Ini (Chatbot)	Jurnal Rekomendasi Musik
Domain	Natural Language Processing (NLP).	Rekomendasi Musik (Collaborative Filtering).

<b>Dataset</b>	GLUE STS-B (kesamaan teks).	Dataset rating pengguna lagu.
<b>Metode Evaluasi</b>	Cosine, Pearson, Spearman.	Cosine, Pearson.

Aspek	Penelitian Ini (Chatbot)	Jurnal Rekomendasi Musik
<b>Tujuan</b>	Mengukur kesamaan antar kalimat dalam chatbot.	Mengukur kesamaan preferensi pengguna untuk rekomendasi musik.
<b>Output</b>	Skor kesamaan antar teks.	Skor kesamaan antar pengguna atau lagu.

### 3. Perbandingan Hasil Evaluasi

Metode	Penelitian Ini (Chatbot)	Jurnal Rekomendasi Musik
<b>Cosine Similarity</b>	0.74 (Chatbot)	MAE = 0.066559, RMSE = 0.086709
<b>Pearson Correlation</b>	0.75 (Chatbot)	MAE = 0.066538, RMSE = 0.086698
<b>Spearman Correlation</b>	0.73 (Chatbot)	Tidak digunakan.

Penjelasan:

- Dalam penelitian ini, Cosine Similarity memiliki skor 0.74 dalam menilai kesamaan teks chatbot, sementara dalam jurnal rekomendasi musik, Cosine digunakan dengan hasil MAE = 0.066559, RMSE = 0.086709.
- Pearson Correlation lebih tinggi dari Cosine di chatbot (0.75), sejalan dengan hasil pada jurnal rekomendasi musik (MAE = 0.066538, RMSE = 0.086698), menunjukkan bahwa Pearson menangkap pola linear lebih baik.
- Spearman Correlation hanya digunakan dalam penelitian ini dengan skor 0.73, sementara dalam jurnal rekomendasi musik tidak diterapkan.

Hasil ini menunjukkan bahwa pendekatan Pearson dan Cosine dapat diadaptasi untuk berbagai aplikasi dengan kebutuhan yang berbeda. Chatbot memerlukan pemahaman konteks teks, sedangkan sistem rekomendasi musik lebih bergantung pada pola numerik dalam data rating pengguna.

### 7. Rekomendasi untuk Penelitian Lanjutan

#### 1. Melakukan fine-tuning LLaMA dengan dataset STS-B dan dataset percakapan

Selain meningkatkan performa model dalam memahami kesamaan teks, fine-tuning pada dataset percakapan dapat membantu model menghasilkan respons yang lebih alami dan relevan dalam interaksi chatbot.

#### 2. Mengeksplorasi teknik pengayaan data atau augmentasi

Data augmentation, seperti parafrase otomatis, sinonim substitution, atau back-translation, dapat digunakan untuk meningkatkan variasi data pelatihan dan memperkuat kemampuan model dalam menangani berbagai ekspresi bahasa.

#### 3. Menguji model pada tugas NLP lainnya

Selain kesamaan teks, model dapat diuji pada tugas lain seperti klasifikasi sentimen, natural language inference (NLI), atau ekstraksi informasi untuk memahami sejauh mana performanya dapat digeneralisasi ke berbagai aplikasi NLP.

## KESIMPULAN

Penelitian ini mengevaluasi performa model bahasa besar (LLM) dalam tugas kesamaan teks menggunakan dataset STS-B dari benchmark GLUE. Fokus utama penelitian ini adalah mengukur relevansi jawaban dengan menggunakan tiga metrik utama: Cosine Similarity, Pearson Correlation, dan Spearman Correlation. Selain itu, penelitian ini

juga mengevaluasi efektivitas chatbot berbasis LLaMA dalam menjawab berbagai kategori pertanyaan.

Hasil eksperimen menunjukkan bahwa model all-mpnet-base-v2 memiliki performa terbaik dalam tugas kesamaan teks, dengan Pearson Correlation dan Spearman Correlation sebesar 0,88, menunjukkan bahwa model ini memiliki pemahaman yang sangat baik terhadap hubungan semantik antar teks. Sementara itu, chatbot berbasis LLaMA menunjukkan performa yang cukup baik dalam menjawab pertanyaan, dengan kategori pertanyaan umum memiliki skor tertinggi (Cosine Similarity 0,79; Pearson Correlation 0,80; Spearman Correlation 0,78; dan STS-B Score 4,6), yang mengindikasikan bahwa chatbot lebih mampu menangani pertanyaan bersifat umum dibandingkan kategori lainnya.

Kontribusi utama penelitian ini terhadap bidang sistem informasi dan NLP adalah sebagai berikut:

1. Evaluasi Model Bahasa dengan Metrik Similaritas: Penelitian ini memberikan analisis komprehensif mengenai Cosine Similarity, Pearson Correlation, dan Spearman Correlation dalam mengukur relevansi respons model bahasa.
2. Implementasi GLUE STS-B untuk Analisis Kesamaan Teks: Penggunaan dataset STS-B dari GLUE menunjukkan efektivitas model dalam memahami kesamaan semantik antar teks.
3. Penggunaan LLaMA dalam Chatbot: Penelitian ini menunjukkan bahwa LLaMA dapat memberikan respons yang cukup relevan, meskipun masih perlu dilakukan optimalisasi lebih lanjut.
4. Pengaruh Kategori Pertanyaan terhadap Performa Chatbot: Hasil menunjukkan bahwa chatbot lebih baik dalam menangani pertanyaan umum dibandingkan dengan kategori sains dan teknologi, yang mengindikasikan bahwa kualitas respons masih bervariasi tergantung pada kompleksitas domain pertanyaan.

### **Saran**

Berdasarkan hasil penelitian ini, terdapat beberapa saran yang dapat dijadikan acuan untuk pengembangan lebih lanjut:

1. Fine- Optimasi Model LLaMA dengan Fine-Tuning

Model LLaMA masih dapat dioptimalkan lebih lanjut dengan teknik fine-tuning, terutama dengan dataset STS-B serta dataset percakapan lainnya. Dengan pendekatan ini, model diharapkan mampu memberikan respons yang lebih relevan, akurat, dan kontekstual terhadap pertanyaan pengguna.

2. Penggunaan Benchmark Tambahan

Selain menggunakan GLUE STS-B, penelitian selanjutnya dapat mempertimbangkan benchmark lain seperti SuperGLUE atau MS MARCO. Benchmark tambahan ini dapat membantu mengevaluasi pemahaman model terhadap teks dalam berbagai konteks yang lebih kompleks dan bervariasi.

3. Pengembangan Sistem Chatbot yang Lebih Adaptif

Hasil penelitian ini menunjukkan bahwa chatbot berbasis LLaMA memberikan respons yang cukup baik, tetapi masih memiliki keterbatasan dalam menjawab pertanyaan dengan tingkat kompleksitas tinggi. Oleh karena itu, integrasi dengan pendekatan Retrieval-Augmented Generation (RAG) atau sistem berbasis pengetahuan dapat dipertimbangkan untuk meningkatkan pemahaman chatbot terhadap pertanyaan pengguna.

### **DAFTAR PUSTAKA**

- Brown, T.B. et al., (2020). Language Models are Few-Shot Learners. Available at: <http://arxiv.org/abs/2005.14165>.
- Perrigo, B. (2024). 'Mark Zuckerberg's Meta unveils open-source AI model LLaMA', TIME Magazine, 24 July. Available at: [https://time.com/7002563/mark-zuckerberg-ai-llama-meta-open-source/?utm\\_source=chatgpt.com](https://time.com/7002563/mark-zuckerberg-ai-llama-meta-open-source/?utm_source=chatgpt.com) (Accessed: 18 December 2024)
- Touvron, H. et al., (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. Available at: <http://arxiv.org/abs/2307.09288>.

- Bender, E.M. et al., (2021). On the dangers of stochastic parrots: Can language models be too big? FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. 3 March 2021 Association for Computing Machinery, Inc, pp. 610–623.
- Naveed, H. et al., (2023). A Comprehensive Overview of Large Language Models. Available at: <http://arxiv.org/abs/2307.06435>.
- Lumbansiantar, S., Dwiasnati, S., Fatonah, N.S., (2023). Penerapan Metode Cosine Similarity Dalam Mendeteksi Plagiarisme Pada Jurnal. Format : Jurnal Ilmiah Teknik Informatika 12, 142. doi:10.22441/format.2023.v12.i2.007
- Wang, A. et al., (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. Available at: <http://arxiv.org/abs/1804.07461>.
- Burgund, D., Nikolovski, S., Galić, D. and Maravić, N., (2023). Pearson Correlation in Determination of Quality of Current Transformers. Sensors, 23(5).
- Touvron, H. et al., (2023). LLaMA: Open and Efficient Foundation Language Models. Available at: <http://arxiv.org/abs/2302.13971>.
- Lee, S. et al., (2023). A Survey on Evaluation Metrics for Machine Translation. Mathematics, 11(4). <https://doi.org/10.3390/math11041006>
- Tholib, A., (2023). Implementasi Algoritma Machine Learning Berbasis Web dengan Framework Streamlit, Universitas Nusantara PGRI Kediri.