Sahabat-AI

# Pedoman Penggunaan Bertanggung Jawab Sahabat-AI

Sahabat-AI Responsible Use Guideline

indosat
OOREDOO HUTCHISON

# Daftar Isi

## *Table of Contents*

**Pendahuluan dan Tujuan** 1
*Introduction and Purpose*

**Lingkup Aplikasi** 6
*Scope of Application*

**Pedoman Penggunaan Bertanggung Jawab** 8
*Responsible Use Guidelines*

**Penggunaan yang Dilarang** 11
*Prohibited Uses*

**Strategi Mitigasi dan Manajemen Risiko** 12
*Mitigation Strategies and Risk Management*

**Informasi Kontak dan Bantuan** 13
*Contact Information and Support*

**Version**
April 2025

# I. Pendahuluan dan Tujuan

### A. Ringkasan terkait Sahabat-AI dan Kemampuannya

**Sahabat-AI** adalah sekumpulan Model Bahasa Besar (LLM) yang telah dilatih secara kontinu *(pre-trained)* dan disesuaikan *(instruct-tuned)* untuk Bahasa Indonesia serta berbagai bahasa daerah. Ekosistem Sahabat-AI diinisiasi dan dikembangkan bersama oleh perusahaan teknologi dan telekomunikasi Indonesia: PT Indosat Tbk ("Indosat Ooredoo Hutchison" or "IOH").

Dalam mengembangkan model tersebut, kami berkolaborasi dengan universitas lokal dan grup media di Indonesia untuk melatih model kita dengan talenta mereka dan/atau menggunakan set data mereka. Universitas ternama tersebut termasuk Universitas Gadjah Mada, Universitas Indonesia, Universitas IPB, dan Institut Teknologi Bandung, serta sejumlah grup media ternama di Indonesia. Kami berencana untuk memperluas kolaborasi kami dengan lebih banyak universitas dan grup media lainnya untuk dapat semakin melatih dan memperkaya model dalam Bahasa Indonesia, agar memastikan ketersediaan konteks lokal dan relevansi budaya. Kami juga didukung oleh pusat riset dan ahli teknologi global seperti **AI** Singapore untuk melatih modelnya dan mendapatkan pemahaman bahasa secara umum.

Ekosistem Sahabat-AI diakselerasi oleh **NVIDIA.**

Kami membangun LLM dalam Bahasa Indonesia dan bahasa-bahasa daerah karena kami percaya bahwa Kecerdasan Artifisial harus mencerminkan kecerdasan dan budaya bangsa kita sendiri serta memungkinkan inklusivitas dan representasi Indonesia pada lanskap global LLM yang telah berkembang di dua wilayah utama: Amerika Serikat dan Tiongkok. Oleh karena itu, model LLM yang berkembang dari negara-negara tersebut mencerminkan pandangan berdasarkan data dan bahasa yang digunakan untuk melatihnya. Hal ini mencerminkan visi dan prinsip **AI** dari Indosat Ooredoo Hutchison**,** yang mengutamakan sistem **AI** yang etis, transparan, dan aman demi kepentingan seluruh pengguna.

Sebagai negara dengan jumlah penduduk terbesar keempat di dunia, budaya Indonesia sangatlah beragam, dengan lebih dari 700 bahasa daerah.

Dengan mayoritas orang Indonesia (77%) lebih memilih untuk menggunakan bahasa daerah mereka (Sumber: Laporan e-Conomy SEA 2024), masih terdapat banyak masyarakat yang belum terlayani oleh **LLM** yang sudah ada.

Melalui Sahabat-AI. kami berharap untuk dapat mempercepat pengembangan dan adopsi penggunaan kecerdasan artifisial di Indonesia, di mana bahasa yang digunakan dalam perdagangan, bisnis, dan masyarakat umumnya tidak terwakilkan dengan baik oleh model global yang tersedia saat ini.

# I. Introduction and Purpose

### A. Overview of Sahabat-AI and its capabilities

**Sahabat-AI** (Indonesian language for "close friends") is a collection of Large Language Models (LLMs) which has been continuously pre-trained and instruct-tuned for the Indonesian language and its various local languages. The Sahabat-AI ecosystem is co-initiated, and co-developed by Indonesian homegrown tech and telecommunication companies: PT Indosat Tbk ("Indosat Ooredoo Hutchison" or "IOH").

In developing the model, we worked with local universities and media groups in Indonesia to train our model with their talents and/or using their datasets. These include top universities such as Universitas Gadjah Mada, Universitas Indonesia, IPB University, Institut Teknologi Bandung, and several key Indonesian media groups. We plan to expand our collaborations with more universities and media groups to further train and enrich the model in Bahasa Indonesia, ensuring the best provision of local context and cultural relevance. We are also supported by research centers and global tech experts such as AI Singapore to train the model to gain general language understanding.

Sahabat-AI ecosystem, accelerated by NVIDIA.

We build LLMs in Indonesian and its local languages as we believe artificial intelligence must reflect our nation's intelligence and culture while enabling inclusivity and representation of Indonesia within the global LLM landscape, which has evolved around two dominant areas: the USA and China. Consequently, these models reflect those worldviews based on data sets and the languages that train them. This reflects the vision and AI principles of Indosat Ooredoo Hutchison, which prioritize ethical, transparent, and secure **AI** systems for the benefit of all users.

Indonesia, the world's fourth most populous country in the world, has a rich and diverse culture with more than 700 local languages.

With the majority of Indonesians (77%) preferring to use local languages (Source: e-Conomy SEA Report 2024). many remain underserved by existing LLMs.

Through Sahabat-AI, we aim to catalyze the development and adoption of AI use cases in Indonesia where the medium of commerce, enterprise and society are primarily the languages underrepresented by existing global models.

Kami juga percaya bahwa semakin banyak orang Indonesia menggunakan dan berkontribusi terhadap Sahabat-AI, semakin baik kemampuan Sahabat-AI dalam memahami konteks lokal dan relevansi budaya. Sehingga, kami berkomitmen untuk mengembangkan Sahabat-AI menjadi ekosistem sumber terbuka agar dapat mempercepat pelatihan model dan memungkinkan konteks yang lebih lokal.

We also believe the more people use and contribute to Sahabat-AI, the better Sahabat-AI is able to comprehend local context and cultural relevance. Therefore, we are committed to developing Sahabat-AI as an open source ecosystem in order to accelerate the model's training and enable more localized context.

Melalui ekosistem sumber terbuka, kami dapat mendemokratisasi akses kecerdasan artifisial bagi semua orang: peneliti, akademisi, perusahaan kecil hingga besar, serta *developer* sehari-hari.

With an open-source ecosystem, we will be able to democratize AI access for everyone: researchers. academics, small to large enterprises, and everyday developers.

**B. Tujuan Pedoman**

**B. Objectives of The Guideline**

Dokumen ini berfungsi sebagai kerangka untuk pengembangan dan pemanfaatan Sahabat-AI secara bertanggungjawab. Tujuan utama dari pedoman ini adalah memberikan arahan praktis bagi *developer* dan pengguna Sahabat-AI, yang secara khusus dioptimalkan untuk bahasa Indonesia dan beragam bahasa daerahnya .

This document serves as a framework for the responsible development and utilization of Sahabat-AI. The primary purpose of this guideline is to provide practical direction for developers and users engaging with Sahabat-AI's language models, which are specifically optimized for Indonesian language and its rich variety of regional languages.

Sahabat-AI berfokus pada pelestarian warisan budaya dan linguistik Indonesia melalui pengembangan **AI** yang bertanggung jawab yang menghormati nuansa dari bahasa daerah sekaligus memastikan aksesibilitas dan kesesuaian budaya bagi semua komunitas. Pedoman Penggunaan Bertanggung Jawab ini menetapkan standar untuk proses desain, pengembangan, dan penerapan model bahasa khusus Indonesia, sekaligus menerapkan langkah-langkah keselamatan yang menyeluruh dan pertimbangan etika yang disesuaikan dengan konteks sosial dan budaya Indonesia. Kami mengacu pada kerangka kebijakan yang berlaku di wilayah Indonesia seperti Surat Edaran Menteri Komunikasi dan Informatika No. 9 Tahun 2023 tentang Etika Kecerdasan Artifisial dan Pedoman ASEAN tentang Tata Kelola dan Etika AI.

Sahabat-AI focuses on preserving Indonesia's cultural and linguistic heritage through responsible AI development that respects the nuances of its languages and dialects while ensuring accessibility and cultural appropriateness for all communities. The Responsible Use Guideline is setting standards for the design, development, and deployment of Indonesia-specific language models, alongside implementing robust safety measures and ethical considerations tailored to the country's sociocultural context. We use references from the existing regulatory frameworks such as the Minister of Communication and Informatics' Circular Letter No. 9 Year 2023 on Ethical Guidelines for Use of Artificial Intelligence and ASEAN Guide on AI Governance and Ethics.

Kami sangat menyarankan *developer* dan pengguna untuk menggunakan pedoman ini sebagai referensi dan tolak ukur bagi mereka yang ingin memanfaatkan Sahabat-AI secara bertanggung jawab. Pengembang diharapkan untuk memastikan aspek kepercayaan dan keamanan dari produk ataupun inovasi mereka yang didukung oleh model Sahabat-AI. Pedoman ini akan terus dikembangkan, dengan tinjauan dan pembaharuan secara berkala se1nng masukan dan perkembangan dari *developer,* pemangku kepentingan, kemajuan teknologi, dan ketentuan regulasi terkait.

We strongly advise developers and users to use this reference as a benchmark for those seeking to utilise Sahabat-AI responsibly. Developers are expected to ensure the trust and safety of their product or innovation that is powered by the Sahabat-AI model. This guideline is intended to be constantly evolving, with regular reviews and updates as we incorporate feedback from developers, stakeholders, technological advancements, and regulatory requirements in the AI landscape.

Melalui pedoman ini, Sahabat-AI bertujuan untuk menjadi mitra terpercaya dalam lanskap kecerdasan artifisial di Indonesia, mendorong inovasi sembari memastikan pengembangan kecerdasan artifisial yang bertanggung jawab dan etis serta memberikan manfaat bagi seluruh masyarakat Indonesia.

Through this guideline, Sahabat-AI aims to become a trusted partner in Indonesia's AI landscape, fostering innovation while ensuring responsible and ethical AI development that benefits all Indonesians.

## C. Prinsip-Prinsip Utama

Panduan Sahabat-AI mengintegrasikan prinsip-prinsip AI yang diadopsi oleh Indosat Ooredoo Hutchison, termasuk keadilan, transparansi, privasi, keamanan, pengawasan manusia, serta kepatuhan terhadap hukum yang berlaku.

Penting bagi *developer* untuk memperhatikan kerangka regulasi dan pedoman yang ada ketika mengembangkan dan menerapkan kecerdasan artifisial. Sebagai LLM yang dikembangkan di Indonesia, Sahabat-AI mempertimbangkan prinsip-prinsip dan rekomendasi yang ditetapkan oleh lembaga berwenang, seperti:

- Surat Edaran Kementerian Komunikasi dan Informatika tentang Etika Kecerdasan Artifisial
- Pedoman **ASEAN** tentang Tata Kelola dan Etika **AI**

### 1. Akuntabilitas dan Integritas

Akuntabilitas manusia sangat pentinq untuk memastikan bahwa sistem **AI** berfungsi dengan baik, mematuhi hukum yang berlaku, dan mengikuti prinsip-prinsip etika. *Developers* dan pengguna harus menjaga integritas sepanjang siklus hidup sistem AI, mulai dari desain hingga penerapan, dan bertindak secara bertanggung jawab jika terjadi malfungsi atau penyalahgunaan. Selain itu, pengoperasian **AI** harus mengutamakan kemampuan untuk membuat keputusan berdasarkan informasi yang dapat dipercaya dan dapat dipertanggungjawabkan. Sehingga memastikan bahwa wawasan atau inovasi yang dihasilkan oleh **AI** dapat diandalkan dan dibagikan kepada publik secara bertanggungjawab.

### 2. Transparansi

Transparansi dalam **AI** mencakup pengungkapan informasi mengenai kapan sistem **AI** digunakan dan tujuan di balik pengambilan keputusan tersebut. Transparansi ini sangat penting untuk membangun kepercayaan publik, karena pengguna perlu memahami proses bagaimana data mereka digunakan dan faktor-faktor yang mempengaruhi keputusan **AI.** Pengungkapan yangjelas sangatlah penting, terutama dengan semakin banyaknya **AI** yang hadir di berbagai industri, dan pengguna perlu menyadari potensi risiko, seperti bias atau malfungsi. Dengan memberitahu pengguna tentang peranan AI dalam proses, individu dapat membuat keputusan berdasarkan informasi yang mumpuni untuk dapat mempertimbangkan apakah mereka akan berinteraksi dengan sistem tersebut.

Penjelasan yang memadai dapat melengkapi transparansi dengan memastikan bahwa keputusan AI disampaikan dengan cara yang mudah dipahami oleh pengguna. Hal ini bisa berupa deskripsi teks sederhana hingga bantuan visual seperti heatmap, yang membantu orang memahami bagaimana AI yang dibangun mencapai kesimpulannya.

### 3. Keamanan dan Keselamatan

Sistem **AI** harus dilindungi dari serangan berbahaya seperti perusakan data dan pembalikan model. Hal tersebut memerlukan penerapan langkah-langkah keamanan siber yang kuat, termasuk enkripsi, otentikasi, pembaruan perangkat lunak secara rutin, serta pengembangan rencana tanggap insiden. Deployer sistem **AI** juga harus melakukan pengujian keamanan seperti penilaian kerentanan dan uji penetrasi, untuk melindungi sistem **AI** dari potensi ancaman digital.

## C. Core Principles

The Sahabat-AI Guideline incorporates the AI principles adopted by Indosat Ooredoo Hutchison, including fairness, transparency, privacy, security, human oversight, and compliance with applicable laws.

It is important that developers take into account existing regulatory frameworks and guidelines when developing and deploying AI. As an Indonesian developed LLM, Sahabat-AI takes into consideration the principles and recommendations set by governing institutions, such as:

- Minister of Communication and Informatics' Circular Letter on **AI** Ethics
- **ASEAN** Guide on **AI** Governance and Ethics

### 1. Accountability and Integrity

Human accountability is crucial to ensure AI systems function properly, comply with relevant laws, and adhere to ethical principles. Developers and users should maintain integrity throughout the entire lifecycle of AI systems, from design to deployment, and act responsibly if any malfunction or misuse occurs. Furthermore, the operation of AI should prioritize the ability to make decisions based on trustworthy and accountable information. Therefore ensuring that anyinsights or innovations generated by AI can be relied upon and responsibly shared with the public.

### 2. Transparency

Transparency in AI involves disclosing information related to when AI systems are being used and the purpose behind their decision-making. This transparency is crucial for building public trust, as users need to understand how their data is being utilized and the factors influencing AI decisions. Clear disclosure is important, especially as AI's presence grows in various industries, and users need to be aware of potential risks, such as bias or malfunction. By informing users about AI's role in processes, individuals can make informed choices about whether to engage with the system.

Explainability complements transparency by ensuring that AI decisions are communicated in a way that is understandable to users. This can range from simple text descriptions to visual aids like heatmaps, helping people grasp how AI reaches its conclusions.

### 3. Security and Safety

AI systems must be secured against malicious attacks like data poisoning and model inversion. This requires deploying robust cybersecurity measures, including encryption, authentication, and regular software updates, as well as developing incident response plans. Deployers should also perform security testing, such as vulnerability assessments and penetration testing, to safeguard **AI** systems against potential digital threats.

Selain itu, keselamatan dalam hal ini mencakup penilaian risiko untuk mengidentifikasi dan mengurangi potensi bahaya, serta memungkinkan intervensi manusia atau penghentian sistem jika **AI** membuat keputusan yang tidak aman, seperti pada kendaraan otonom.

Additionally, safety involves conducting risk assessments to identify and mitigate potential dangers, allowing for human intervention or system disengagement if the **AI** makes unsafe decisions, such as in autonomous vehicles.

**4. Kemanusiaan**

Sistem AI harus mengutamakan nilai-nilai yang berfokus pada manusia dan mempertimbangkan manfaat keseluruhan bagi masyarakat, sembari melindungi individu dari bahaya. Sejak proses desain hingga penerapan, sistem **AI** harus dibuat dengan fokus pada manfaat yang diberikan kepada manusia, termasuk memastikan bahwa sistem ini tidak mengeksploitasi populasi rentan atau membuat keputusan yang merugikan. Penting untuk memahami interaksi dan persepsi pengguna untuk menghindari hasil yang negatif. **AI** tersebut dapat dicapai dengan menguji sistem pada berbagai kelompok pengguna dan menerima umpan balik.

**4. Human-centricity**

AI systems should prioritize human-centred values and consider the overall benefit of society while protecting individuals from harm. From design to deployment, AI systems must be created with a focus on human benefit, including to ensure they do not exploit vulnerable populations or make harmful decisions. It is essential to understand user interactions and perceptions to avoid negative outcomes, which can be achieved by testing systems with diverse user groups and incorporating feedback.

Selain itu, AI tidak boleh digunakan untuk tujuan jahat ataupun praktik manipulatif, seperti penerapan pola gelap yang dapat menipu pengguna untuk membuat pilihan yang merugikan. Deployer juga harus mempertimbangkan dampak sosial yang lebih luas dari AI, terutama terkait dengan hak kekayaan intelektual, dengan melakukan penilaian dampak dan mencari cara untuk mengintegrasikan **AI** dengan memberikan atribusi yang akurat serta menghindari pelanggaran informasi yang bersifat kepemilikan.

Additionally, AI should not be used for malicious purposes or manipulative practices, such as deployment of dark patterns that deceive users into making harmful choices. Deployers must also consider AI's broader societal impact, especially regarding intellectual property rights, by conducting impact assessments and exploring ways to integrate AI by providing the right attributions and avoiding any violation of proprietary information.

Terlebih, sangatlah penting untuk dapat memastikan kepatuhan terhadap prinsip-prinsip hak asasi manusia, hubungan sosial, kepercayaan yang dianut, serta pendapat atau pemikiran setiap orang.

Furthermore, it is also crucial to ensure adherence to the principles of human rights, rights to societal relationships, rights of belief, as well as rights to the freedom of expression or thoughts.

**5. Privasi dan Pengelolaan Data**

Sistem **AI** harus memastikan mekanisme perlindungan privasi dan data yang kuat sepanjang siklus hidupnya. Hal tersebut mencakup pengaturan protokol yang jelas tentang praktik pemrosesan data, memastikan kepatuhan terhadap peraturan perundang-undangan relevan serta pedoman etika, seperti Undang-Undang No. 27 Tahun 2022 tentang Perlindungan Data Pribadi dan Surat Edaran Menteri Komunikasi dan Informatika No. 9 Tahun 2023 tentang Etika Kecerdasan Artifisial.

**5. Privacy and Data Governance**

AI systems must ensure robust privacy and data protection mechanisms throughout their lifecycle. This involves setting clear protocols on data processing practices, ensuring compliance with relevant laws and ethical guidelines, such as Indonesia's Law No. 27 Year 2022 on Personal Data Protection and the Minister of Communication and Informatics' Circular Letter No. 9 Year 2023 on Artificial Intelligence Ethics.

Ketika sistem AI mengumpulkan data pribadi, organisasi harus memiliki transparansi penuh mengenai praktik data mereka. Organisasi perlu dengan jelas memberitahukan jenis-jenis data yang dikumpulkan, bagaimana data tersebut digunakan, diproses, dibagikan, ditransfer, disimpan dan dihapus, dan juga siapa saja yang memiliki akses terhadap data pada Pemberitahuan Privasi. Selain itu, organisasi perlu memberitahukan individu terkait dengan hak-hak mereka, termasuk hak mengakses, memperbaiki ataupun mengajukan keberatan terhadap pemrosesan data. Organisasijuga harus menilai risiko yang terkait dengan pemrosesan data dan menerapkan langkah-langkah untuk memastikan akurasi, integritas dan keadilan sembari mencegah keputusan **AI** yang bias atau diskriminatif

When AI systems collect personal data, organizations must be fully transparent about their data practices. They should clearly outline in their Privacy Notice the types of data collected, how it is used, processed, shared, transferred, stored, and deleted, as well as who has access to it. Additionally, organizations must inform individuals of their rights, such as rights to access, correct or object to data processing. Organizations should also assess risks associated with data processing and implement measures to ensure data accuracy, integrity and fairness while preventing biased or discriminatory AI decisions.

Sistem **AI** harus memastikan pengumpulan data pribadi hanya mencakup data yang benar-benar diperlukan sesuai dengan dasar-dasar dan tujuan yang sah dan hanya menyimpan data tersebut selama dibutuhkan untuk mencapai tujuan yang dimaksud.

AI systems should collect only the necessary personal data with a legitimate basis and purpose and retain it only as long as needed to fulfil its intended purpose.

*Developer* juga harus memprioritaskan tata kelola data yang kuat, termasuk manajemen siklus hidup dan audit, untuk memastikan keadilan, keamanan, dan kepatuhan.

**6. Ketangguhan**

Sistem **AI** harus cukup tangguh untuk dapat menangani kesalahan, input yang tidak terduga, dan kondisi dunia nyata yang dinamis, serta menjaga kinerja yang konsisten. Mereka harus memiliki ketangguhan terhadap perubahan data yang tidak terduga dan tidak menunjukkan perilaku yang berbahaya, serta memastikan bahwa sistem tetap berfungsi sebagaimana mestinya.

Mengingat bahwa sistem **AI** tidak sempurna, *deployer* harus menerapkan kontrol akses yang sesuai dan perlindungan untuk sistem sensitif guna mengurangi risiko dari performa yang tidak dapat diandalkan. Pengujian yang ketat sebelum penerapan sangat penting untuk memastikan ketahanan dan konsistensi di berbagai situasi. Selain itu, pendokumentasian sumber data, pencatatan langkah-langkah pemrosesan data, dan penjagaan data *lineage* dapat mendukung pemecahan masalah dan memastikan sistem **AI** beroperasi dengan andal.

Developers should also prioritize strong data governance, including lifecycle management and audits, to ensure fairness, security, and compliance.

**6. Robustness**

AI systems must be robust enough to handle errors, unexpected inputs, and dynamic real-world conditions while maintaining consistent performance. They should be resilient to unforeseen data changes and not exhibit dangerous behavior, ensuring they continue to function as intended.

Given that AI systems are not infallible, deployers must implement proper access control and protection for sensitive systems to mitigate risks from unreliable performance. Rigorous testing before deployment is essential to ensure robustness and consistency across diverse situations. Additionally, documenting data sources, tracking data processing steps, and maintaining data lineage can aid in troubleshooting and ensuring AI systems operate reliably.

# II. Lingkup Aplikasi

**A. Definisi**

anduan ini berfungsi sebagai sumber informasi yang utamanya ditujukan untuk *developer.* Panduan 1n1 memberikan saran praktis tentang cara membangun dan meluncurkan produk yang didukung oleh Sahabat-AI dengan cara yang bertanggung jawab. Panduan ini memberikan wawasan mengenai pertimbangan penting yang perlu diperhatikan *developer* sepanjang proses pengembangan, mulai dari desain awal hingga tahap peluncuran.

Selain itu, panduan ini juga menguraikan berbagai strategi mitigasi untuk mengatasi potensi risiko, dengan memfokuskan pada teknik penyelarasan di tingkat model dan sistem. Meskipun ditujukan untuk *developer,* prinsip-prinsip yang dijelaskan di sini juga relevan untuk pemangku kepentingan lainnya yang terlibat dalam siklus pengembangan AI. Pihak-pihak ini dapat memperoleh manfaat dari informasi mengenai cara mengintegrasikan praktik-praktik yang bertanggung jawab dan pendekatan mitigasi risiko, sehingga dapat memastikan produk akhir sesuai dengan standar etika, peraturan perundang-undangan, dan keselamatan.

**B. Pengecualian atau Batasan**

Meskipun pedoman 1n1 memberikan rekomendasi penggunaan Sahabat-AI yang bertanggung jawab, terdapat pengecualian dan keterbatasan dalam penerapannya. Pedoman ini tidak mencakup semua kemungkinan skenario atau kompleksitas teknis yang mungkin timbul dalam pengembangan dan penerapan **LLM.** Oleh karena itu, *developer* perlu melakukan penilaian dan penanganan risiko-risiko unik dan spesifik terhadap kasus penggunaan masing-masing.

Pedoman inijuga tidak dimaksudkan untuk menggantikan ataupun mengesampingkan persyaratan peraturan perundang-undangan, kebijakan, atau ketentuan yang mungkin berlaku di sektor-sektor tertentu, seperti kesehatan, keuangan, atau pendidikan. Selain itu, dokumen ini tidak memberikan spesifikasi teknis yang mendalam untuk implementasi LLM, melainkan menggarisbawahi praktik terbaik dan prinsip-prinsip tingkat tinggi yang perlu dipertimbangkan.

# 11. Scope of Application

**A. Definition**

This guide serves as a resource primarily aimed at developers, offering practical advice on how to responsibly build and deploy Sahabat-AI-powered products. It provides insights into key considerations that developers should take into account throughout the development process, from the initial design to deployment.

The guide also outlines various mitigation strategies to address potential risks, focusing on alignment techniques at both the model and system levels. While primarily targeted at developers, the principles outlined here are also relevant to other stakeholders involved in the AI development lifecycle. These groups can all benefit from understanding how to integrate responsible practices and risk-mitigation approaches, ensuring the final product aligns with ethical, legal, and safety standards.

**B. Exclusions or Limitations**

While this guideline provides recommendations for the responsible use of Sahabat-AI, there are certain exclusions and limitations to their application. This guideline does not cover all possible scenarios or technical complexities that may arise in the development and deployment of LLMs, and as such, developers should assess and address unique risks specific to their usecase.

The guideline is also not intended to replace or supersede any legal, regulatory, or industry-specific requirements that may apply to specific sectors, such as healthcare, finance, or education. Additionally, this document does not provide exhaustive technical specifications for the implementation of LLMs but rather outlines best practices and high-level principles to be considered.

Sahabat-AI diberikan sebagaimana adanya. Indosat Ooredoo Hutchison tidak mewakilkan atau menjamin bahwa keandalan, ketepatan waktu, kualitas, kesesuaian, ketersediaan, keakuratan, kelengkapan, ataupun keamanan Sahabat-AI akan memenuhi kebutuhan dan harapan Pengguna. Pengguna bertanggung jawab sepenuhnya untuk memutuskan apakah Sahabat-AI layak digunakan, diperbanyak, dimodifikasi, dijalankan, ditampilkan, atau didistribusikan ("Penggunaan"). Pengguna juga bertanggung jawab sepenuhnya atas semua risiko yang timbul akibat penggunaan atau pendistribusian Sahabat-AI dan penggunaan hak dan izin yang diberikan dalam pedoman ini. Indosat Ooredoo Hutchison tidak bertanggung jawab atas segala kerugian atau kerusakan, ataupun biaya dalam bentuk apa pun, baik secara langsung, tidak langsung, insidental, khusus, maupun konsekuensial, yang timbul dari atau sehubungan dengan Penggunaan atau ketergantungan Pengguna terhadap Sahabat-AI.

Terakhir, pedoman 1n1 utamanya ditujukan untuk *developer.* Namun, kami menyambut baik masukan dari pihak lain, seperti pelajar, periset, dan badan pemerintah / publik, untuk dapat memberikan input berdasarkan kerangka kerja atau kebijakan tambahan untuk memastikan penggunaan yang bertanggung jawab terhadap ekosistem Sahabat-AI. *Developer* juga harus mempertimbangkan bahwa sektor **AI** terus berkembang, dan terdapat kemungkinan pedoman ini perlu kembali ditinjau dan diperbarui secara berkala untuk mencerminkan kemajuan industri dan teknologi, risiko-risiko baru, dan harapan masyarakat yang berkembang.

Sahabat-AI is provided "as is". Indosat Ooredoo Hutchison do not represent or warrant that the reliability, timeliness, quality, suitability, availability, accuracy, completeness, or security of Sahabat-AI will fulfil users' needs and expectations. Users are solely responsible for deciding if it is suitable to use, reproduce, modify, perform, display, or distribute Sahabat-AI ("Usage"). Users are also solely responsible for all risks that come with using or distributing Sahabat-AI and using the rights and permissions given in this guideline. Indosat Ooredoo Hutchison shall not be liable for any losses or damages, or expense of any kind, whether direct, indirect, incidental, special, or consequential, arising out of or in connection with the User's Usage or reliance on Sahabat-**AI.**

Lastly, the guideline is primarily aimed at developers. However, we welcome other stakeholders, such as students, researchers, and government / public agencies, to provide input based on different frameworks or additional policies to ensure responsible use of the Sahabat-AI ecosystem. Developers should also be mindful that the field of AI is evolving, and this guideline may need to be periodically reviewed and updated to reflect new industry and technology advancements, emerging risks, and evolving societal expectations.

# 111. Pedoman Penggunaan Bertanggung Jawab

## A. *Application Design*

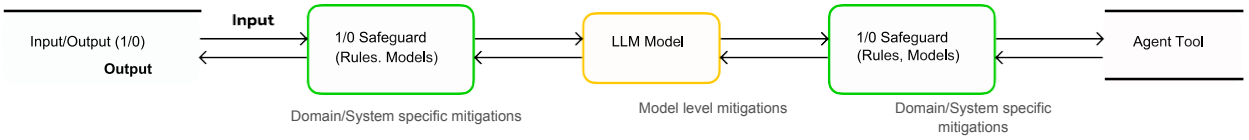
Layaknya database SOL mana pun, dimana tanggung jawab *developer* aplikasi adalah untuk memberikan perlindungan terhadap SOL *injection,* penggunaan model Sahabat memerlukan langkah-langkah pencegahan serupa. Tanggung jawab *developer* aplikasi LLM adalah untuk melindungi terhadap *prompt injection* dan potensi penyalahgunaan lainnya. Integrasi yang dirancang dengan baik harus mencakup beberapa lapisan perlindungan, memastikan penggunaan yang bertanggung jawab melampaui upaya mitigasi pada level model yang diterapkan selama pelatihan. Integrasi model Sahabat-AI yang khusus harus mencakup langkah-langkah perlindungan berikut pada level aplikasi dan model, dengan langkah mitigasi sesuai dan diterapkan selama desain aplikasi.

# 111. Responsible Use Guideline

## A. Application Design

Like in any SOL database, where it is the responsibility of the application developer to protect against SOL injection, using the Sahabat model requires similar precautions. It is the LLM application developer's responsibility to safeguard against prompt injection and other potential misuse. **A** well-designed integration should include multiple layers of protection, ensuring responsible use beyond just model-level mitigations applied during training. A typical Sahabat-AI model integration should include the following safeguards at both the application and model levels. with appropriate mitigation steps incorporated during application design.

**Responsible Sahabat Model Deployment**




- **Rekayasa Prompt:** Untuk memastikan model berperilaku sesuai dengan yang diinginkan, sangatlah penting untuk merancang *prompt* secara hati-hati. Hal ini mencegah model melakukan penyimpangan dari tujuan yang ditetapkan. Contohnya termasuk memanfaatkan pemanggilan fungsi, menggunakan input dengan template tertentu, dan minimalisasi input pengguna secara bebas.

- **Perlindungan Input untuk Model:** Berbagai aturan atau model dapat diterapkan pada *input prompt* pengguna untuk memastikan kepatuhan terhadap budaya, bahasa, dan hukum setempat. Contohnya termasuk menggunakan *rule engine* untuk menyaring pertanyaan yang mengandung kata-kata kasar, menggunakan model LLM yang lebih kecil untuk mendeteksi niat pengguna dan memblokir permintaan berbahaya, serta membatasi bahasa input yang didukung.

- **Prompt Engineering:** To ensure the model behaves as intended, it is vital to design prompts carefully. This prevents the model from deviating from its goal. Examples include leveraging function calling, using templated inputs, and minimizing free-flow user input.

- **Input Safeguards for the Model:** Various rules or models can be applied to user input prompts to ensure cultural, linguistic, and legal compliance. Examples include using a rule engine to filter out queries containing curse words, employing a smaller LLM model to detect user intent and block malicious requests, and restricting supported input dialects.

- **Mitigasi di Level Model:** Selama pre-training atau fine-tuning, model harus diselaraskan. Hal ini akan dibahas pada bagian-bagian selanjutnya.

- **Perlindungan *Tooling Input* untuk Model:** Jika kasus penggunaan melibatkan tool call, sangat penting untuk melindungi apa yang diteruskan ke agent tool. Hal ini mencegah LLM untuk memanggil alat yang berada di luar cakupan aplikasi tersebut (misalnya, pemanggilan API DELETE oleh pengguna sedangkan hanya pemanggilan API GET yang diperbolehkan).

- **Perlindungan *Tooling Output* untuk Model:** Untuk mencegah respons pemanggilan alat yang dapat menyebabkan LLM menghasilkan konten yang dilarang, sangatlah penting untuk menyaring respons yang tidak diinginkan ataupun berbahaya. Model atau aturan serupa yang digunakan untuk perlindungan input dapat diterapkan untuk tujuan ini. Contohnya termasuk membatasi bahasa output yang didukung dan menghindari kebocoran data yang tidak diinginkan.

- **Perlindungan *Output* Model:** Karena ini adalah perlindungan terakhir sebelum merespons pengguna, model atau aturan perlindungan yang paling ketat harus digunakan untuk memastikan model merespons secara bertanggungjawab.

**Pemantauan:**
Selain menetapkan langkah-langkah perlindungan di berbagai lapisan, pemantauan yang berkelanjutan sangat penting:

- **Pemantauan Kinerja Perlindungan Secara *Real-time*** Setiap lapisan perlindungan yang ditambahkan dapat berpotensi menurunkan pengalaman pengguna. Oleh karena itu, sangat penting untuk terus memantau perhitungan jumlah filter masuk/keluar dari setiap perlindungan, sekaligus dengan waktu tanggapannya.

- **Pemantauan Filter Perlindungan:** Selain pemantauan kinerja sistem yang objektif untuk berbagai perlindungan, penting juga untuk mengambil sampel dan memantau sampel positif dan negatif dari perlindungan input dan output. Haltersebut membantu memahami keterbatasan sistem saat ini dan dapat memperbaiki model atau aturan yang berlaku saat ini.

- **Menuju Mitigasi di Level Model:** Pelabelan sampel dari perlindungan dapat dimanfaatkan di masa depan untuk melatih model itu sendiri ataupun untuk meningkatkan kemampuan model perlindungan.

- **Model-Level Mitigations:** During pre-training or fine-tuning, the model should be aligned. This is discussed in later sections.

- **Tooling Input Safeguards for the Model:** If the use case involves tool calling, it is important to protect what is passed to the agent tool. This prevents the LLM from calling tools that are outside the scope of the application (e.g., calling a DELETE user API when only GET API calls are permitted).

- **Tooling Output Safeguards for the Model:** To prevent tool call responses from causing the LLM to generate harmful content, it is crucial to filter out unwanted or malicious responses. Similar models or rules as those used for input safeguards can be employed for this purpose. Examples include restricting supported output dialects and avoiding unwanted data leakage.

- **Model Output Safeguards:** As this is the final safeguard before responding to the user, the strictest safeguard model or rule should be used to ensure the model responds responsibly.

**Monitoring:**
Apart from establishing these safeguards at different layers, ongoing monitoring is essential:

- **Real-time Safeguard Performance Monitoring:** Each added safeguard layer can potentially degrade the user experience. It is important to continuously monitor each safeguard's filtering in/out count, along with response time.

- **Safeguard Filter Monitoring:** In addition to objective system performance monitoring of different safeguards, it is also important to sample and monitor positive and negative samples of input and output safeguards. This helps to understand current system limitations and improve existing models or rules.

- **Moving Towards Model-Level Mitigation:** Labeling of samples from safeguards can be leveraged in the future to train the model itself or further improve safeguard models.

## B. *Safety Alignment*

Sahabat-AI telah melalui berbagai langkah penyelarasan keamanan selama pengembangannya. Namun, seperti LLM lainnya, tetap terdapat risiko seperti halusinasi yang perlu dikelola. Oleh karena itu, pertimbangan tambahan harus diperhatikan dengan mengikuti desain aplikasi yang telah disebutkan untuk mengelola potensi risiko keamanan dalam pemanfaatan Sahabat-AI.

Siapa pun yang mengunduh dan menggunakan model **AI** harus memastikan bahwa *developer* menguji model tersebut secara menyeluruh, menyesuaikannya dengan kebutuhan dan tantangan spesifik mereka, dan memiliki tim khusus untuk memastikan keselarasan kepercayaan dan keamanan.

## C . *Model Evaluation*

Evaluasi yang ketat sangat disarankan untuk memastikan bahwa model menghasilkan konten yang andal dan aman, serta sesuai dengan kebijakan sistem. Proses ini bersifat iteratif hingga model dapat mencapai kinerja yang memadai dalam kasus penggunaan yang dimaksud dan memenuhi standar keamanan. Para *developer* dianjurkan untuk mengambil langkah-langkah yang diperlukan guna memberikan hasil evaluasi yang akurat. Hal ini dapat mencakup melakukan evaluasi pada set pengujian ketahanan yang mencerminkan data pada lingkungan produksi, mencegah kebocoran data pelatihan terhadap set pengujian, dan menentukan metrik yang relevan dengan hati-hati.

Evaluasi dapat dilakukan secara otomatis maupun manual. Evaluasi otomatis biasanya menghitung metrik otomatis (misalnya, ROUGE, BLEU) atau menggunakan model yang kuat untuk menilai tanggapan model. Meskipun proses ini cepat dan relatif murah, evaluasi otomatis memiliki keterbatasan yang dapat diatasi dengan memanfaatkan *annotator* manusia untuk memverifikasi tanggap model secara manual.

Terdapat berbagai tolak ukur publik yang dapat digunakan untuk menilai kemampuan bahasa umum dan pengetahuan faktual. SEA HELM adalah tolak ukur otomatis yang mencakup pemahaman, penalaran, dan pembuatan bahasa. SEA HELM mendukung Bahasa Indonesia dan beberapa bahasa daerahnya. Dalam tolak ukur keselamatan, dataset dan platform yang tersedia sebagian besar berfokus pada Bahasa Inggris. Oleh karena itu, para *developer* disarankan untuk mengidentifikasi kerentanan dan potensi risiko pada modelnya dengan menyusun *prompt* yang dapat menghasilkan tanggapan yang tidak aman.

## B. Safety Alignment

Sahabat-AI has gone through various steps of safety alignment during its development. However, as with any LLM there will remain risks such as hallucination that needs to be managed. Hence, additional consideration should be taken into account by following the application design mentioned to manage the potential safety in utilizing Sahabat-AI.

Anyone who downloads and uses the AI model must ensure that their developers engineers test the model thoroughly, tailor them to their specific needs and challenges, and have a dedicated team to ensure trust and safety alignment.

## C. Model Evaluation

Rigorous evaluation is strongly recommended to ensure the model generates reliable and safe content that aligns with the system's policies. This is an iterative process until the model achieves satisfactory performance in the intended use cases and safety standards. Developers are encouraged to take necessary actions to provide accurate evaluation results. This may include conducting the evaluation on holdout test sets that reflect data in the production environment, preventing training data leakage to the test sets, and carefully determining the relevant metrics.

Evaluation can be done automatically and manually. Automatic evaluation typically calculates automatic metrics (e.g., ROUGE, BLEU) or utilizes strong models to judge the model's responses. While the process is fast and relatively cheap, automatic evaluation has limitations that can be addressed by leveraging human annotators to manually verify the model's responses.

There are complimentary public benchmarks to assess general language capability and factual knowledge. SEA HELM is an automatic benchmark that covers language understanding, reasoning, and generation. It supports Indonesian and some of its local languages. For safety benchmarks, existing datasets and platforms mostly focus on English. Therefore, developers are suggested to identify model vulnerabilities and potential risks by constructing prompts that maygenerate unsafe responses.




## IV. Penggunaan yang Dilarang

Pengembangan dan penerapan sistem yang didukung oleh Sahabat-AI secara bertanggung jawab memerlukan kepatuhan ketat terhadap prinsip etika dan standar hukum. Untuk memastikan teknologi ini digunakan untuk tujuan yang positif dan sah, penggunaan-penggunaan berikut ini secara tegas dilarang.

**A. Pembuatan Konten yang Berbahaya atau Jahat**

Pengembangan, penerapan, atau penggunaan sistem yang didukung oleh Sahabat-AI untuk menghasilkan konten yang merugikan, kasar, menyesatkan, atau ilegal sangat dilarang. Ini termasuk namun tidak terbatas pada pembuatan konten ujaran kebencian, hasutan untuk kekerasan, informasi palsu, ataupun konten yang dimaksudkan untuk menipu, memanipulasi, atau mengelabui individu. Kami mencatat bahwa tindakan rasisme, penyebaran disinformasi/hoaks, bahasa ofensif, adalah beberapa contoh yang tercantum dalam Surat Edaran Menteri Komunikasi dan Informatika No. 9 Tahun 2023 tentang Etika Kecerdasan Artifisial dan Pedoman ASEAN tentang Tata Kelola dan Etika AI yang harus dihindari dalam pengembangan dan penerapan AI.

**B. Eksploitasi Individu Rentan**

Sahabat-AI tidak boleh digunakan untuk mengeksploitasi, memanipulasi, atau mengambil keuntungan dari individu atau kelompok yang rentan. Hal tersebut termasuk menggunakan Sahabat-AI untuk menargetkan individu berdasarkan data pribadi sensitif, memanipulasi opini, atau menciptakan interaksi yang menyesatkan yang dapat menyebabkan kerugian secara finansial, emosional, atau sosial.

Individu atau kelompok rentan dalam hal ini dapat mencakup anak-anak, orang lanjut usia, penyandang disabilitas yang bergantung pada teknologi, komunitas berpenghasilan rendah, serta kelompok adat yang budayanya rentan disalahgunakan.

**C. Pelanggaran Privasi dan Hukum Perlindungan Data**

Pengembangan sistem yang didukung oleh Sahabat-AI yang mengumpulkan, memproses, atau membagikan data pribadi secara ilegal tanpa persetujuan yang sah, atau yang melanggar peraturan perlindungan data, dilarang. Termasuk penggunaan Sahabat-AI untuk melakukan pengawasan yang tidak sah, pengambilan data tanpa izin, ataupun pembagian informasi pribadi tanpa persetujuan eksplisit dari individu.

## IV. Prohibited Uses

The responsible development and deployment of Sahabat-AI-powered systems require strict adherence to ethical principles and legal standards. To ensure these technologies are used for positive and lawful purposes, the following uses are explicitly prohibited.

**A. Creation of Harmful or Malicious Content**

The development, deployment, or use of Sahabat-AI-powered systems for generating harmful, abusive, misleading, or illegal content is strictly prohibited. This includes but is not limited to generating hate speech, incitement to violence, false information, or content intended to deceive, manipulate, or defraud individuals. We note that acts of racism, the spreading of disinformation or misinformation, offensive language, are some examples listed in the Minister of Communication and Informatics' Circular Letter No. 9 Year 2023 on Ethical Guidelines for Use of Artificial Intelligence and ASEAN Guide on AI Governance and Ethics that should be avoided during AI development and deployment.

**B. Exploitation of Vulnerable Individuals**

Sahabat-AI must not be used to exploit, manipulate, or take advantage of vulnerable individuals or groups. This includes using Sahabat-AI to target individuals based on sensitive personal data, manipulate opinions, or create deceptive interactions that could lead to financial, emotional, or social harm.

Vulnerable individuals or groups may include children, elderly individuals, persons with disabilities who rely on assistive technologies, low-income communities, and indigenous groups whose culture could be misused.

**C. Violation of Privacy and Data Protection Laws**

The development of Sahabat-AI-powered systems that unlawfully collect, process, or share personal data without proper consent, or those that violate data protection laws, is prohibited. This includes the use of Sahabat-AI for unauthorized surveillance, data scraping, or sharing of private information without the explicit consent of individuals.

# V. Strategi Mitigasi dan Manajemen Risiko

Pengembangan dan penerapan sistem yang didukung oleh Sahabat-AI secara bertanggung jawab memerlukan kepatuhan ketat terhadap prinsip etika dan standar hukum. Untuk memastikan teknologi ini digunakan untuk tujuan yang positif dan sah, penggunaan-penggunaan berikut ini secara tegas dilarang.

### A. Strategi Mitigasi

**Penyaringan Konten & Moderasi:** Terapkan mekanisme penyaringan konten yang kuat untuk mendeteksi dan memblokir *output* yang berbahaya, bias, ataupun tidak pantas. Hal tersebut bisa termasuk menggunakan penyaring berbasis aturan, klasifikasi dari pembelajaran mesin, dan moderasi dengan keterlibatan manusia.

**Rekayasa Prompt & Validasi Input:** Melakukan desain terhadap *prompt* dan proses validasi input untuk dapat membimbing interaksi pengguna dan mencegah masukan yang berbahaya atau yang berpotensi disalahgunakan (misalnya *prompt* untuk melakukan serangan injeksi).

**Mekanisme Penjelasan & Transparansi:** Terapkan mekanisme untuk memberikan konteks dan penjelasan kepada pengguna tentang *output* dari LLM, sehingga membantu mereka memahami keterbatasan dan potensi bias (misalnya melalui kartu model).

**Pemantauan & Evaluasi Berkelanjutan:** Terapkan sistem untuk pemantauan berkelanjutan terhadap kinerja aplikasi LLM, umpan balik pengguna, dan potensi penyalahgunaan. Lakukan evaluasi secara teratur terhadap efektivitas strategi mitigasi dan lakukan perbaikanjika diperlukan.

### B. Manajemen Risiko

**Penilaian Risiko secara Proaktif:** Lakukan penilaian risiko menyeluruh sebelum meluncurkan aplikasi yang didukung oleh Sahabat-AI. Hal ini harus melibatkan identifikasi potensi kerugian, bias, kerentanan keamanan, serta penggunaan yang tidak diinginkan.

**Penanganan Masalah Keamanan:** Ketika penggunaan Sahabat-AI menunjukkan adanya kekurangan dalam aspek ketentuan keamanan untuk penggunanya, penyempurnaan model sebaiknya dipertimbangkan dengan melatih model pada set data yang terfokus pada keamanan. Strategi penyelarasan keamanan juga perlu dipertimbangkan untuk mengarahkan model terhadap perilaku yang lebih aman dan bertanggungjawab.

# V. Mitigation Strategies and Risk Management

The responsible development and deployment of Sahabat-AI-powered systems require strict adherence to ethical principles and legal standards. To ensure these technologies are used for positive and lawful purposes, the following uses are explicitly prohibited.

### A. Mitigation Stretagies

**Content Filtering & Moderation:** Implement robust content filtering mechanism to detect and block harmful, biased, or inappropriate outputs. This can involve rule-based filters, machine learning classifiers, and human-in-the-loop moderation.

**Prompt Engineering & Input Validation:** Design prompts and input validation processes to guide user interactions and prevent malicious or harmful input that could lead to misuse (e.g. prompt injection attack).

**Explainability & Transparency Mechanisms:** Implement mechanisms to provide users with context and explanations about LLM outputs, helping them understand limitations and potential biases (e.g. model cards).

**Continuous Monitoring & Evaluation:** Implement systems for continuous monitoring of LLM application performance, user feedback and potential misuse. Regularly evaluate the effectiveness of mitigation strategies and improve as needed.

### B. Risk Management

**Proactive Risk Assessment:** Conduct thorough risk assessments before deploying any Sahabat-AI-powered application. This should involve identifying potential harms, biases, security vulnerabilities, and unintended uses.

**Addressing safety issues:** When Sahabat-AI use exhibits insufficient safety for user requirements, fine-tuning should be considered to train models specifically on curated safety-focused dataset. Safety alignment strategies should also be considered to guide the model towards safer and more responsible behavior.

## Informasi Kontak dan Bantuan

Pertanyaan terkait dengan model Sahabat-AI, dan/atau kesempatan kolaborasi, silakan untuk mengirimkan surat elektronik ke alamat **support@sahabat-ai.com.** Kami akan dengan senang hati membantu dan mengeksplorasi potensi sinergi.

## Contact Information and Support

For any inquiries regarding the Sahabat-AI model, and/or potential collaboration opportunities, please send an email to **support@sahabat-ai.com.** We would be happy to assist you and explore possible synergies.



# Sahabat-AI

Brought to you by

indosat
OOREDOO HUTCHISON