



Next-Gen AI Customer Service

Gregorius Reynaldi Pratama

11 January 2026

AI UNTUK
SEMUA

Meet The Team



Gregorius Reynaldi Pratama



<https://github.com/GregReynaldi>

Background

Evolusi Layanan Pelanggan Digital:

- ✓ Ekspektasi pelanggan modern menuntut respon yang tidak hanya cepat, tetapi juga intuitif. Peralihan dari teks statis ke interaksi berbasis suara (Voice-first) menuntut sistem yang mampu memahami keluhan pelanggan secara alami melalui input audio maupun teks secara 24/7.

Limitasi Model Tradisional & Fenomena Hallucination:

- ✓ Model bahasa besar (LLM) berisiko mengalami "halusinasi" atau memberikan jawaban yang tidak sesuai fakta perusahaan. Penggunaan dataset Customer Support Tickets menjadi penting untuk melatih model dalam memahami masalah spesifik.

Sinergi BERT, RAG, dan LangChain:

- ✓ Proyek ini menggabungkan kekuatan BERT untuk pemahaman semantik yang presisi dengan arsitektur Retrieval-Augmented Generation (RAG). RAG memungkinkan sistem untuk "melihat" dokumen referensi atau basis pengetahuan internal secara real-time, sehingga jawaban yang dihasilkan selalu akurat dan relevan.

Interaksi Masa Depan: Voice Recognition & AI Agent:

- ✓ Integrasi teknologi Speech-to-Text (STT) memungkinkan pelanggan untuk menyampaikan keluhan secara lisan. Voice Recognition bertindak sebagai gerbang utama, mengubah audio menjadi teks.



Tujuan Project

Automated Classification:

- ✓ Menggunakan BERT untuk mengkategorikan tiket keluhan secara otomatis dengan akurasi tinggi.

Accurate Solutions:

- ✓ Mengintegrasikan RAG & LangChain untuk memberikan jawaban solusi yang faktual berdasarkan dokumen referensi (bukan asumsi model).

Voice Interaction:

- ✓ Mengimplementasikan Speech Recognition agar pelanggan bisa menyampaikan keluhan melalui input suara.

Operational Efficiency:

- ✓ Mempercepat waktu respon (response time) dan mengurangi beban kerja manual pada layanan pelanggan.

Enhanced Accessibility:

- ✓ Meningkatkan aksesibilitas layanan melalui antarmuka multi-modal (suara & teks) yang user-friendly.



Cakupan Penelitian

Klasifikasi Tiket Otomatis via BERT

- ✓ Fokus penggunaan model BERT & mengategorikan tiket dukungan pelanggan. Sistem dirancang untuk memahami konteks bahasa alami yang kompleks guna memastikan setiap keluhan diarahkan ke departemen yang tepat secara otomatis.

Integrasi RAG dan LangChain

- ✓ Cakupan ini mencakup pembangunan basis pengetahuan menggunakan metode Retrieval-Augmented Generation (RAG). Dengan bantuan framework LangChain, sistem tidak hanya mengklasifikasi masalah, tetapi juga melakukan pencarian dokumen referensi internal secara cerdas untuk menghasilkan jawaban yang faktual, akurat, dan bukan sekadar asumsi model.

Antarmuka Berbasis Suara (Voice Input)

- ✓ Sistem mendukung input teks dan suara melalui teknologi Speech-to-Text. Keluhan verbal dikonversi menjadi data terstruktur untuk diproses oleh pipeline AI, guna meningkatkan aksesibilitas dan kemudahan interaksi bagi pelanggan.

Analisis Performa dan Optimasi Model

- ✓ Penelitian mencakup evaluasi metrik akurasi dan loss untuk mengukur kinerja model. Fokus analisis terletak pada penanganan overfitting serta efektivitas sistem dalam menghasilkan respon faktual dari dokumen referensi yang tersedia.



Data Collection

Sumber Data:

Menggunakan Multilingual Customer Support Tickets Dataset dari Kaggle, khususnya subset Bahasa Inggris, yang berisi tiket dukungan pelanggan dari berbagai industri. Dataset ini mencakup pertanyaan, jawaban, dan subjek tiket yang memberikan informasi tentang interaksi pelanggan dengan layanan dukungan.

Struktur Dataset:

- Queue: Menentukan departemen yang menerima tiket, seperti Technical Support, Customer Service, Billing, dll.
- Priority: Menunjukkan urgensi tiket, dengan nilai seperti Low, Medium, dan Critical.
- Language: Bahasa yang digunakan dalam tiket, seperti EN (Inggris), DE (Jerman), dll.
- Subject: Subjek dari tiket pelanggan.
- Body: Isi email atau pertanyaan pelanggan.
- Answer: Respon dari agen dukungan.
- Type: Jenis tiket, seperti Incident, Request, Problem, dll.
- Tags: Kategori tambahan yang diberikan pada tiket untuk mempermudah klasifikasi, misalnya Bug, Feature Request, dll.



Exploratory Data Analysis - Dataset Raw

```
<class 'pandas.core.frame.DataFrame'>
Index: 28587 entries, Wesentlicher Sich
Data columns (total 15 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   body         28587 non-null    object  
 1   answer       28580 non-null    object  
 2   type         28587 non-null    object  
 3   queue        28587 non-null    object  
 4   priority     28587 non-null    object  
 5   language     28587 non-null    object  
 6   version      28587 non-null    int64   
 7   tag_1        28587 non-null    object  
 8   tag_2        28574 non-null    object  
 9   tag_3        28451 non-null    object  
 10  tag_4        25529 non-null    object  
 11  tag_5        14545 non-null    object  
 12  tag_6        5874 non-null    object  
 13  tag_7        2040 non-null    object  
 14  tag_8        565 non-null     object  
dtypes: int64(1), object(14)
memory usage: 3.5+ MB
```

body	0
answer	7
type	0
queue	0
priority	0
language	0
version	0
tag_1	0
tag_2	13
tag_3	136
tag_4	3058
tag_5	14042
tag_6	22713
tag_7	26547
tag_8	28022
	dtype: int64

Shape Dataset : (28587, 15)

Jumlah Entri dan Kolom:

- Dataset memiliki 28.587 entri dan 15 kolom, dengan shape (28.587, 15).

Tipe Data Kolom:

- Kolom dalam dataset terdiri dari tipe data object (untuk teks dan kategori) dan int64 (untuk versi)

Nilai Non-Null:

- Semua kolom utama (body, answer, type, queue, priority, language, version) tidak memiliki nilai kosong (non-null).
- tag_1 hingga tag_4 memiliki sebagian besar entri terisi, dengan tag_1 dan tag_2 yang tidak memiliki banyak nilai kosong.
- tag_8 mencatatkan nilai non-null terbanyak (28.022), sedangkan tag_6 dan tag_7 juga memiliki lebih dari 22.000 nilai terisi.

Penurunan Nilai Non-Null:

- Kolom body, answer, type, queue, priority, dan language memiliki nilai penuh (0 nilai kosong).
- Kolom tag_1 hingga tag_8 menunjukkan distribusi nilai kosong yang bervariasi, dengan tag_2 dan tag_3 yang lebih banyak kosong dibandingkan tag lainnya.

Exploratory Data Analysis - Dataset Clean

```
<class 'pandas.core.frame.DataFrame'>
Index: 16335 entries, 1 to 28586
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype  
---  --          -----          object 
 0   subject     13729 non-null   object 
 1   body        16335 non-null   object 
 2   answer       16335 non-null   object 
 3   type         16335 non-null   object 
 4   queue        16335 non-null   object 
 5   priority     16335 non-null   object 
 6   language     16335 non-null   object 
 7   version      16335 non-null   int64  
 8   tag_1        16335 non-null   object 
 9   tag_2        16329 non-null   object 
 10  tag_3        16266 non-null   object 
 11  tag_4        14626 non-null   object 
 12  tag_5        8415 non-null   object 
 13  tag_6        3370 non-null   object 
 14  tag_7        1134 non-null   object 
 15  tag_8        281 non-null    object 
dtypes: int64(1), object(15)
memory usage: 2.1+ MB
```

subject	2606
body	0
answer	0
type	0
queue	0
priority	0
language	0
version	0
tag_1	0
tag_2	6
tag_3	69
tag_4	1709
tag_5	7920
tag_6	12965
tag_7	15201
tag_8	16054
dtype:	int64

Shape Dataset : (16335, 16)

Jumlah Entri dan Kolom:

- Dataset terdiri dari 16.335 entri dan 16 kolom, dengan shape (16.335, 16).

Tipe Data Kolom:

- Mayoritas kolom bertipe object (teks), kecuali kolom version yang bertipe int64.

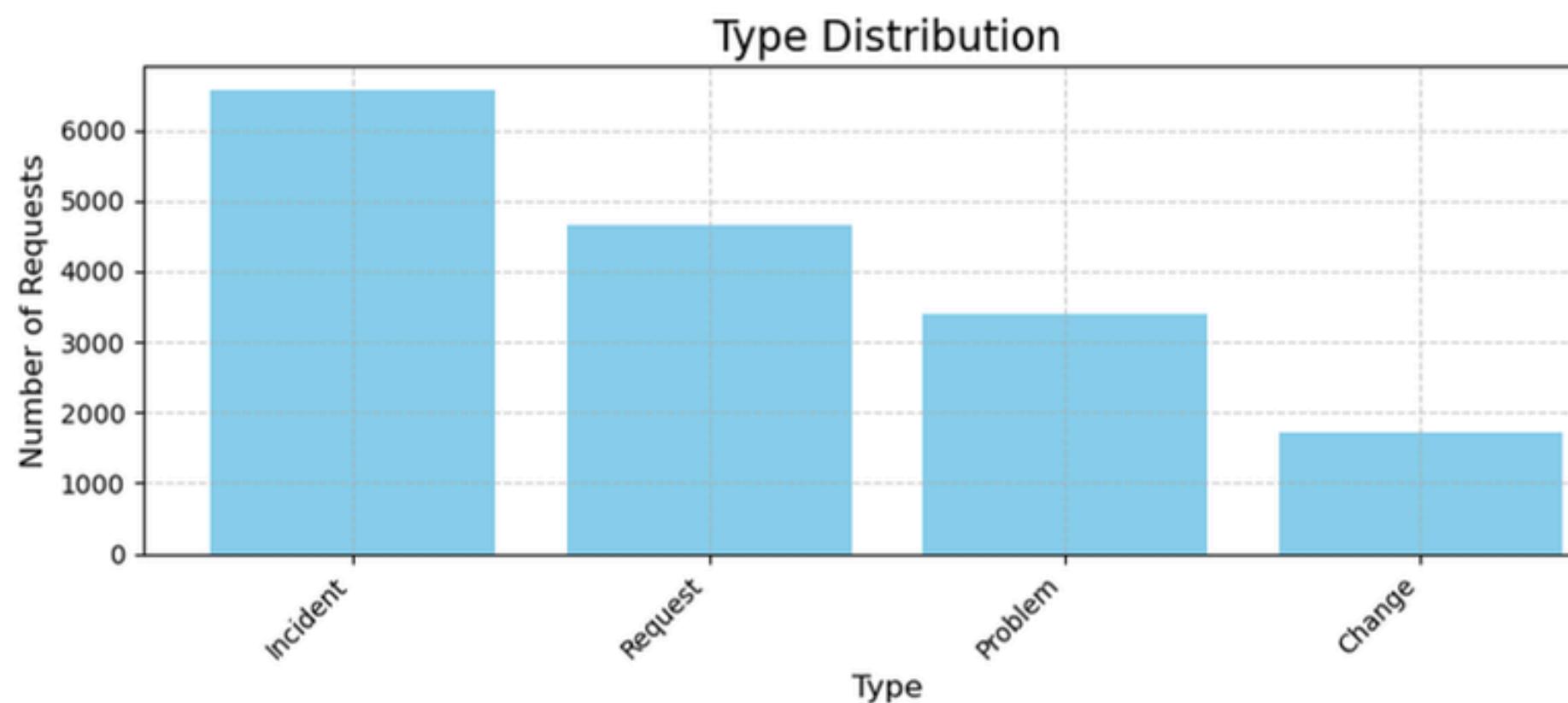
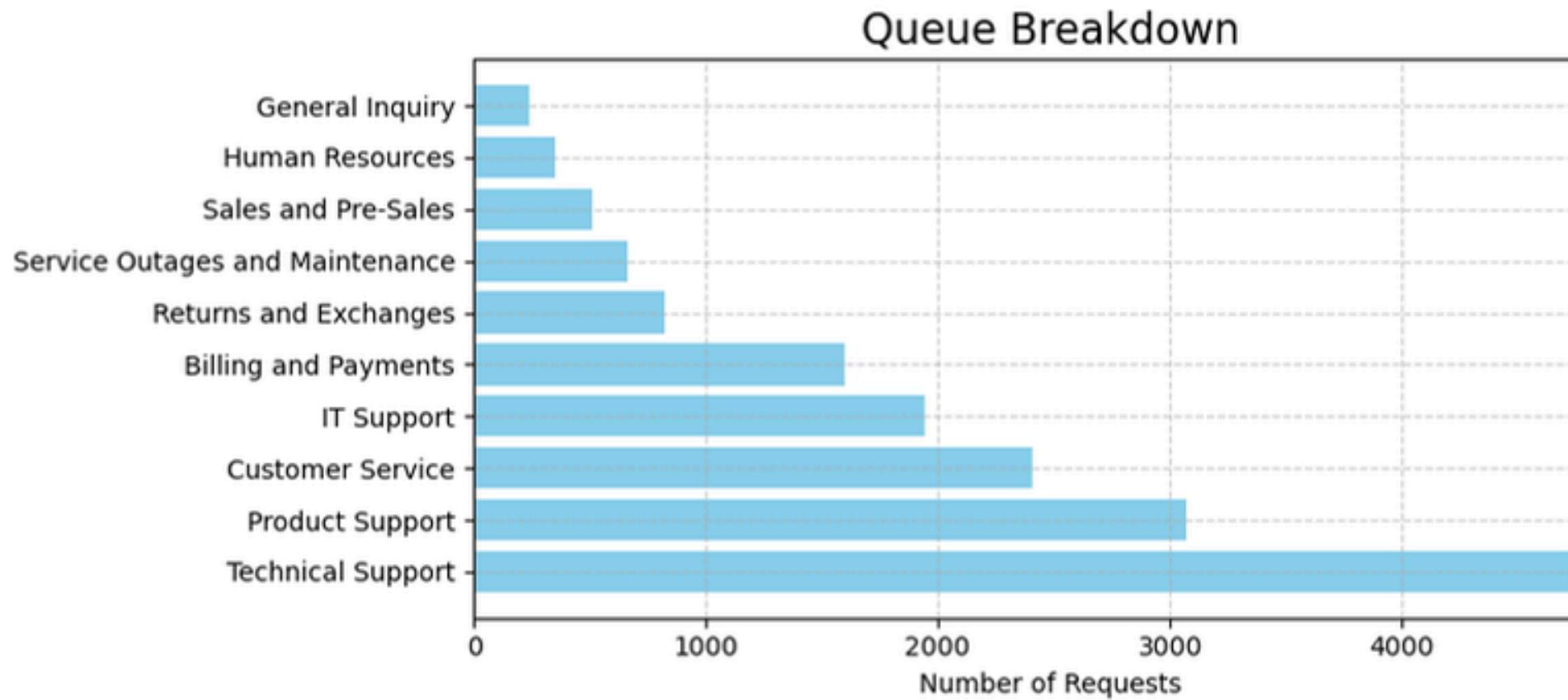
Nilai Non-Null:

- Sebagian besar kolom hampir tanpa nilai kosong, kecuali tag_8 yang hanya terisi 281 nilai. Kolom tag_5 memiliki nilai non-null terbanyak (7.920).

Penurunan Nilai Non-Null:

- Kolom subject memiliki 2.606 nilai non-null, sementara kolom seperti body dan answer penuh. Kolom tag memiliki distribusi nilai kosong yang bervariasi.

Exploratory Data Analysis - Query + Type Distribution



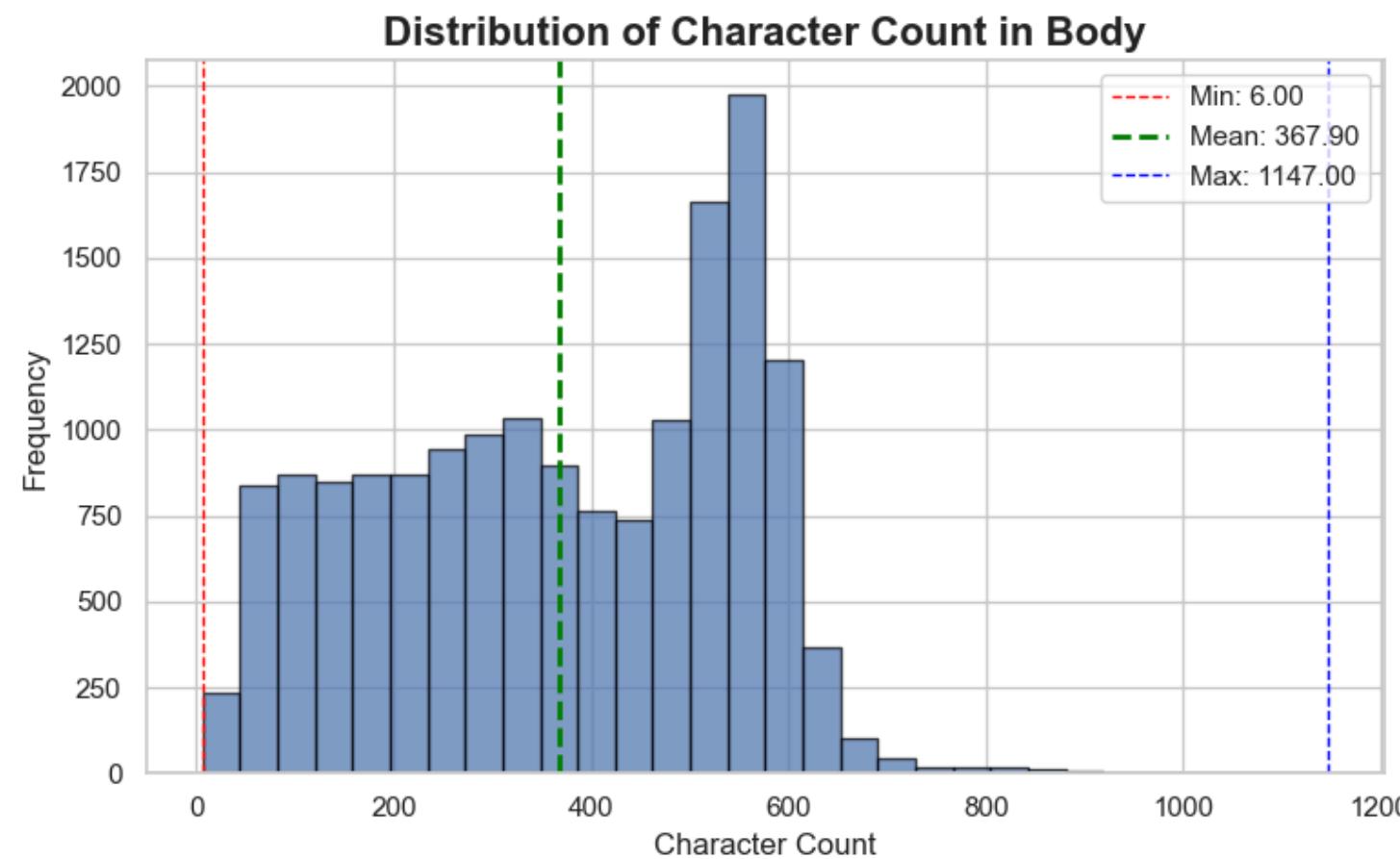
Distribusi Jenis Permintaan

- Insiden adalah jenis permintaan yang paling banyak diterima, lebih dari 6000 permintaan.
- Permintaan menempati posisi kedua, menunjukkan tingkat kebutuhan yang cukup tinggi di luar masalah teknis.
- Masalah dan Perubahan menunjukkan volume permintaan yang lebih rendah, namun tetap signifikan dibandingkan dengan jenis lainnya.

Rincian Antrian Permintaan

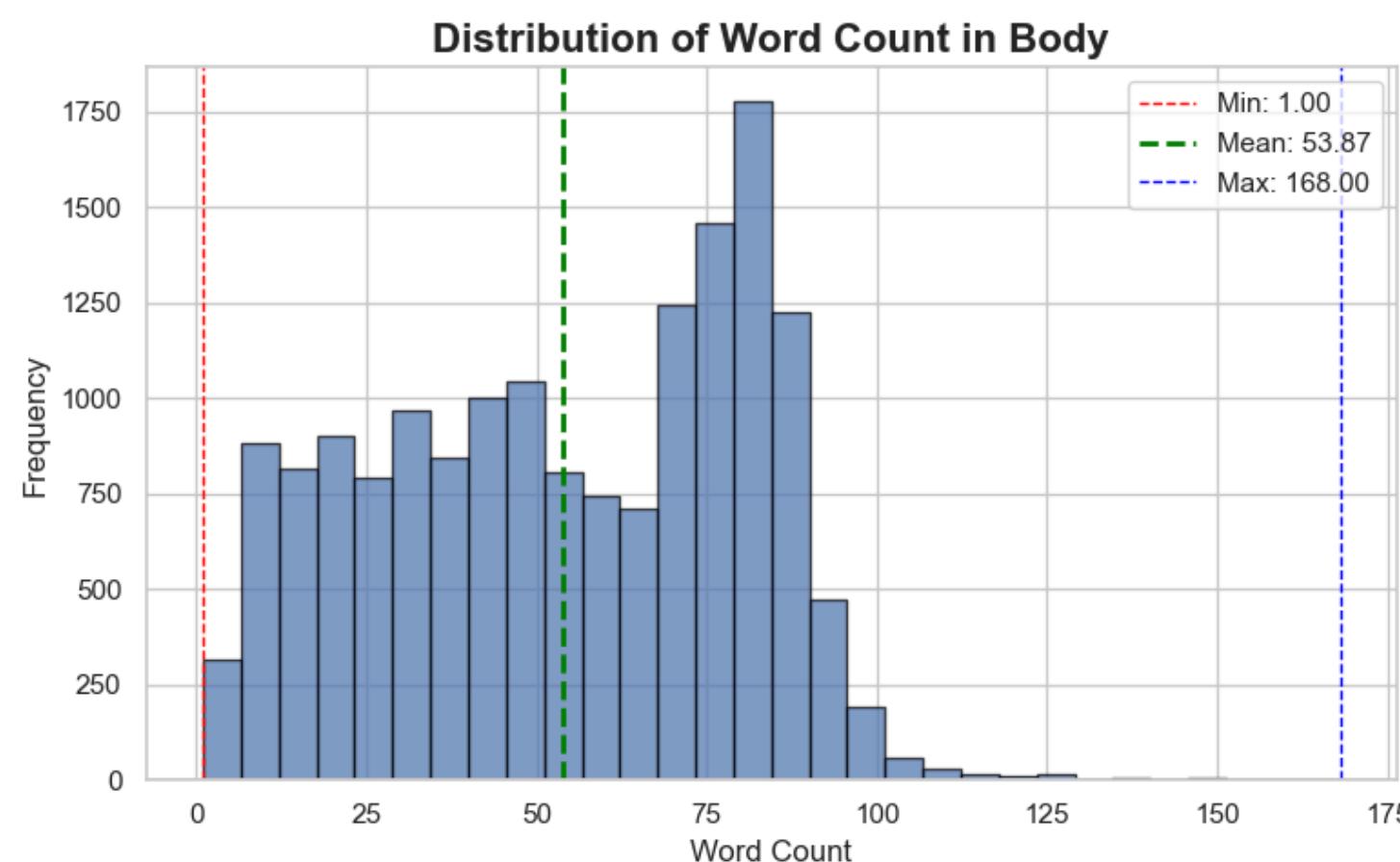
- Dukungan Teknis menjadi kategori permintaan terbesar dengan lebih dari 4000 permintaan, yaitu fokus utama pada pemeliharaan teknis/masalah perangkat.
- Dukungan Produk dan Layanan Pelanggan menunjukkan permintaan yang relatif tinggi, namun jauh di bawah kategori teknis.
- Permintaan untuk Penjualan dan Pra-Penjualan, Sumber Daya Manusia, serta Pertanyaan Umum relatif lebih sedikit, yang mungkin menunjukkan fokus permintaan yang lebih terpusat pada masalah teknis dan dukungan operasional.

Exploratory Data Analysis - Word + Character Freq



Distribusi Jumlah Kata

- Min: 1 kata, menunjukkan bahwa beberapa entri mungkin sangat singkat.
- Rata-rata: 53,87 kata, menunjukkan kebanyakan entri memiliki panjang sedang.
- Maksimum: 168 kata, yang menunjukkan bahwa beberapa entri lebih panjang namun tidak terlalu jauh dari rata-rata.
- Distribusi: Sebagian besar entri memiliki jumlah kata antara 50 hingga 75, dengan puncak terbesar di sekitar 70 kata.



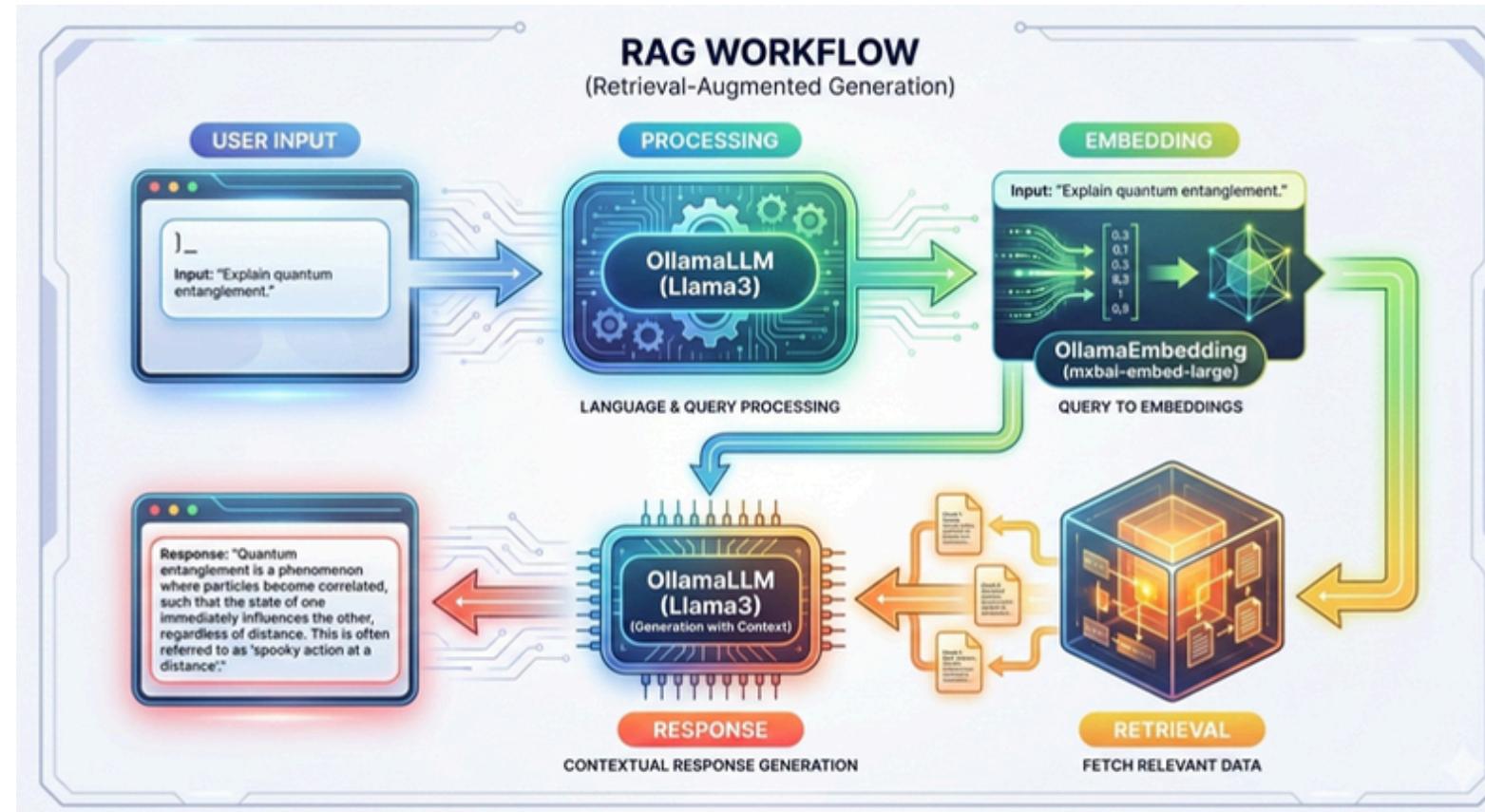
Distribusi Jumlah Karakter

- Min: 6 karakter, mengindikasikan bahwa entri sangat singkat mungkin hanya berisi satu kata.
- Rata-rata: 367,90 karakter, yang menggambarkan panjang teks secara umum.
- Maksimum: 1147 karakter, menunjukkan bahwa beberapa entri sangat panjang.
- Distribusi: Mayoritas entri memiliki jumlah karakter antara 300 hingga 600 karakter, dengan puncak terbesar di sekitar 550 karakter.

RAG



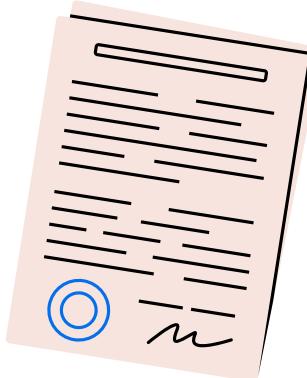
RAG (Retrieval-Augmented Generation)



Component	Chroma	OllamaLLM (Llama3)	OllamaEmbedding (mxbai-embed-)
Description	Vector database for retrieval.	Large language model for text	Embedding model for semantic
Key Feature	Efficient retrieval of embeddings.	Text generation based on context.	Converts text into embeddings.
Context Size	Scalable for large datasets.	Up to 4096 tokens.	Only 512 tokens Context Size
Use in RAG	Powers retrieval in RAG.	Powers generation in RAG.	Provides embeddings for retrieval.
Performance	Fast, scalable retrieval.	High-quality text generation.	Accurate and efficient

- **Input Pengguna**: Proses dimulai dengan pengguna memasukkan kueri.
- **Pemrosesan (OllamaLLM - Llama3)**: Kueri diproses melalui OllamaLLM, di mana kueri dianalisis dan dipersiapkan untuk tahap pengambilan data.
- **Embedding (mxbai-embed-large)**: Kueri yang telah diproses diubah menjadi embedding menggunakan Ollama Embedding, yang mengubah query menjadi representasi vektor untuk pemahaman semantik.
- **Retrieval**: ChromaDB atau sistem retrieval lainnya mencari data relevan berdasarkan embedding kueri, mencari informasi yang mirip dalam basis data.
- **Generasi Respons**: OllamaLLM (Llama3) menghasilkan respons yang relevan dan akurat berdasarkan kueri yang dimasukkan serta data yang diperoleh dari tahap retrieval.

RAG (Retrieval Augmented Generation) - Example



```
prompt = """
You are a professional Customer Support Assistant at Reynaldi Company.
Analyze the provided Document Context to answer the user's inquiry.

REPLACEMENT RULES:
1. If the context or your answer contains ANY company name, replace it with: Reynaldi Company.
2. If the context or your answer contains ANY phone number, replace it with: +86-13028896826.
3. If the context or your answer contains ANY email address, replace it with: gregoriusreynaldi@gmail.com.

RESPONSE RULES:
1. START the response with a professional greeting: "Dear Customer,".
2. Use ONLY the information provided in the "Document Context".
3. If the answer is NOT in the context, state: "I'm sorry, I don't have specific information about that in our records."
4. END the response with:
  "Regards,
  Customer Service Reynaldi Company"

Document Context:
{context}

Question:
{question}

Your Answer:
"""

q = "Good evening, can u give me brief introduction about the HIPAA?? And also can u give advice on data encryption for me??"
```



```
Welcome in Vector Embeddings Document Part
Model Embeddings Process Already Done
Good evening! I'd be happy to help you with that.

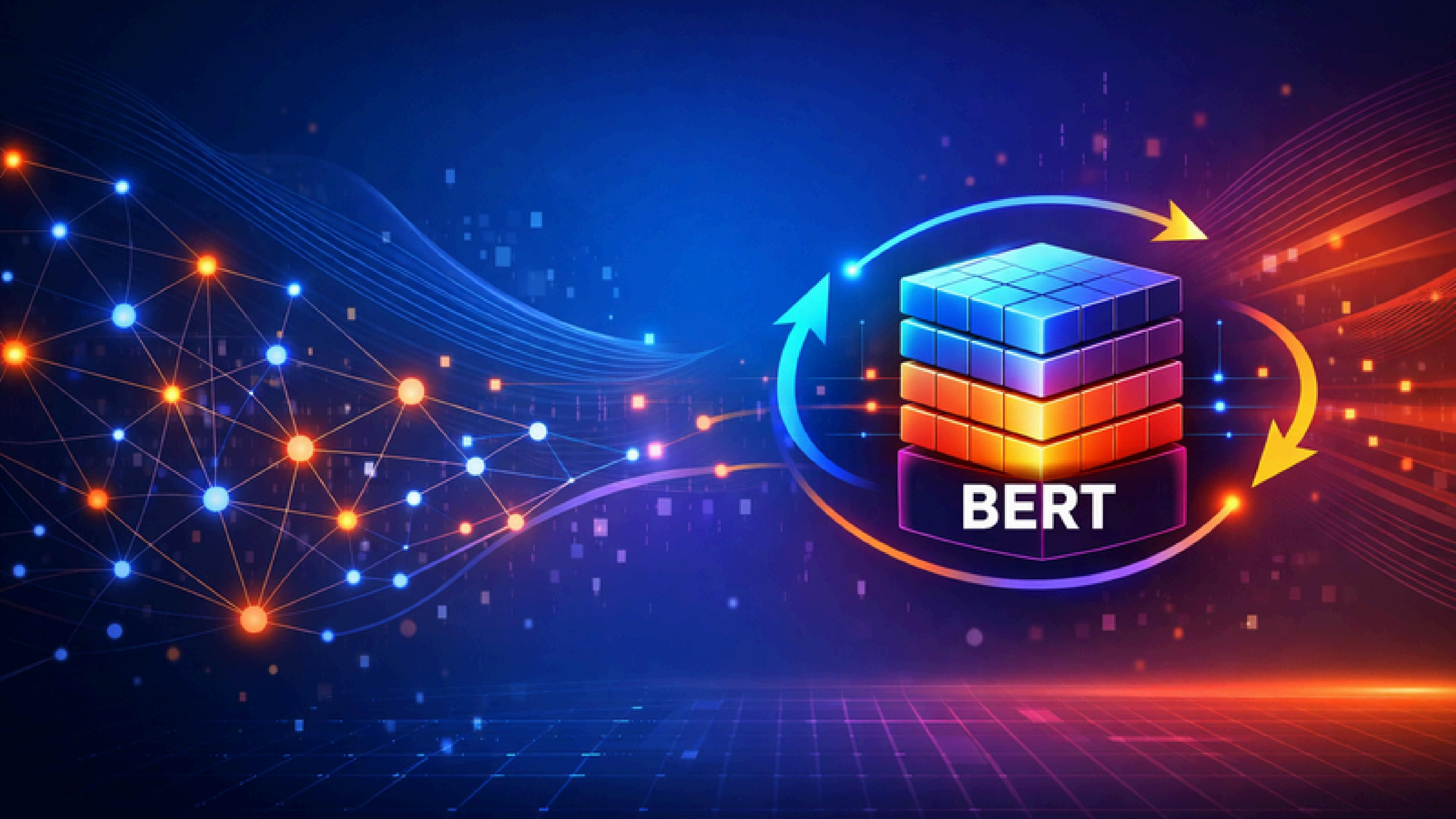
Regarding HIPAA (Health Insurance Portability and Accountability Act), it's a federal law that aims to protect sensitive patient information. Specifically, it sets standards for the security and confidentiality of electronic protected health information (ePHI). Compliance with HIPAA regulations is crucial for healthcare organizations, medical records, and other entities handling personal health data.

Now, regarding data encryption, I can offer some general advice based on the provided documents:

1. **Encryption at rest**: It's recommended to encrypt your data when it's stored or "at rest." This ensures that even if unauthorized access is gained to your storage devices or databases, sensitive information remains protected.
2. **Review compliance with HIPAA requirements**: Make sure you're compliant with HIPAA regulations by reviewing the guidelines and implementing necessary measures.

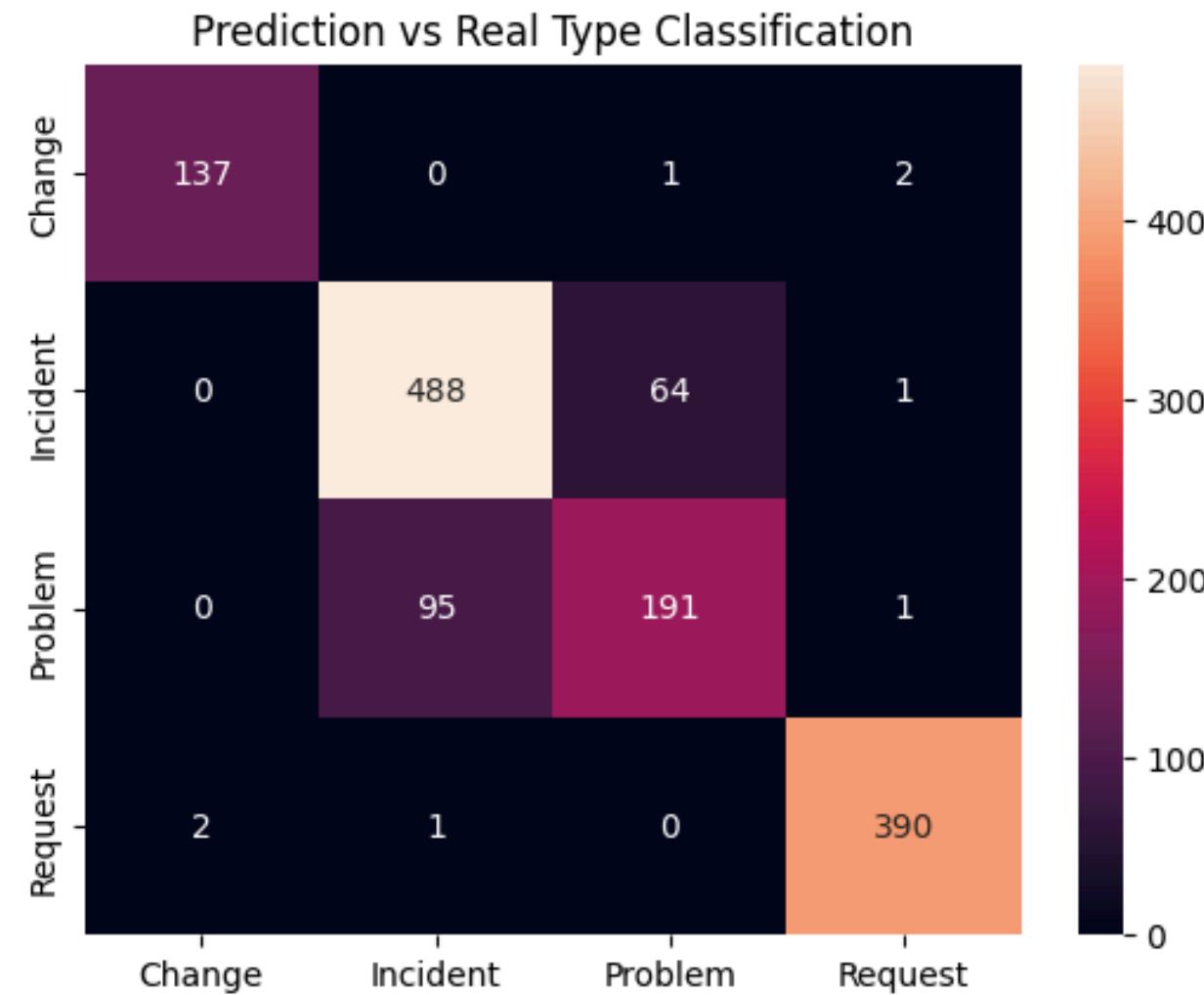
For additional security, considering third-party tools or plugins that enhance data protection can be beneficial.

I hope this helps! If you have more specific questions or need further guidance, feel free to ask.
```



BERT

Type Classification



Step	Training Loss	Validation Loss	Accuracy
500	0.549200	0.387113	0.803350
1000	0.372200	0.354261	0.838310
1500	0.332600	0.332116	0.847779
2000	0.265500	0.334405	0.853605
2500	0.232200	0.318681	0.878369
3000	0.191200	0.345505	0.873999
3500	0.162100	0.358586	0.875455

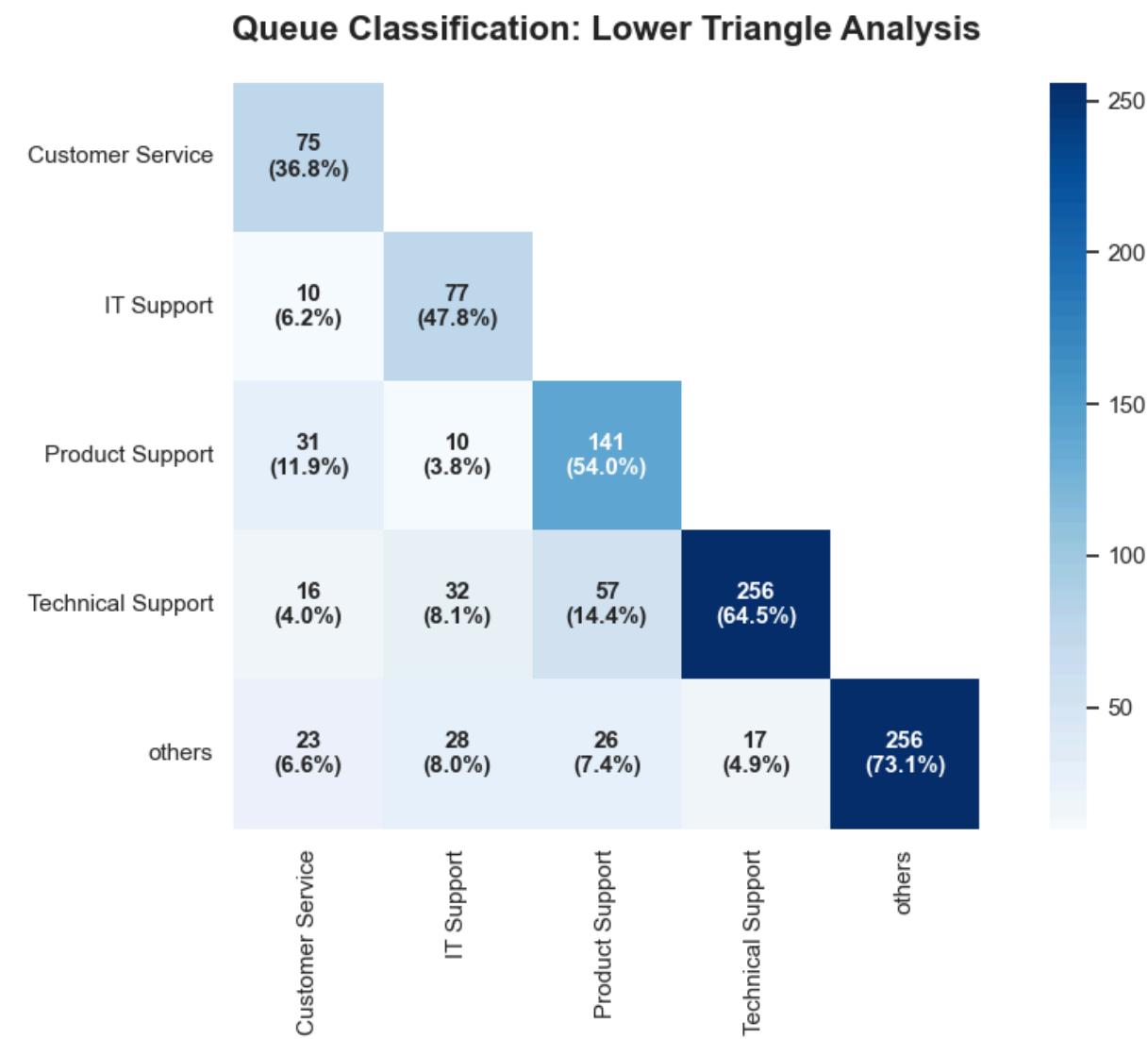
Tabel Training dan Validation:

- Training Loss terus menurun dari 0.5492 pada langkah ke-500 menjadi 0.1621 pada langkah ke-3500, menunjukkan peningkatan yang signifikan dalam kemampuan model untuk meminimalkan kesalahan selama pelatihan.
- Validation Loss juga menurun dengan stabil dari 0.3871 pada langkah ke-500 menjadi 0.3585 pada langkah ke-3500, yang menandakan bahwa model tidak hanya belajar dengan baik pada data pelatihan, tetapi juga menunjukkan perbaikan pada data validasi.
- Akurasi meningkat dari 80.34% pada langkah ke-500 menjadi 87.46% pada langkah ke-3500, menunjukkan kinerja yang semakin baik dalam mengklasifikasikan data.

Confussion Matrix:

- **Change:** Model berhasil memprediksi 137 kasus Change dengan benar, tetapi ada 2 kesalahan yang diprediksi sebagai Request.
- **Incident:** 488 kasus Incident berhasil diprediksi dengan benar, dengan sedikit kesalahan yang diklasifikasikan sebagai Problem
- **Problem:** Model memprediksi 191 kasus Problem dengan benar, tetapi ada 95 kasus yang salah diklasifikasikan sebagai Incident.
- **Request:** 390 kasus Request diprediksi dengan benar, sementara 2 diprediksi sebagai Change dan 1 sebagai Problem.

Queue Classification



Step	Training Loss	Validation Loss	Accuracy
500	1.438500	1.357294	0.426074
1000	1.317000	1.305262	0.440641
1500	1.229100	1.238262	0.479971
2000	1.051800	1.203814	0.537509
2500	0.908700	1.232961	0.546977
3000	0.775300	1.180232	0.582666
3500	0.695400	1.189860	0.586307

Tabel Training dan Validation:

Training Loss menunjukkan penurunan yang konsisten dari 1.438500 pada langkah ke-500 menjadi 0.695400 pada langkah ke-3500, menunjukkan bahwa model semakin belajar dengan baik. Validation Loss sedikit fluktuatif, namun secara umum tetap stabil dengan nilai mendekati 1.19 pada akhir pelatihan. Akurasi meningkat secara bertahap dari 42.61% pada langkah ke-500 menjadi 58.63% pada langkah ke-3500.

Evaluasi Akhir:

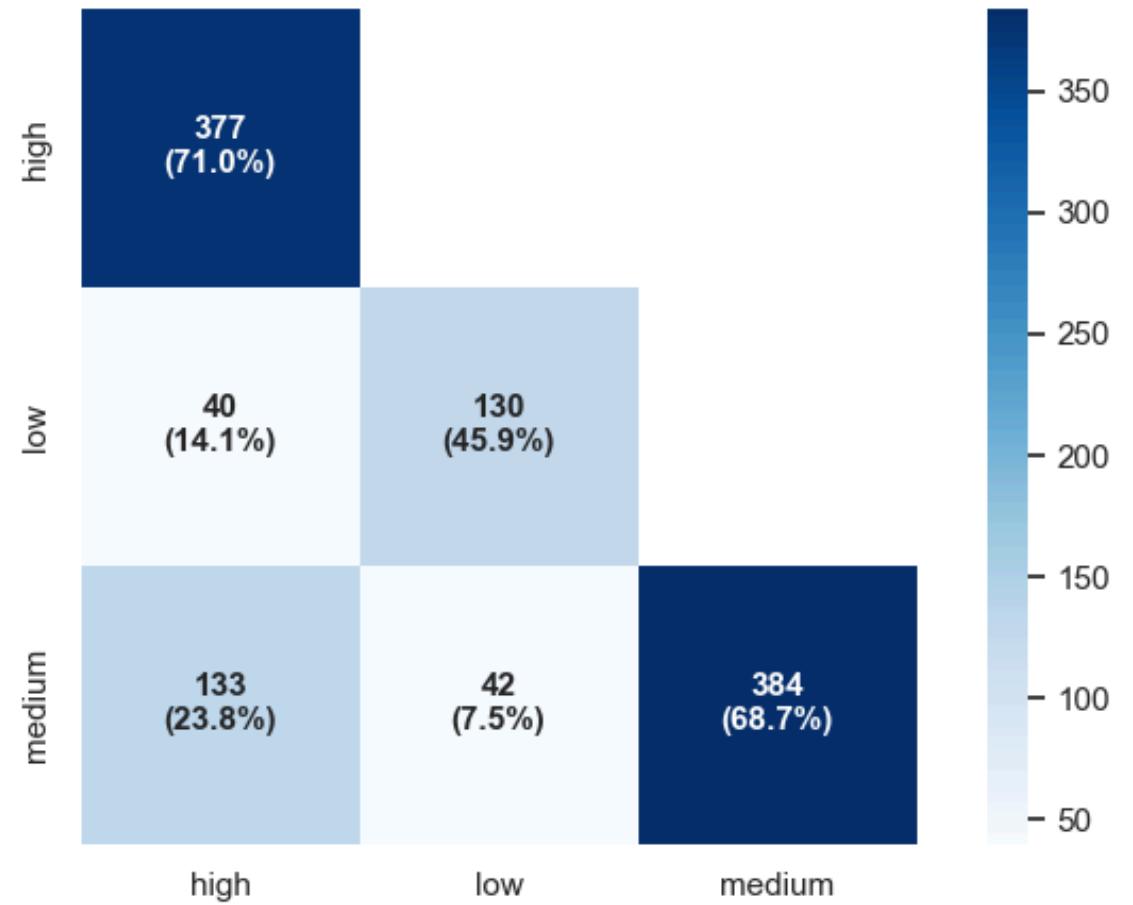
Loss evaluasi (eval_loss) berada di 1.19, menunjukkan kinerja model pada data yang tidak terlihat. Akurasi evaluasi (eval_accuracy) mencapai 58.63%, yang menunjukkan tingkat prediksi yang lebih baik pada data validasi. Waktu pelatihan total (eval_runtime) adalah 8.012 detik. Samples per second mencapai 171.369, menunjukkan kecepatan pemrosesan data yang baik.

Kesimpulan:

Model menunjukkan peningkatan akurasi yang konsisten selama pelatihan, meskipun loss validasi tetap cukup tinggi. Ini menunjukkan bahwa meskipun model belajar dengan baik, masih ada ruang untuk perbaikan dalam akurasi prediksi pada data yang tidak terlihat.

Level Classification

Level Classification: Lower Triangle Analysis



Tabel Training dan Validation:

- Training Loss menunjukkan penurunan yang konsisten dari 1.0562 pada langkah ke-500 menjadi 0.4352 pada langkah ke-3500, yang menunjukkan bahwa model semakin belajar dengan baik dalam meminimalkan kesalahan.
- Validation Loss menunjukkan fluktuasi, dimulai dari 1.0248 pada langkah ke-500 dan turun ke 0.9275 pada langkah ke-3500. Ini menunjukkan bahwa model cenderung mengalami penurunan kesalahan di data validasi meskipun ada beberapa variasi.
- Akurasi meningkat secara bertahap dari 46.47% pada langkah ke-500 menjadi 64.89% pada langkah ke-3500, menunjukkan bahwa model semakin baik dalam klasifikasi data.

Evaluasi Akhir:

- Loss evaluasi (eval_loss) mencapai 0.9275, menunjukkan kinerja model pada data yang tidak terlihat cukup baik, meskipun ada ruang untuk perbaikan.
- Akurasi evaluasi (eval_accuracy) mencapai 64.89%, yang menunjukkan peningkatan akurasi yang signifikan pada data validasi.
- Waktu pelatihan total (eval_runtime) adalah 8.012 detik, menunjukkan bahwa model dapat diproses dalam waktu yang efisien.
- Samples per second mencapai 171.369, menunjukkan kecepatan pemrosesan data yang sangat baik.

Step	Training Loss	Validation Loss	Accuracy
500	1.056200	1.024832	0.464676
1000	0.990100	0.987159	0.490168
1500	0.923100	0.962638	0.506919
2000	0.756000	0.902998	0.588492
2500	0.635900	0.936809	0.608157
3000	0.527800	0.894870	0.643117
3500	0.435200	0.927522	0.648944



ASR

Speech Recognition Part...



Whispers

Komponen	Whisper
Arsitektur	Berbasis Transformer, dilatih untuk multibahasa dan multimodalitas (suara dan teks). Memiliki Encoder dan Decoder
Model Ukuran	Tersedia dalam berbagai ukuran (Tiny, Base, Small, Medium, Large) untuk fleksibilitas akurasi dan kecepatan.
Pre-training	Dilatih pada dataset multibahasa besar, memungkinkan pengenalan suara dari berbagai bahasa tanpa pelabelan bahasa khusus.
Deteksi Bahasa	Mendeteksi bahasa dari input audio secara otomatis tanpa perlu pengaturan manual.
Penggunaan	Ideal untuk aplikasi yang memerlukan pengenalan suara multibahasa atau pengenalan aksen yang beragam.
Kecepatan & Akurasi	Model kecil lebih cepat, sementara model besar memberikan akurasi lebih tinggi, namun lebih memerlukan sumber daya.
Output	Menghasilkan transkrip teks dari input audio dan juga dapat digunakan untuk penerjemahan otomatis.

Model yang Digunakan (openai/whisper-large-v3):

- Jenis Model: Model Automatic Speech Recognition (ASR) berbasis Transformer.
- Ukuran Model: Whisper-large-v3 adalah varian terbesar dengan performa terbaik di antara model Whisper, menawarkan akurasi tinggi untuk pengenalan ucapan.
- Sampling Rate: 16,000 Hz, audio yang masuk harus diproses pada frekuensi sampel ini untuk mendapatkan hasil terbaik.

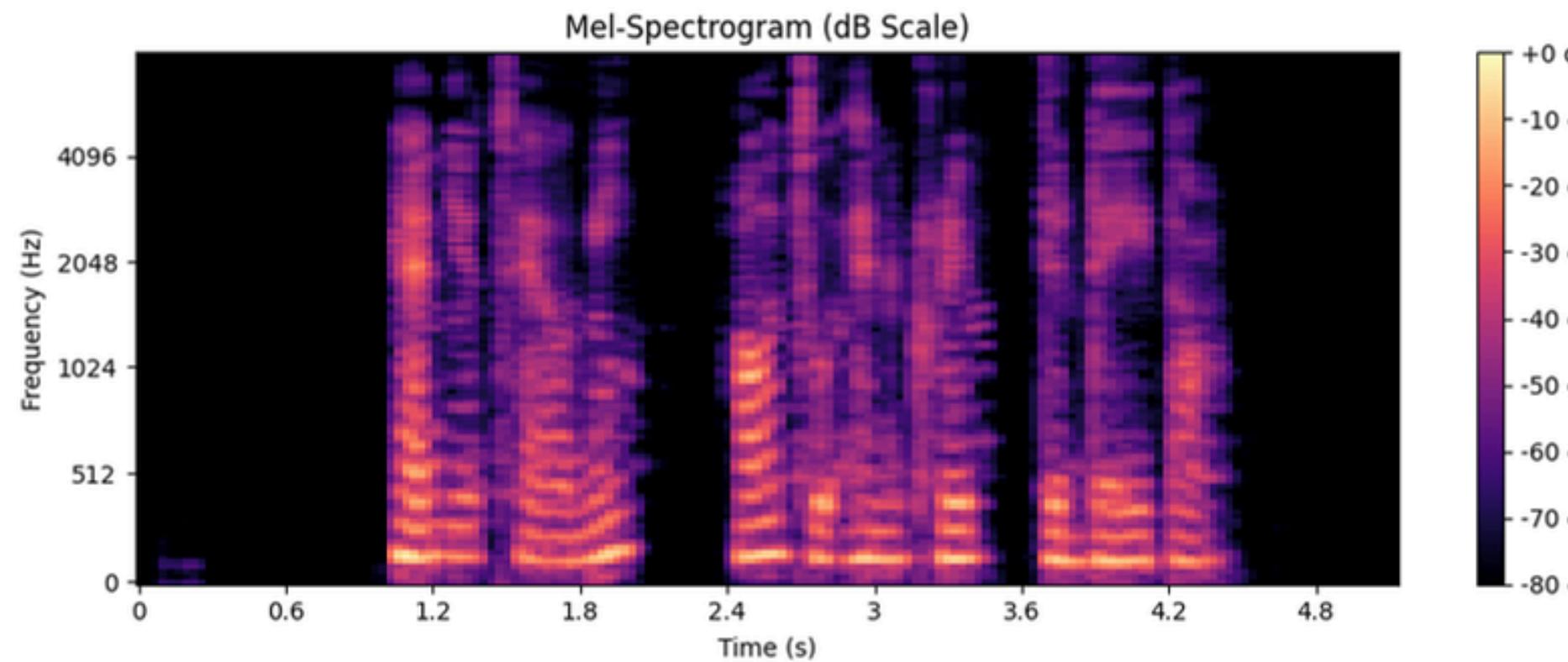
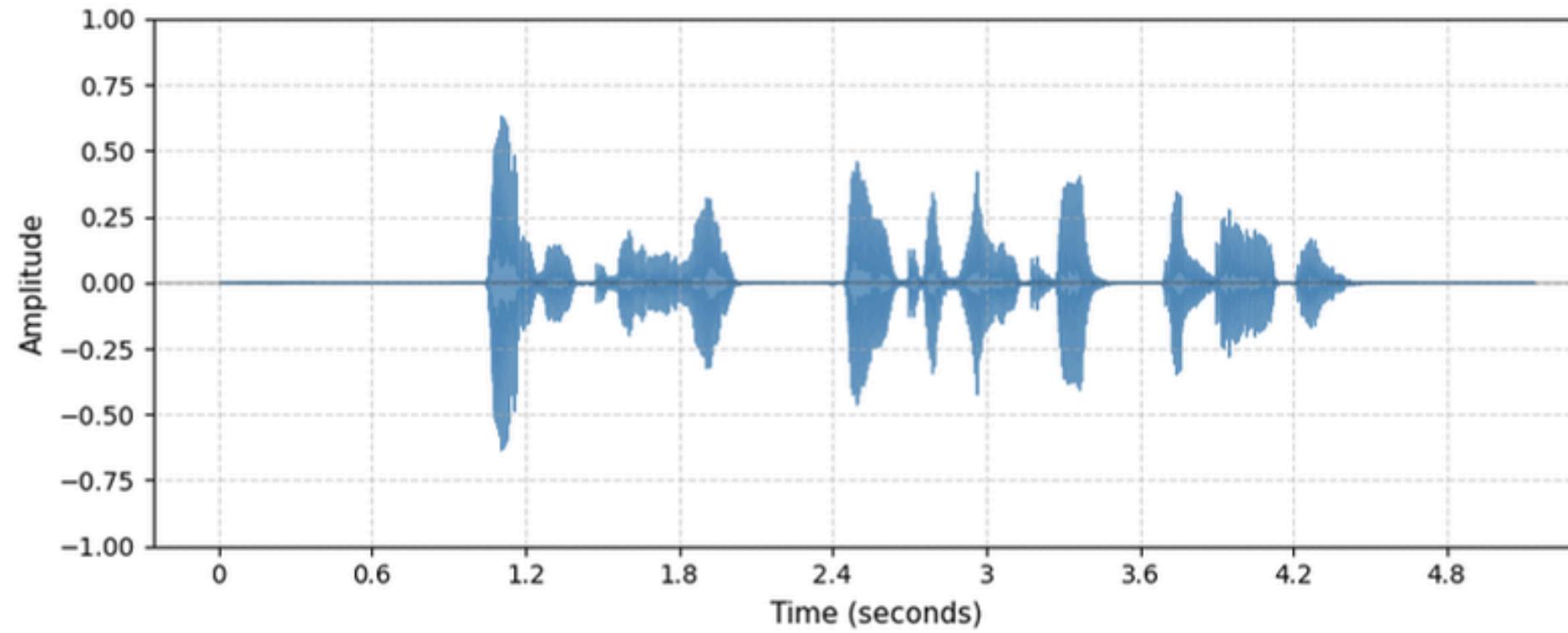
Arsitektur:

- Encoder-Decoder : Whisper menggunakan arsitektur Transformer dengan encoder-decoder yang digabungkan dalam satu model.
- Multibahasa: Mampu mengenali berbagai bahasa dan aksen tanpa pelatihan ulang untuk bahasa tertentu.

Whispers

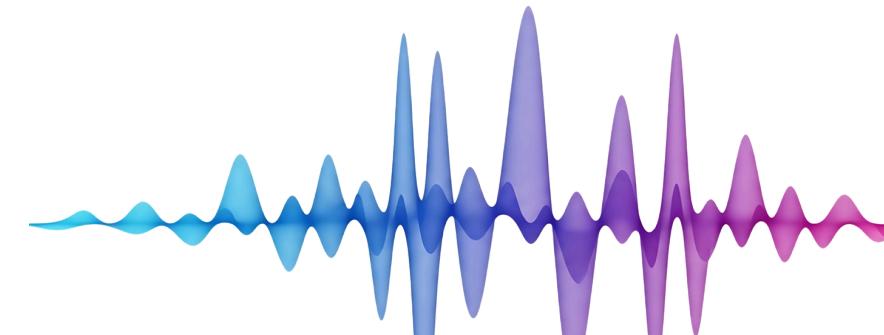


Time-Domain Waveform: Audio Signal Analysis



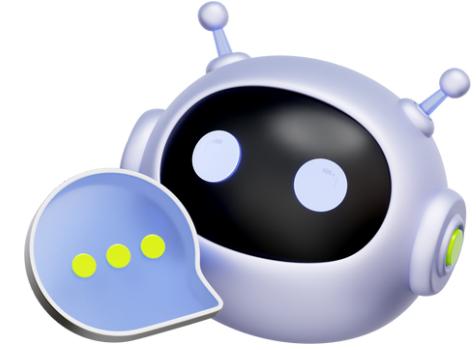
'Can you tell me how to encrypt the data?'

- Time-Domain Waveform menggambarkan perubahan amplitudo sinyal audio terhadap waktu.
- Mel-Spectrogram menunjukkan analisis frekuensi suara dalam skala dB.
- Transkripsi Teks: "Can you tell me how to encrypt the data?" merupakan hasil dari pemrosesan audio menjadi teks.





VISUALIZATION OF THE APP



Advanced RAG Assistant

Ask questions or record voice to get intelligent responses

Chat Interface

can you tell me what is SaaS and is there any relation with your company product with it?
21/01/45

Dear Customer, Thank you for your inquiry regarding SaaS. Our software, offered by Reynaldi Company, provides benefits such as streamlined workflows and enhanced security. We will provide documentation and setup instructions to support your understanding of our product. Our SaaS is designed to seamlessly integrate various products, enhancing overall compatibility and functionality. For detailed information, we provide comprehensive documentation and guides for setting up these integrations on our website. You can also contact us at +86-13028896826 for assistance in discussing your specific needs. Regards, Customer Service
Reynaldi Company
21/01/46

Type your question here... Send

Voice Recording

Click to start recording
00:00

Transcription
Transcribed text will appear here... Use in Chat

Classification Results

Level
medium

Queue
Customer Service

Type
Requests

Advanced RAG Assistant

Ask questions or record voice to get intelligent responses

Chat Interface

question based on the available information.
21/01/42

I am facing overheating issues in the server room due to a malfunctioning PC fan. Despite multiple restart cycles and cleaning efforts, can you assist with resolving this issue to prevent potential hardware damage?
21/01/53

Based on the document context, it seems that the problem remains unresolved after trying several restart cycles and basic cleaning to remove dust accumulation. In this case, a professional inspection of the fan may be necessary, possibly including replacement. Please confirm the make and model of the affected system, and let us know your preferred time for a technician to visit.
21/01/55

Type your question here... Send

Voice Recording

Recording completed
00:00

Transcription
I am facing overheating issues in the server room due to a malfunctioning PC Use in Chat

Classification Results

Level
high

Queue
IT Support

Type
Incident

Thank You